

NATIONAL ACADEMY OF SCIENCES OF UKRAINE
INSTITUTE FOR INFORMATION REGISTRATION PROBLEMS

**A.G. Dodonov, D.V. Lande,
V.V. Prishchepa, V.G. Putyatin**

**COMPUTER
COMPETITIVE INTELLIGENCE**

Kyiv – 2021

UDC 004.5

BBC 22.18, 32.81, 60.54

C95

A.G. Dodonov, D.V. Lande, V.V. Prishchepa, V.G. Putyatin
Computer competitive intelligence. – Kyiv: "Engineering", 2021. –
354 p.

The book is devoted to the issues of computer competitive intelligence, intelligence in the open resources of the Internet. Computer competitive intelligence covers automated procedures for collecting and analytical processing of information that are carried out to support managerial decision-making, increase competitiveness exclusively from open sources in computer networks – websites, blogosphere, social networks, instant messengers, databases. The book deals with various issues of information and analytical activities in the network environment. Elements of information theory, social network analysis, information and mathematical modeling are considered as the theoretical foundations of computer competitive intelligence.

For a wide range of professionals in the field of information technology and security.

*Recommended for publication by the Academic Council of the Institute for Information Registration Problems of the National Academy of Sciences of Ukraine
(Protocol No. 9 dated February 17, 2021)*

Reviewers:

Corresponding member National Academy of Sciences of Ukraine, Doctor of Technical Sciences, Professor *V.V. Mohor*
Doctor of Technical Sciences, Professor *A.Ya. Matov*
Doctor of Law, Professor *K.I. Belyakov*

ISBN 978-966-2344-79-0

©A.G. Dodonov, D.V. Lande,
V.V. Prishchepa, V.G. Putyatin,
2021

Table of contents

Introduction	6
1. Competitive intelligence and OSINT	13
1.1. Competitive Intelligence Tasks	13
1.2. Features of computer competitive intelligence	15
1.3. Problems of computer competitive intelligence	16
1.4. OSINT - open source intelligence	18
1.4.1. <i>OSINT as an intelligence discipline</i>	18
1.4.2. <i>Applications for OSINT</i>	22
1.4.3. <i>OSINT Technologies</i>	24
1.4.4. <i>International experience</i>	26
2. Computer technologies of competitive intelligence	28
2.1. Finding information on the Internet	33
2.2. Information space monitoring	39
2.3. Text Mining , Information Extraction	42
2.4. Domain Models	46
2.5. Big Concept Data	52
2.5.1. <i>Concept of Big Data</i>	52
2.5.2. <i>Big Data Techniques</i>	56
2.5.3. <i>Big data technologies and tools</i>	64
2.6. Mathematical foundations	85
2.6.1. <i>Time series</i>	87

2.6.2. <i>Correlation analysis</i>	93
2.6.3. <i>Fourier analysis</i>	98
2.6.4. <i>Wavelet analysis</i>	101
2.6.5. <i>Pattern correlation</i>	113
2.6.6. <i>Fractal analysis</i>	116
2.6.7. <i>Multifractal analysis</i>	125
2.6.8. <i>Network models</i>	135
2.7. Implemented competitive intelligence technologies	151
3. Sources of information	183
3.1. Websites	186
3.2. Social networks, blogs	191
3.2.1. <i>Major social networks</i>	194
3.2.2. <i>Social media monitoring</i>	197
3.2.3. <i>Social media analysis</i>	201
3.3. Deep web, special databases	203
3.3.1. <i>Understanding the Deep Web</i>	204
3.3.2. <i>Types of Deep Web Resources</i>	206
3.3.3. <i>Deep Web Services</i>	211
3.3.4. <i>Special databases</i>	213
4. Reputation analysis	218
4.1. Reputation management problem	218
4.2. Modeling reputation in networks	223
4.3. Rating of Internet resources	230

5. Legal issues of competitive intelligence	239
5.1. Competitive intelligence in the legal field	239
5.2. Competitive Intelligence and Trade Secret Protection	242
5.3. Competitive Intelligence	244
and protection of personal data	244
5.4. Competitive intelligence and copyright protection	251
6. Opposition to information operations	253
6.1. Information influence, attacks and operations	257
6.2. Stages of information operations	259
6.3. Information Operations Modeling	263
6.4. Identification of information operations	276
6.5. Ways to counter information operations	285
6.6. Examples of information operations	286
Conclusion	292
Brief glossary	294
Literature	309
Competitive intelligence websites	316
Addresses of mentioned web resources	317

Introduction

Computer competitive intelligence (Computer Competitive Intelligence) covers the procedures for collecting and processing information carried out to support managerial decision-making, increase the competitiveness of organizations exclusively from open sources from computer networks, most of which are built on top of the Internet, the so-called overlays. Therefore, the term Internet intelligence is often used as a synonym for competitive intelligence. Thus, this book is actually devoted to the problems of competitive intelligence, but with one significant limitation – all sources of information necessary for conducting intelligence activities are open and available on computer networks. Moreover, most of the tools, information processing programs, are also freely available through modern computer networks. In the English-language literature, this type of intelligence is usually called open source intelligence (Open Sources INTelligence, OSINT) [Bird, 2007], which can also be considered a synonym for the term “competitive intelligence”. However, it should be noted that in foreign literature, the use of OSINT is largely limited to use in the public sphere. But it is for OSINT technologies that the largest number of methods, techniques and technologies have been created.

Intelligence information can be obtained from official sources, other open sources, the media, announcements, advertising, intra-company, banking, government reports, databases, experts, by obtaining (collecting), analyzing or special processing of data, texts. True, at the same time, the amount of heterogeneous information that needs to be processed in order to obtain grains of knowledge is huge, and therefore, at present, competitive intelligence is unthinkable without the use of specialized information technologies, the practical application of the modern concept of big data (Big data).

According to the former director of the US Central Intelligence Agency (CIA) R. Hillenkert, “80% of intelligence information is obtained from sources such as books, magazines, scientific and technical reviews, photographs, commercial analytical reports, newspapers, television and radio programs...”.

According to other estimates, in any intelligence service, from 35 to 95% of all information is obtained from open sources.

At the same time, the share of costs for working with open sources, for example, in the US intelligence budget, is only about 1%.

It is known that for business structures, 95% of useful information is provided by competitive intelligence, 4.1% of information can be legally obtained from government agencies. Only large companies can afford full -fledged business intelligence in the markets, but the possibilities of competitive intelligence are available to almost everyone [https://trademaster.ua/articles/312620].

The importance of open source intelligence was noted by US President Lyndon Johnson (Lyndon Baines Johnson) On June 30 1966, when he delivered a speech at the swearing-in ceremony by CIA Director Richard M. Helms: "The highest achievement is not the result of secret information quietly recounted, but results from patient, hourly study of printed sources."

According to the well-established erroneous opinion, all useful intelligence information is obtained from secret sources through intelligence or operational means – in fact, this is not the case. The well-known confession of Admiral Zaharias, Deputy Chief of Intelligence of the US Navy during the Second World War, refutes this. So, according to him 95 % of information the intelligence of the naval forces drew from open sources, 4 % – from the official ones, and only 1 % – from confidential sources. In fairness, it must be said that often this one percent is the golden missing link that allows you to put together a holistic picture of a disparate mosaic of all intelligence. And if such a ratio is true for military intelligence, then it will be even more correct for competitive intelligence for business structures.

At the same time, an analysis of the declassified CIA report for 1987 "Enterprise-Level Computing in the Soviet Economy" (SOV C87-10043) gives an idea of what a colossal amount of data the analysts had to process. To compile the report, 347 open sources were constantly scanned throughout the year ; to create a one-page summary, an information array of approximately 7 million words was processed daily.

It is well known that the main difference between competitive intelligence and industrial espionage is legitimacy and adherence to ethical standards [Dudikhin, 2004]. Here, this provi-

sion is brought to the absolute – only all sources of information in this case are available and legal.

Internet intelligence, intelligence from Internet sources, as well as all competitive intelligence, is a special type of information and analytical work that allows you to collect versatile business information without using those specific methods of operational-search activities that are the exclusive prerogative law enforcement.

At the same time, the methods of conducting Internet intelligence, the techniques and technologies for its implementation are very close to those used in traditional intelligence activities by special services.

The use of Internet intelligence in a commercial company is justified not only by information security considerations, but it is also important for solving management and marketing problems in that it provides:

- monitoring the reputation of the company (from the point of view of customers, competitors, government agencies);
- active participation in the formation of the company's image, the information field around the company;
- tracking the emergence of a new competitor, technology or distribution channel;
- identification of possible mergers and acquisitions;
- assessment of potential risks in investments;
- outpacing the steps of competitors in the framework of marketing campaigns;
- ahead of competitors in tenders;
- identification of information leakage channels.

The shaky line between the concepts of competitive intelligence and industrial espionage lies in the legitimacy, legality of the methods and means used in the process of collecting targeted information [Lande, Prishchepa, 2007]. It should also be noted that there is a very subtle difference between business intelligence (Business Intelligence, BI) and competitive intelligence. From publications and descriptions of systems where these terms are mentioned, it can be concluded that business intelligence is aimed more at studying "internal" marketing, financial, economic information and information about customers, while

competitive intelligence more often covers processes related to mining "external" information and knowledge directly about competitors.

The founder of modern business intelligence is Xerox, which faced competition from Japanese manufacturers [Prescott, 2003]. In the early 1970s, after the Japanese entered the American market, Xerox managers noticed that the company began to lose its position in the market. The situation was corrected by changes based on the collection of up-to-date information about the market and competitors. Xerox, thanks to its Japanese branch, created a system for evaluating and analyzing work (benchmarking), and then adapted and applied intelligence technologies to business. At the same time, one of the main conditions for organizing this process was the relentless observance of the law, since the company's reputation could collapse much earlier than the economic advantages of industrial espionage could be taken advantage of. Soon, these methods of work began to be adopted by other American companies. Then business intelligence began to be applied in Europe, and later throughout the world.

Ignoring business intelligence opportunities at the initial stage was costly even for the largest companies [Jilad, 2010]. So after creating a camera that produced a finished picture, Polaroid began to rest on its laurels. When the company's analytical department presented a report that pointed to the prospects for the development of the photographic industry and the dawn of the digital age, the company's management called this information "futuristic nonsense." Some time passed, and in October 2001 Polaroid filed its first bankruptcy procedure.

the 1970s, the "Big Three" American car manufacturers did not respond to the entry of Japanese car manufacturers into the market. However, the Americans themselves chose small, economical and reliable Japanese cars, and American corporations suffered significant losses.

business intelligence has learned from the open press that America's last guitar factory may be closing due to cheaper Korean instruments, and the US government is preparing to protect its makers with customs duties. Knowing this in time, Samsung representatives managed to import a large number of guitars into the United States, and as a result of the introduction of import duties, they also raised prices for this musical instrument.

Today's development of information technology has made computer competitive intelligence available even to relatively small companies, today it is widespread at all levels of the economy.

In practice, the conceptual base of competitive intelligence has not yet been fully formed, until a distinction is made between the terms “business” or “economic” intelligence, and competitive intelligence is mistakenly understood as the whole range of activities related to information and analytical support for managing business risks, identifying threats, opportunities. and other factors influencing obtaining competitive advantages in business.

In the arsenal of those who today are fully engaged in competitive intelligence, there is no special equipment, spy equipment. Their main tool is a computer connected to the Internet. The activities of competitive intelligence units (services) of companies are increasingly based on the latest achievements in the field of artificial intelligence, combined with developments in the fields of psychology, sociology, and economics.

The tangible benefits derived from the use of competitive intelligence confirm the results of a survey conducted back in 1999. among the top 500 US companies. Almost 90% of companies confirmed that they have created competitive intelligence units. At the same time, corporations spend on exploration on average 1–1.5% of their turnover and are quite profitable [Lande, Prishchepa, 2007].

Numerous professional associations (communities) of specialists in the field of competitive intelligence are currently being created. The most well-known of these communities organizing conferences and trainings are Strategic and Competitive Intelligence Professionals , SCIP in the USA (www.scip.org) (Fig. 1) and Competia in Canada (www.competia.com).

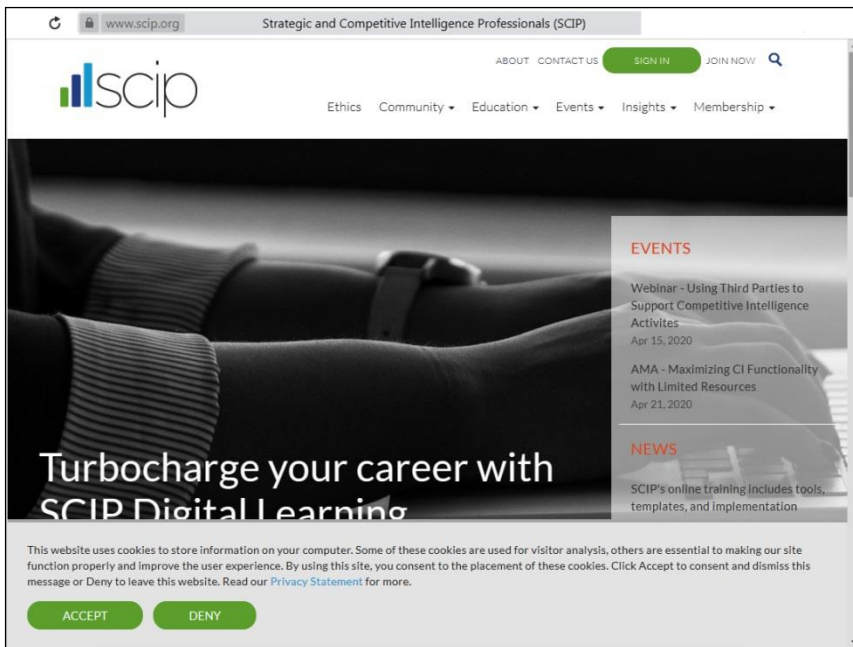


Figure 1. Fragment of the SCIP website (www.scip.org)

In Ukraine, the public organization "Society of Analysts and Competitive Intelligence Professionals" is well-known (<https://www.scip.org.ua/>). In Ukraine, specialists in the field of competitive intelligence are also being trained at the Kharkiv National University of Radio Electronics, where they train masters in the specialty "Consolidated Information".

With the beginning of Russian aggression against Ukraine, the issues of obtaining information from open sources and debunking fakes have acquired a new meaning and relevance. This gave rise to new projects and learning platforms, the purpose of which was the formation and dissemination of skills and abilities to work with information online, check information, etc. One of these projects was OSINT Academy and free online OSINT course (Open Source Intelligence) of the Institute of Post-Industrial Society [OSINT, 2018].

Currently, competitive intelligence is not limited to the study of competitors, but analyzes the entire environment surrounding

an organization or enterprise. The political situation, peculiarities of legislation, personnel transfers, new technologies, the company's own customers and suppliers, etc. are studied, experts on special issues are selected.

1. Competitive intelligence and OSINT

1.1. Competitive Intelligence Tasks

The main tasks of Internet intelligence as a segment of competitive intelligence [Kochergov, 2009] are:

1. Information support for the process of making managerial decisions at the strategic and tactical levels;
2. Early warning, i.e. drawing the attention of decision makers to threats that could potentially harm the business;
3. Forecast and prevention of possible threats to business;
4. Identification (together with the security service) of competitors' attempts to gain access to the company's classified information.
5. Identification of favorable business opportunities;
6. Risk management, ensuring the company's effective response to rapid changes in the environment, Internet space ;
7. Industrial counterintelligence, anticipation of competitors' intelligence activities in a network environment, analytical support of the company's security service.

The above tasks of competitive intelligence are key, they serve to achieve the fundamental goal of competitive intelligence – to ensure the security of the company, the realization of the fact that its fate is in the hands of decision makers, that the company will not suddenly become a victim of someone's hostile actions.

In addition, within the framework of computer competitive intelligence, the following tasks should be solved:

- collection and timely provision of management and business units of the company with reliable and comprehensive information from network sources about the "external" and "internal" environment of the enterprise;
- identification of risk factors, threats that may affect the economic interests of the business or interfere with its normal functioning;
- identification of new opportunities and other factors influencing obtaining competitive advantages;

- strengthening of favorable and localization of unfavorable factors of the competitive environment for the activities of the business structure ;
- development of forecasts and recommendations on the impact of the competitive environment on the activities of the business structure.

Competitive intelligence is becoming a modern direction in the study of the behavior of competitors in the market, allowing you to create models of the market, its participants, determine the characteristics and optimize the tactics and strategies for the development of business entities in certain markets. To solve its problems, it is necessary to use effective methods of working with information and its elements. Information thus becomes both the object of market research and the basis for creating its model.

Above, the tasks of computer competitive intelligence are formulated, designed for the legitimate activities of the relevant structures. The entire system of competitive intelligence should allow the company's management not only to quickly respond to changes in the situation on the markets, but also to assess further opportunities for its development. Competitive intelligence provides a transition from traditional decision-making based on insufficient information to knowledge-based management. At the same time, it also provides risk reduction, business security, as well as the acquisition of competitive advantages. The modern system of competitive intelligence allows not only to monitor information, but also to model the strategy of competitors, identify their partners, suppliers, and understand the terms of cooperation.

The main tasks of competitive intelligence systems are to find and summarize information about competitors, markets, products, business trends and operations on such main objects:

- partners, shareholders, subcontractors, allies, contractors, customers, competitors;
- company combinations, mergers, acquisitions, crisis situations, etc.;
- personnel composition of the company, partners, competitors, etc., as well as personnel changes, their dynamics;

- trade turnover, budget and its distribution by points;
- concluded contracts, agreements or arrangements.

Interest in conducting competitive intelligence is caused by the scope of activities of companies, their spheres of influence and interests. This knowledge can be used, for example, to influence the positions of partners and opponents. Of great importance is information related to the policy of competitors, their intentions, strengths and weaknesses, products and services, prices, advertising campaigns, and other market parameters.

1.2. Features of computer competitive intelligence

Modern open network resources, websites, social networks, video services, instant messengers are now becoming the main source of information and an effective tool for competitive intelligence. They allow not only to track the actions of competing companies in real time, but also to identify the latest trends on topics of interest. Let us name just some of the ways to use Internet resources to solve competitive intelligence problems [Lande, 2019]:

1. Receiving news on the target topic.

Modern online news services such as Google News, Yahoo News, UAPort. net, social networks such as Twitter, FaceBook, Reddit allow you to receive news selected in accordance with the information needs of users. For example, when using the social network Twitter, you can use the search mode and enter a query, such as "bankruptcy". The user will then receive a list of messages, in some cases provided with user accounts, whose messages are relevant to the query entered. Thus, it is possible to define experts who can be grouped according to their information needs. Then, following the opinion of a group of experts, one can get a fairly wide coverage of the problem, several points of view, new information resources.

2. Identification of trends.

Information resources (websites, social networks, etc.) selected using search capabilities can be used to manually or using special analytical tools to identify trends in the selected area.

3. Receiving distribution of targeted documents by subscription (messengers, e-mail, SMS).

Many of the news aggregators and social networks (in particular, Twitter) provide the opportunity for high-quality personalized periodic mailings, covering messages, comments, and expert channels.

4. Building networks of information links, cognitive maps

For the tasks of competitive intelligence, it is important not only to obtain targeted information (messages), but also to understand the links that are found in the analysis of information. Not only the object of analysis is important, but also the information resources associated with it, profiles in social networks, "friends", discussion groups, etc. In some cases, you can see who is the subscriber of these profiles, who is interested in the same topic and, therefore, can become a new source for obtaining targeted information.

5. Getting answers to questions.

Social networks, forums, blogs can be used as a way to get answers to specific questions, including those on competitive intelligence methodology. If the question is posed correctly, then with a high probability you can get an answer to it from other users.

6. Filtering "garbage".

Competitive intelligence is not always interested in well-known, often false data and information (Fake News), interesting to most, and social networks are focused on such data. When using network resources as a powerful base for competitive information, special attention should be paid to processing requests, selecting sources, experts, and establishing links.

1.3. Problems of computer competitive intelligence

Let's note some problems connected with computer competitive intelligence.

First and the most significant problem is that the colossal volumes of information on the Internet, in particular in social networks, make it difficult to find and select the information that is really needed. By itself, raw, unaggregated and unverified data cannot provide quality support for competitive intelligence decision making.

Currently, search engines, in particular, the Google system, indexes more than a trillion documents, the volumes are con-

stantly growing. Along with this, according to Eric Emerson Schmidt, chairman of the board of directors of Google from 2001 to 2011, even such a powerful search engine as Google will be able to index all the information available today only after about 300 years.

Traditional Internet search engines do a great job with simple one-time queries, but, as a rule, are poorly applicable for competitive intelligence needs. According to some estimates [Dodonov, 2014], more than 97% of open information critical for competitive intelligence cannot be found using traditional information search systems.

The second problem of computer competitive intelligence is that information on the Internet is dynamic: it is posted, modified and deleted. A partial solution to these problems is possible with the use of content monitoring systems for information flows on the Internet.

The third problem that needs to be solved for the purposes of competitive intelligence is the automatic extraction of concepts from formalized information arrays (tables, databases), as well as unstructured texts. A promising direction for solving this problem in competitive intelligence systems is the use of Knowledge Discovery, Data Mining and Text Mining technologies [Lande, 2005, Nada new], [Pechenkin, 2004].

The fourth problem is the ability to automatically identify non-obvious patterns and relationships recorded in documents. Currently, there are several ways to solve the problems of extracting concepts from texts and identifying their relationships, both practical and theoretical. One of these ways is the construction of matrices and graphs of connections between concepts, models of subject areas, cognitive maps, to which appropriate mathematical methods can be applied. As a rule, the nodes of these graphs are coefficients that are proportional to the number of documents corresponding to the concepts under study.

The fifth problem is the search for information in the "hidden" Internet, which contains an incomparably greater amount of data potentially interesting for competitive intelligence than in the open part of the network. Not all potentially open "unclassified" information is readily available, rather the opposite. Obtaining the necessary information in each specific case is a difficult task. According to experts, only about 10-15% of the necessary

information is available on the Internet in a ready-made form, the remaining 85-90% can be obtained as a result of comparison, aggregation and analysis of numerous disparate data.

So, the Internet contains most of the information necessary for competitive intelligence, but the question of finding and effectively using it remains open. The reason is the inherent disadvantages of the Internet [Lande, 2005]:

- disproportionate increase in the level of information noise;
- dominance of parasitic information;
- weak structured and coherent information;
- dynamism of information;
- lack of information integrity;
- repeated duplication of information;
- lack of possibility of semantic search;
- limited access to the "hidden" web.

Despite this, the possibilities of the Internet are highly valued by experts in the field of competitive intelligence.

1.4. OSINT – open source intelligence

1.4.1. OSINT as an intelligence discipline

As one of the synonyms for the concept of competitive intelligence, often used in law enforcement agencies of various states, the concept of “open source intelligence” OSINT (Open source intelligence) is used. This is one of the areas of intelligence, which includes the search, selection and extraction of intelligence information obtained from publicly available sources (not necessarily computer or network), as well as the analysis of this information.

OSINT is based on two main concepts:

- open source is a source of information that provides it without the requirement of maintaining its confidentiality, i.e. provides information that is not protected from public disclosure. Open sources refer to the public information environment, and have no access restrictions for individuals;

- public information is information published or made available for general use; available to the public.

According to CIA analyst Sherman Kent in 1947, politicians receive from open sources up to 80 percent of the information they need to make decisions in peacetime. Later, Lieutenant General Samuel Wilson, head of the US Department of Defense Intelligence Agency in 1976-1977, noted that "90 percent of intelligence comes from open sources and only 10 comes from the work of agents."

American security researcher Mark M. Lowenthal defines open information as "any information that can be obtained from open collections: all types of media, government reports and other documents, scientific studies and reports, commercial information providers, the Internet, etc. The main characteristic of open information is that it does not require illegal collection methods to obtain it and that it can be obtained by means that are fully consistent with the copyrights and commercial terms of the providers.

The world community is increasingly using information from open sources to solve a wide range of problems. OSINT materials serve as the basis for all methods of conducting intelligence as a store of intelligence data, their analyzer and disseminator.

According to [ATP, 2012], open source intelligence OSINT is one of the methods of conducting intelligence, which makes a significant contribution to the planning of military operations, and also provides all the necessary information during their conduct. Also defined:

1) Open source intelligence (OSINT) is one of the methods of conducting intelligence by collecting information from open sources, analyzing it, preparing it and providing the final product to higher management in a timely manner in order to solve certain intelligence tasks.

2) OSINT is an intelligence method developed from the collection and analysis of publicly available information and is not under the direct control of the US government. OSINT is the result of a systematic collection, processing and analysis of the necessary public information.

In particular, the role of OSINT in exploration is determined by a number of aspects, including the speed of receipt, volume, quality, clarity, ease of further use, cost of obtaining, etc. The following factors influence the planning and preparation of OSINT maintenance:

- Effective information support. Most of the necessary reference materials about the objects of information operations are obtained from open sources. This is mainly achieved by collecting information from the media. The accumulation of data from open sources is the main function of OSINT.
- Relevance. The availability, depth and scope of publicly available information make it possible to find the necessary information without the involvement of specialized human and technical means of intelligence.
- Simplification of data mining processes. OSINT provides the necessary information, eliminating the need to involve unnecessary technical and human intelligence methods.
- Depth of data analysis. As a formal part of the intelligence process, OSINT allows management to deeply analyze publicly available information in order to make appropriate decisions.
- Efficiency. A sharp reduction in the time of access to information on the Internet. Reduction of man-hours associated with the search for information, people and their relationships based on open sources. Get valuable operational information quickly. The rapidly changing environment during crises is most fully reflected in CNN's ongoing coverage from the scene.
- Volume. Possibility of mass monitoring of certain sources of information in order to search for interesting content, people and events. As experience shows, well-assembled pieces of information from open sources in the aggregate can be equivalent or even more significant than professional intelligence reports.
- Quality. Compared with the reports of special agents, information from open sources turns out to be preferable, if only because it is devoid of subjectivity, not diluted with lies.

- Clarity. So if in the case of using OSINT the reliability of open sources is both clear and unclear, then in the case of secretly obtained data, the degree of their reliability is always in doubt.
- Ease of use. It is customary to surround any secrets with barriers of secrecy stamps, special access modes. As for OSINT data, they can be easily transferred to any interested authorities. It is possible to conduct a comprehensive investigation based on data from the Internet
- PFigure The cost of data mining in OSINT is minimal, determined only by the cost of the service used.

In particular, today, the software and technological solutions offered for OSINT provide:

- collection of data from social networks such as Facebook, Twitter or Youtube, analysis of the collected data;
- extracting the essence of events from the collected content;
- aggregation of information received from the Internet;
- information influence on the Internet;
- assessment of the reliability of information;
- monitoring and recognition of identity on the Internet, including using geolocation;
- work with information obtained from segments of the web space invisible with the help of traditional network search engines (dark web, hidden web, deep web).

1.4.2. Applications of OSINT

There are many applications of OSINT, some of which are:

Intelligence service

Open sources contain a huge amount of information that is necessary and meets the requirements of intelligence agencies, both public and private, commercial, providing an understanding of objective and subjective factors associated, for example, with the activities of competitors. At the same time, of course, in order to increase the effectiveness of intelligence activities, open information is used in combination with other, in particular, undercover resources.

The American Open Source Intelligence Community initiative (known as the National Open Source Enterprise) is expressed by Intelligence Community Directive 301, which is promulgated by the Director of National Intelligence [DNI, 2006]. The directive establishes the powers and responsibilities of the Assistant Deputy Director of National Open Source Intelligence (ADDNI/OS), the DNI Open Source Center, and the National Open Source Committee.

OSINT in the military

The following are US military units that participate in OSINT activities as an example:

- Unified Combatant Command ;
- Defense Intelligence Agency ;
- National Geospatial-Intelligence Agency ;
- US Army Foreign Military Studies Office ;
- EUCOM JAC Molesworth ;
- Foreign Media Monitoring in Support of Information Operations, US Strategic Command .

National security

The US Department of Homeland Security maintains an active open source intelligence unit. On February 14, 2007, the "Domestic Open Source Enterprise" was established to support the OSINT department and work with state, local and tribal partners.

Justice

The OSINT law enforcement community applies open source intelligence to predict, prevent, investigate crime and prosecute criminals, including terrorists. In addition, information processing centers (Fusion Centers) in the US are increasingly using OSINT to support their intelligence and investigations. Such centers were originally created under the auspices of the Department of Homeland Security (DHS) and the Department of Justice and allowed the exchange of strategic information between the CIA, the FBI, the Department of Defense, the Ministry of Emergency Situations, as well as local administrations, etc.

Examples of successful OSINT law enforcement are Scotland Yard OSINT; Royal Canadian Mounted Police (RCMP) OSINT.

The New York City Police Department (NYPD) includes the OSINT division as the Los Angeles County Sheriff's Department, located in the Bureau of Emergency Operations and associated with the Los Angeles Joint Regional Intelligence Center.

In terms of law enforcement, OSINT can be used to combat such phenomena as:

- Organized crime and gangs;
- Pedophilia;
- Identity theft and extortion;
- Laundering of money;
- Crime in the field of infringement of intellectual property;
- Activities of extremist organizations.

At the same time, with the help of OSINT, engagement is identified and influence is increased on the Internet:

- Identification of key figures and activists;
- Real-time competitor monitoring ;
- Restriction of dissemination of information;
- Formation of public opinion;
- Identification of extremist organizations;
- Risks for public transport;
- Sanctions and legal requirements;
- Analysis of enemy databases (HME, IED, TTPs);
- Geolocation of targets;

- Support for military operations.

Cyber security

OSINT provides support for cybersecurity processes, in particular, answers can be given to such questions from the field of protecting telecommunications networks by obtaining information:

- Who is attacking your organization?
- What are their motives?
- How are they organized?
- What tools are used?

Business

OSINT in business includes commercial intelligence, intellectual intelligence, and business intelligence, and is often the core practice area of private intelligence agencies.

Businesses may use information brokers and private investigators to collect and analyze relevant information for business purposes, which may include the media, the deep web, the next generation web, and commercial content.

1.4.3. OSINT technologies

OSINT is a very diverse form of information gathering and analysis. When running OSINT, precautions must often be taken when collecting information from the Internet. This can be done in the form of using VPNs for anonymity and inconspicuous collection of information, proxy servers in a distributed network environment. Source evaluation becomes important to the overall OSINT collection and analysis process. The OSINT analyst needs intelligent analysis to identify the true or false process that will affect the prediction of the future. Finally, analysts must find a use for evaluative analysis so that its results can be incorporated into a finished classified, unclassified, or proprietary intellectual product.

Information gathering in OSINT is generally different from data gathering in other intelligence disciplines, where obtaining raw information to be analyzed can be a major difficulty. In OSINT, the main difficulty is identifying relevant, reliable sources from the vast amount of publicly available information.

Stages of OSINT

The OSINT process consists of four stages: planning, preparation, collection and production of the final material – analytics and four main processes: analysis, extraction and accumulation of intelligence, evaluation and distribution by direction. The process of conducting intelligence, as well as the processes of preparing response information operations (planning, preparation, execution and debriefing), overlap and repeat in accordance with the requirements of practice.

As outlined in the Field Reconnaissance Manual, OSINT improves efficiency and supports the reconnaissance and other operations.

On Fig. 2 shows a typical flow diagram of the OSINT maintenance process.

Intelligence gathering synchronizes and integrates the processes of planning, use of forces and means, processing and distribution of system elements to support combat operations, which is a combined intelligence and operational function.

After analysis, information obtained from various sources becomes intelligence data that contains the necessary information about the enemy, threats, climate, weather conditions, terrain, etc.

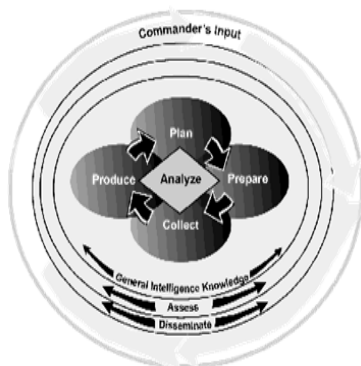


Figure 2 – Typical diagram of the process of maintaining OSINT : Plan ⇒ Preparation ⇒ Collection ⇒ Production. General intelligence knowledge. Grade. Distribution [APT, 2012]

It has been established that such elements of the OSINT structure as a constant flow of information, hardware, software, security of communications and databases cover the means of:

- ensuring the availability of intelligence data. Making intelligence available is a process by which intelligence organizations actively and quickly gain access to intelligence;
- development and maintenance of an automated intelligence network. The main task is to provide information systems that provide communication, joint analysis and processing, dissemination of materials and creation of conditions for the availability of intelligence data;
- creating and maintaining access. This task entails establishing, providing and maintaining access to classified and unclassified programs, databases, networks, systems and other Internet resources for the troops of allied states, joint forces, national agencies and international organizations ;
- creating and maintaining databases. This task involves the creation and maintenance of unclassified and secret databases. The creation and maintenance of a database contributes to rapid analysis, reporting, processing, distribution, and the conduct of long-term hostilities.

1.4.4. International experience

Open source intelligence improves the performance of the entire intelligence community, from the national to the tactical levels. Below is a list of some organizations that are engaged in the extraction, accumulation, use, analysis, and dissemination of information from open sources in the United States.

- Open Source Defense Council (DOSC);
- US Armed Forces Intelligence and Security Command (INSCOM);
- Department of the Army Intelligence Information Service (DA IIS);
- Director of National Intelligence at the Open Source Center (DNI OSC);

- Open Source Academy;
- Advanced Systems Department (ASD);
- FBI;
- Federal Research Division (FRD), Library of Congress.

Along with the widespread use of OSINT in the United States, we will give more examples of the use of this technology in other countries.

Germany's foreign intelligence service, the Federal Intelligence Service, also takes advantage of Open Source Intelligence in the Abteilung Gesamtlage/FIZ and Unterstützende Fachdienste (GU) units.

In Australia, the open source expert is the Office of National Assessments, which is one of the government's intelligence agencies. In the UK, there is an information service BBC Monitoring, focused on the collection of openly available information by journalists. The analysis of the data collected by the BBC is carried out by the subscribers of this service, including employees of the secret British intelligence services.

2. Computer technologies of competitive intelligence

Computer competitive intelligence uses various tools in its arsenal, the most developed of which are specialized information and analytical systems. The scheme of operation of a typical information-analytical system of computer competitive intelligence (IAS KKR) is shown in Fig. 3.

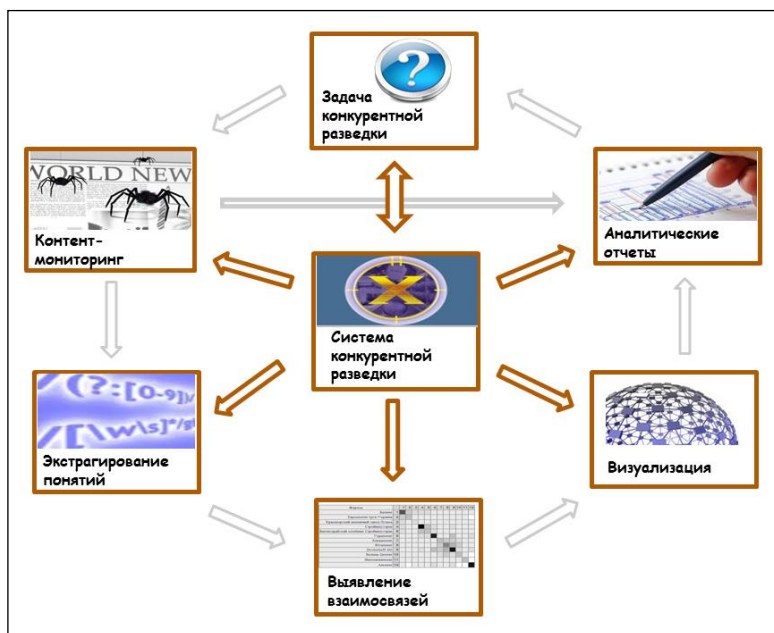


Figure 3 – Scheme of operation of a typical IAS KKR

The information-analytical system of computer competitive intelligence includes the following components:

- complexes for content monitoring of information from open networks (web space, social, peer-to-peer networks, etc.);
- means of extracting concepts (companies, persons, events, etc.) from full-text documents;

- means of identifying and visualizing information links, identifying anomalies, non-obvious patterns;
- means of generating analytical documents that are provided to decision makers (DM).

The content part, the information base of the information and analytical system of competitive intelligence is formed by the content monitoring complex. The features of modern content monitoring complexes are that they must cover huge amounts of information from dynamically increasing information flows in networks in the presence of noise information, a large part of poorly accessible resources, the so-called "hidden Internet". In some cases, the implementation of this complex can be transferred to the so-called "data collection processors", companies that are engaged in the targeted collection of large amounts of information from social media according to customer requirements.

The formation of a database (DB) of the IAS KKR occurs by connecting to the Internet and collecting (according to certain criteria and accounts) information from certain information resources (the list below can be expanded):

Web sites;

Blogs:

Twitter ;
livejournal.

Social media:

Facebook;
Instagram;
LinkedIN
Reddit
medium.

Video services :

YouTube ;
RuTube.

Messengers :

telegram ;
Viber.

In addition, it should be possible for the administrator of the IAS KKR content monitoring complex to configure automatic

scanning and primary processing modules, and, if necessary, create service accounts through which access to certain information resources will be organized.

With the help of content monitoring complexes within the framework of competitive intelligence, as a rule, the following tasks are solved:

- monitoring the activities of partners, competitors, regulators;
- control of media presence and media activity of market participants;
- finding information about market participants;
- identifying new products in the markets;
- identification of new players in the markets;
- organization of a retrospective information fund of documents for their subsequent use in analytical activities.

The process of turning raw data into knowledge and bringing it to end users is commonly called the intelligence cycle. In its classical sense, the reconnaissance cycle (reconnaissance cycle) is usually divided into five main stages:

- target designation, planning, identification of information sources;
- collection, extraction of data;
- processing intelligence data (intelligence data) – turning them into intelligence information;
- analysis and synthesis of intelligence information – its transformation into knowledge – conclusions, recommendations, decisions;
- bringing information to end users.

Some key features of the above stages should also be noted, namely:

- target designation and planning should be divided into three levels – strategic, tactical and operational;
- at the stage of collecting information, it is extremely important to use as many independent and primary sources as possible;

- the data processing process involves accounting, classification, selection, verification and evaluation of the obtained information;
- the intelligence cycle, in some cases, may not require deep study, for example, in a limited time, may not be complete and end with the issuance of knowledge to consumers in the form of final conclusions, recommendations or draft decisions, but simply processed information in the form of information notes;
- an intelligence document should not contain references to confidential sources of information, as this may lead to their decryption;
- conclusions and recommendations should be clear, concise and unambiguous, and forecasts should be probabilistic;
- bringing information to end consumers should be in a form adapted to the perception of the customer and in a form easily accessible to their understanding (it is interesting to note that the CIA, for example, provided US President R. Reagan with daily information in the form of a video film that was filmed every day, since the former film actor perceived such presentation of information more adequately).

So, open sources are the most accessible channel of information, when using them, the objectivity of the information obtained increases, however, the labor costs for extracting the necessary information increase sharply. Therefore, specialized techniques and systems must be applied in computer competitive intelligence. Such specialized techniques and systems have been created by scientists for the benefit of intelligence agencies over the years, both in the West and in the former Soviet Union. Over the past 10–20 years, a significant amount of world information has been transferred from paper to electronic form, the widespread use and growth of the Internet, modern information technologies have made competitive intelligence on the Internet one of the most promising areas of intelligence activity. The fact that almost all the special services of the world do this only confirms the prospects of this direction.

To search and collect information in computer networks in the interests of intelligence around the world, special monitoring data collection systems are used, data collection processors that use special software systems (in computer slang they are called "robots" or "spiders"). The robot program itself bypasses the specified addresses (URL) on the Internet according to a given schedule, downloads data from them, and then extracts the necessary information from them using a whole arsenal of linguistic, semantic and statistical analysis tools. Such software systems automatically intercept any information put on monitoring as soon as it appears in the available network segment.

Information, has become widespread. Extracting (deep analysis of texts / extraction of knowledge from information). The unique features of these concepts and technologies is that they can be used to extract from raw data previously unknown, non-obvious, useful in practice and accessible for interpretation knowledge necessary for decision-making in various fields of activity. Such technologies were used mainly in special services. One of the first declassified such complexes was the French TAIGA system (Traitement Automatique de l'Information Geopolitique d'Actualite – an automatic system for processing relevant geopolitical information) [Doronin, 2003]. This software package was used for 11 years in the interests of French intelligence, after which it was replaced by a more modern one, declassified and allowed for commercial use. The new, more advanced Noemic complex, put into service with French intelligence, is capable of processing information at a speed of more than 1 billion characters per second. The American analogue of these Topic software systems, which means "Theme", has also already been declassified and transferred for commercial use.

Similar analytical systems were created in the former USSR, in particular in Russia. Suffice it to recall such well-known FAPSI systems as "Barometer", "Elbrus". They were engaged in the processing of Russian and foreign press, statistical and operational information.

The creation and use of such systems continues. So, for example, the Radian 6 system (www.radian6.com) is designed to track brand mentions in social networks in real time, taking into

account the tone and to participate in ongoing discussions. Another system, Alterian SM2, also allows you to track brand mentions, as well as localize the places of discussion and determine the demographic characteristics of social network users. As of 2021, the leading systems are ActivTrak, ChartMogul, Cluvio, Databox , Matomo Analytics , Metabase , Tableau (<https://blog.capterra.com/top-8-free-and-open-source-business-intelligence-software/>). In Ukraine, dozens of information and analytical systems of competitive intelligence have also been created and are being developed, which will be discussed below.

At first glance, it may seem that all of the examples listed are systems that are either used by government agencies or are too expensive to be used by “average” companies. In fact, everything is not quite right. On the modern market there are a number of both Western commercial products and domestic products capable of performing similar tasks in one way or another in the interests of competitive intelligence of commercial structures.

2.1. Searching for information on the Internet

In order to get the grains of information needed by the user on the Web, it is necessary to process huge arrays of raw data. Naturally, special search tools are used to facilitate this task.

Searching for information on the Internet only by browsing individual websites, firstly, is selective and / or random (in addition, information on individual sites can be very subjective or even custom-made), and secondly, it is not productive.

All available means of searching for information on the Internet can be conditionally divided into several subgroups, namely:

- means of searching for information on individual sites;
- collections of links, directories;
- search engines;
- metasearch systems;
- monitoring and content analysis systems;
- extractors of objects, events and facts;
- Knowledge Discovery, Data Mining, Text Mining systems ;
- specialized competitive intelligence systems;

- integrated systems.

All directories, search engines and metasearch engines are websites with specialized databases that store information about other websites and documents stored on them. Upon request, such systems receive a list of hyperlinks, and sometimes a brief description of documents (snippets). As a rule, the search can be performed by keywords and phrases. By clicking on the hyperlink found as a result of the request, the user gets to the original document. Naturally, if the document has changed over time or the website has ceased to exist, then the document originally indexed by the search engine may not be found after a while.

The main difference between search engines and directories is the presence of an automatic "robot" that constantly scans the web space and accumulates new information in the index files of the database. As a rule, information is entered into catalogs manually – either by the owners of the sites, or by the maintenance personnel of the catalogs themselves. The use of such systems is usually free of charge.

Metasearch engines are systems that integrate search results across different search engines. Since individual search engines index different segments of the Web in different ways, then, naturally, the search result using a metasearch engine will be more complete than using one single search engine. The second search advantage of such systems is that a single query provides a search in many search engines without requiring multiple repetitions of the same query.

Monitoring systems provide regular search and "download" of information on given topics and from given sites, as well as analysis of the content of "downloaded" documents. Such systems, as a rule, have a developed query language, which allows you to significantly detail and specify queries in comparison with conventional search engines. In addition, such systems store full texts of source documents in their databases, which ensures the safety of these documents in time and the possibility of their processing and content analysis, both in the current time and in retrospect. A significant advantage of such systems is also that complex queries, consisting of tens or hundreds of search words and expressions, once compiled by an expert analyst, can be saved as a cataloged query or rubric and later called automatically or manually from the saved list for searching. or analysis.

With the help of content analysis, such systems make it possible to establish intersecting links between topics, concepts and objects set for monitoring, to identify the emotional coloring of documents, to analyze the dynamics of the appearance of certain documents over time, to conduct a comparative analysis of information activity on various topics, and much more..

If monitoring systems, as filtering systems, can extract known objects from the information flow, then extractors of objects, events and facts can extract from the information flow objects that are unknown in advance, events or facts that only correspond to a predefined type, for example, geographical concepts, persons, structures and organizations, events (traffic accidents, disasters, international meetings). In this case, facts can be classified as ordinary or unusual. An example of an ordinary fact in this case can be considered the departure of cars outside the city, and an example of an unusual fact is the departure of a car without license plates outside the same city limits.

Systems such as Knowledge Discovery, Data Mining and Text Mining technologies, are able to identify new knowledge and patterns. Such a system, for example, can independently, without human intervention, draw a conclusion about the fact of acquaintance between people, based on the data available in the system about their graduation from the same school and the same class in the same locality. True, the very rules by which such a system draws conclusions are nevertheless created and set by people so far.

Specialized systems for competitive intelligence may include one or more of the search tools listed above, adapted to these specific tasks. In addition, the needs of competitive intelligence involve the use as sources of information, in addition to full-text documents, also databases available on the Internet, company-owned documents, tables and databases, as well as formalized and non-formalized documents and databases obtained from other sources.

In the countries of the European Union, an ordinary person is registered in more than 300 databases, such as registration (place of residence), insurance, driver's license, banks, credit bureaus, information, rating, recruiting agencies, employment agencies, medical and police records, supermarkets, clubs, customer relationship management systems for commercial firms

(the so-called CRM systems), etc. In the interests of competitive intelligence and marketing, not only markets for goods and services are analyzed, but also the tastes and preferences of individual customers. The information on legal entities stored in various databases is even more extensive.

For the purposes of business intelligence, it is necessary to analyze data from all available sources of information, but within the framework of this work, sources of information that are not presented on the Internet will not be considered.

Integrated competitive intelligence tools include not only all available search tools, but also a bank of identified (obtained) and logically interconnected data, information and knowledge.

From the point of view of creating information and analytical systems, such a system should conceptually imply the implementation of the following three principles:

- a single information space of interconnected concepts – objects and facts, regardless of the type of their sources or content;
- maintaining links of concepts with relevant data and information sources;
- historical-spatial model of the system data bank, which assumes that all accounting objects have the attributes of time and place.

To be fair, it should be noted that, according to Fuld's Intelligence Software Report, there are no well-known commercial versions of full-fledged integrated systems that allow solving the entire range of competitive intelligence tasks, at least in the West.

According to A. Masalovich, an expert in the field of intelligence, out of 23 types of search tasks that are of interest to intelligence analysts, traditional search systems solve only one satisfactorily. Search engines do a great job with simple one-time queries. When the subject area is complex or too broad (for example, “politics”, “economics”), or, conversely, is extremely narrow and remote in time (for example, the terms of a deal of some companies five years ago), and it is required to summarize all information topics and occasions this topic, evaluate them in temporal dynamics, find relationships with other objects, make a complete picture of the object of interest, select a non-standard event from the general array, then you can make sure that:

- the issuance of search engines is either overloaded with thousands of useless links, or vice versa is insufficient;
- information on the Internet is not stored for a long time, the necessary information present on the target site a month ago may not be found there today;
- the search engine does not save the links viewed by the analyst, and each time he has to start routine work from scratch after a forced break;
- the search engine does not always distinguish really important information from informational noise;
- the search engine is not always able to generalize or compare information in terms of meaning or other meaningful criteria;
- search engines do not cover some web resources or certain types of information (for example, information from databases), and some web resources, on the contrary, are always shown on the first pages of the issue, although their content is not interesting to the authors of requests;
- search engines can search for information only on a directly entered request and cannot always repeat them automatically at a given time without user intervention.

According to experts [Kuznetsov, 2006], a significant part of business-critical information from the Internet cannot be found using traditional information retrieval systems. More precisely, network information retrieval systems do not fully cope with the tasks of competitive intelligence. Therefore, specialized systems are being developed that are focused on the tasks of network analytics and competitive intelligence. A list of such public systems, for example, is given at (http://hrazvedka.ru/category/poisk_soft). Here is a description of some of them:

Website-Finder (www.softpedia.com) is a program that allows you to search for websites that are poorly indexed by the Google search engine. For every request you are given 30 results. The program is easy to use, there is a free version.

Global supplier Directory by Solusource ([www.worldindustrialreporter.com / solusource](http://www.worldindustrialreporter.com/solusource)) is a competitive

intelligence web interface from Thomas. Allows you to search for information available in Thomas' retrospective databases (over 100 years of coverage) by company, product and industry.

dtSearch (www.dtsearch.com) is a search program that allows you to process terabytes of text, both on a local disk and in a network environment. Supports static and dynamic data. Allows you to search in all MS Office formats.

InfoNgen (www.infongen.com) is an aggregator of over 35,000 online sources that can be easily customized to unique topics. It combines monitoring, filtering and aggregation of information at the request of a particular user. Provides information in eight languages, provides translation into English.

Sentinel Vizualizer (www.fmsasg.com) is one of the world's best Sentinel Vizualizers for visualizing connections and relationships.

Web Content Extractor (newprosoft.com) – "Web Content Extractor" is the most powerful, easy to use website data extraction software.

Screen-Scraper (screen-scraper.com) – allows you to automatically extract all information from web pages, download the vast majority of file formats, automatically enter data into various forms. Works under all major platforms, has a fully functional free and very powerful professional versions.

Attackindex (attackindex.com) is a system that allows you to answer the questions: is there an information attack against the user or has there been a natural surge of interest in the event; when the inform operation began, how intense and large-scale it was; which sites and social media accounts are used for the attack; who initiated the information operation and how its participants are connected (Fig. 4).

The screenshot shows the top section of the Attack Index website. At the top left is the logo for 'ATTACK INDEX BE ON YOUR GUARD'. To its right are language selection buttons for 'Ru', 'Ua', and 'En'. Further right are navigation links for 'Новости', 'Исследования', and 'Войти'. The main banner features a background image of a server room and the headline 'КОНТРОЛИРУЙ И КОРРЕКТИРУЙ ИНФОРМАЦИЮ О СЕБЕ'. Below this, a section titled 'Узнайте:' contains a list of six bullet points. At the bottom of the banner is a line graph with the title 'КАК ЭТО РАБОТАЕТ?' and a short paragraph of text.

ATTACK INDEX
BE ON YOUR GUARD

Ru Ua En

Новости Исследования Войти

КОНТРОЛИРУЙ И КОРРЕКТИРУЙ ИНФОРМАЦИЮ О СЕБЕ

Узнайте:

- Ведется ли против вас информационная атака или произошел естественный всплеск интереса к событию.
- Когда началась информ операция, насколько она интенсивна и масштабна
- Какие сайты и аккаунты в соцсетях используются для атаки.
- Есть ли сликеры и кто они, какие эмоции о вас они транслируют обществу или вашим клиентам
- Кто стал инициатором информационной операции и как связаны её участники.
- Как устранить или нивелировать последствия атаки
- Типичные сценарии противодействия информационным угрозам

КАК ЭТО РАБОТАЕТ?

Сервис использует технологии больших данных (Big Data), комплекс аналитических алгоритмов AttackIndex, а также инструменты сбора и визуализации данных об атаке. Мы используем открытые источники, базы мониторинга и массивы данных из онлайн-СМИ и социальных сетей. Анализ можно проводить за любой период вплоть до 10 лет. Тестовая версия дает возможность исследования за последний квартал.

Fig. 4 – Fragment pages Attack Index website (attackindex.com)

Photoinvestigator (photoinvestigator.co) is a service for extracting metadata and other information from photographs.

Visual.ly (visual.ly) is a search engine for infographics on the web.

CIRadar (www.ciradar.com/Competitive-Analysis.aspx) is a commercial English-language system for searching information for competitive intelligence in the deep web. Implemented as a web service.

2.2. Information space monitoring

Modern methods of content monitoring are the adaptation of the classical methods of content analysis and text mining (Text Mining) to the conditions for the formation and development of dynamic information arrays, for example, information flows from the Internet. The first typical task of content monitoring is the construction of diagrams of the dynamics of the emergence of concepts (reflection of events) over time.

On the example of the oil products market, let's consider how documents containing the maximum amount of price infor-

mation on this market can be identified from the arrays of textual information from the Internet.

Let us consider how the content monitoring system InfoStream [Grigoriev, 2007] monitors publications related to the Russian-Ukrainian gas crisis of 2008–2009. To do this, a request was made for **"gases ~ to riz & geo.UA"**, entered through the system web interface.

The diagram that appeared after processing the request shows that the crisis peaked in mid-January 2009 (Fig. 5) and was associated with the signing of the relevant agreement in the Kremlin and the reaction of the Secretariat of the President of Ukraine (Fig. 6).

In addition, you can switch to the "Plots" mode, which provides for clustering search results based on weight criteria, which allows you to give the user only the most significant chains of documents. Therefore, a sufficiently high level of correspondence between the issued documents and the needs expressed by the request is ensured. To obtain a list of the main topics related to the oil products market, the query **"(petroleum | gasoline) & prices"** was introduced, which was specified by special features **"numb.medium | numb.large"**, meaning in the InfoStream system an average or high level of presence in digital documents. information (Fig. 7). After that, it is enough to switch to the view mode and analyze the documents, links to which are issued by the system (Fig. 8).

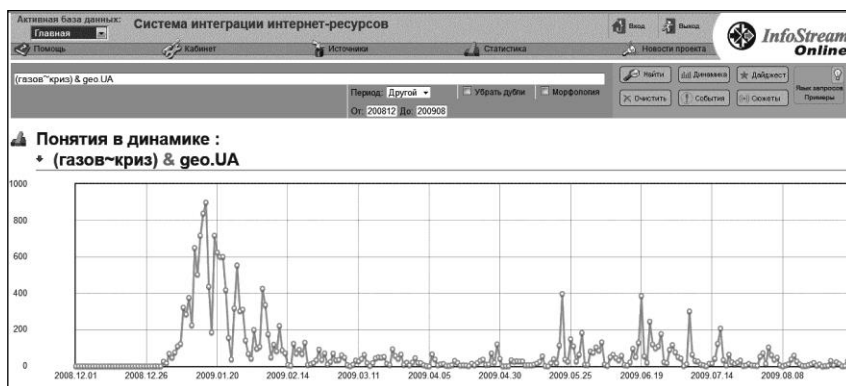


Figure 5 – Diagram of the dynamics of the concept in time

Обзор основных сюжетов
 (газов-криз) & geo.UA & (2009.01.16);
 документов - 903, сюжетов - 112

Секретариат Президента: газовый кризис начался с приходом Тимошенко в Кабинет Министров
 2009.01.16 17:35
 Заместитель главы Секретариата Президента Роман Бессмертный заявляет, что начало сегодняшнего газового кризиса следует искать в 2005 году с приходом Юлии Тимошенко на премьерскую должность. Бессмертный отметил, что на момент прихода Юлии Тимошенко на премьерскую должность в 2005 году между Украиной и Россией была полностью сформирована договорная правовая база и необходимо было лишь ежегодно подписывать

Дубли - Похожие документы - Оригинал
Всего в сюжете сообщений: 90
Первое сообщение: ХайВэй, 2009.01.16 01:33
Ключевые слова: ГАЗ ПРЕЗИДЕНТ УКРАИНЫ РОССИЙСКИЙ ТИМОШЕНКО ЕВРОП МЕДВЕД ПРЕМЬЕР ГАЗОВ ТРАНЗИТ КРИЗИС РЕШЕН УКРАИНСКИЙ СЕКРЕТАРИАТ БЕЗСМЕРТН ЕВРОПЕЙСК МОСКВ МЕЖДУНАРОДН КОНФЛИКТ КИЕВ

2009.01.19 07:30 Российские национал - патриоты готовятся "стирнуть" у коммунистов последний козырь - результаты Всесоюзного референдума 17 марта 1991 года *Славянская Европа*
 2009.01.18 00:23 Особое мнение: "украинский вопрос" придется решать и без Путина с Медведевым *"Forum.msk.ru"*
 2009.01.16 23:53 Тимошенко взяла на себя ответственность за газовые переговоры *Политика de facto*
 2009.01.16 22:20 Мини-саммит в Киеве *EuroNews*
 2009.01.16 21:36 Тимошенко взяла на себя ответственность за газовые переговоры *"Lenta.Ru"*
 2009.01.16 20:35 Секретариат Ющенко просит прокуратуру проверить Тимошенко *"Комсомольская Правда" в Украине*
 2009.01.16 20:31 У Ющенко снова вспомнили о любви между Тимошенко и Путиным *Настоящий Дозор*
 2009.01.16 20:30 Европейский бизнес готов разделить с Россией риск по транзиту топлива через Украину *Первый канал*
 2009.01.16 20:23 Тимошенко берет на себя преодоление газового кризиса. Заявление Премьера *Главное*
 2009.01.16 20:10 Банковая вновь призывает ГПУ "проверить" Тимошенко *Цензор Нет*
 2009.01.16 20:06 Бессмертный требует судить Тимошенко как врага нации *Обозреватель*
 2009.01.16 20:02 Секретариат просит ГПУ начать шить дело Тимошенко *From-UA.com*

Figure 6 – Main storyline on request

Обзор основных сюжетов
 (нефтепродукт | бензин) & цены) & (большая цифровая насыщенность));
 документов - 38, сюжетов - 8

1. **Средние розничные цены на топливо в Украине на утро 25 июня**
 По итогам 21 июня зафиксировали следующие розничные цены на топливо: Топливо Цена, грн. Min/Max Бензин А-7680 10.23 9.80/10.50 Бензин А-92 10.54 10.20/10.79 Бензин А-95 10.91 10.50/11.29 Бензин А-95+ 11.69 10.85/12.59 Дизельное топливо 9.97 9.29/10.29 По данным консалтинговой группы "А-95".
 Сюжет полностью (14)

2013.06.24 04:49 Бензин в Приморье стал дешевле *Delta.RU*
 14
 2013.06.27 15:30 Цены на крупнооптовом рынке нефтепродуктов Украины 27 июня понизились. *Passage.com.ua*

2. **Розничная цена сжиженного газа стала меньше**
 По результатам мониторинга рынка нефтепродуктов в Украине 25 июня 2013 г. отмечается снижение розничных цен на сжиженный газ. Сжиженный нефтяной газ СПБГ, используемый в качестве моторного топлива, за 25.06.2013 г. на АГЗС в Украине подешевел на 0,46% (2,4 коп./л) до 5,34 грн/л.
 Сюжет полностью (8)

2013.06.25 10:29 Цены на автомобильное топливо в Крыму на 25 июня 2013 г. *События Крыма*
 8
 2013.06.27 16:42 Сжиженный газ за неделю подешевел на 1,5% *Терминал*

3. **Цены на бензин и дизтопливо в Киеве 25 июня незначительно изменились**
 РКБ-Украина Розничные цены на бензин и дизельное топливо 25 июня 2013 г. по сравнению с предыдущим торговым днем незначительно изменились. Об этом свидетельствуют данные мониторинга ценового департамента "Консалтинговой группы А-95".
 Бензин А-80 А-92 А-95 А-95+ ДТ ЦРО КЮ 10,33 10,90 10,83 11,23 10,08 5,09 (+0,10) ВОС 10,49 10,76 11,29 12,39 10,29 5,29 (+0,20) Лукойл 10,30 10,89 11,14 12,99 10,19 5,00 ТНК н/л 10,59 10,99 11,99 10,08 Украина 10,10 10,40 10,70
 Сюжет полностью (5)

2013.06.25 13:25 Цены на бензин и дизтопливо в Киеве 25 июня не изменились *РКБ-Украина*
 5
 2013.06.27 13:55 Цены на бензин и дизтопливо в Киеве 27 июня незначительно изменились *РКБ-Украина*

Figure 7 – Fragment of the chain of main plots

Цены на топливо на 27.06.2013

По данным консалтинговой группы "А-95", средние на бензин и дизельное топливо на АЗС в Днепропетровске, грн./л на 27 июня 2013 года:

Компания	A-80	A-92	A-95	ДТ (Л-02,62)
Укрнафта	10,10	10,40	10,70	9,80
Веста Сервис	10,10	10,34	10,59	9,59
Формула Ритейл		10,59	10,94	9,99
Лукойл		10,69	11,14	10,19
Нефтек	10,29	10,59	10,99	9,99
Альфа-Нафта		10,35	10,65	9,80
Средняя по области	10,16	10,49	10,84	9,89

Средние цены на топливо по областям Украины:

Область	A-80	A-92	A-95	ДТ (Л-02,62)
АР Крым	10,24	10,58	10,98	9,99
Винницкая	10,25	10,61	11,00	10,02
Волынская	10,23	10,54	10,90	9,95
Днепропетровская	10,24	10,45	10,77	9,85
Донецкая	10,25	10,70	11,16	10,14
Житомирская	10,29	10,49	10,85	9,95
Закарпатская	10,23	10,54	10,91	9,86
Запорожская	10,20	10,47	10,82	9,89
Ивано-Франковская	10,23	10,61	11,01	10,05
Киев	10,23	10,60	10,98	10,04
Кировградская	10,24	10,53	10,90	9,96
Луганская	10,25	10,62	10,94	10,00
Львовская	10,23	10,60	11,00	10,05

Figure 8 – Document with price information

2.3. Text Mining, Information Extraction

The task that must be constantly solved when conducting competitive intelligence is the automatic extraction of concepts and facts from formalized arrays of information (tables, databases) and unstructured texts presented in the web space, the identification of deep connections between individual concepts. For this, it is supposed to use in competitive intelligence systems Knowledge discovery technologies, the concept of deep analysis of data and texts (Data Mining, Text Mining).

An important task of the Text Mining technology is related to the extraction of its characteristic elements or properties from the text, which can be used as document metadata, keywords, and annotations. Another task is to assign a document to certain categories from a predefined classification scheme. Text Mining also provides a new level of semantic document search.

According to the current methodology, the main elements of Text Mining include [Lande et al., 2009]: classification (Classification), clustering (Clustering), building semantic networks, ex-

tracting facts, concepts (Feature Extraction), summarizing (Summarization), responses to queries (Question Answering), thematic indexing (Thematic Indexing) and search by keywords (Keyword Searching). Also, in some cases, this set is supplemented by means of supporting and creating taxonomy (Taxonomies), thesauri (Thesauri), and ontologies (Ontology).

Text classification uses statistical correlations to create rules for placing documents into specific categories. The classification problem is a classical recognition problem, where, according to some control sample, the system assigns a new object to one category or another. The peculiarity of the Text Mining concept lies in the fact that the number of objects and their attributes can be very large – the use of intelligent mechanisms for optimizing the classification process is envisaged.

Clustering is based on the features of documents, the use of linguistic and mathematical methods without the use of predefined categories. The result of clustering can be a taxonomy or a visual map that provides efficient coverage of large amounts of data. Clustering in Text Mining is considered as a process of selecting compact subgroups of objects with similar properties. Clustering tools allow you to find features and divide objects into subgroups based on these features. Clustering usually precedes classification because it allows you to define groups of objects.

When constructing semantic networks, it is supposed to analyze the relationships between concepts extracted from documents. Concepts correspond to the appearance of certain descriptors (key phrases) in documents. Links between concepts can be established in the simplest case by taking into account the statistics of their joint mention in various documents.

Extraction or extraction of facts (concepts) is intended to obtain some facts from the text in order to improve classification, search, clustering and building semantic networks.

Automatic Text Summarization [Khan, 2000] is the compilation of summaries of materials, annotations or digests, i.e. extracting the most important information from one or more documents and generating concise, understandable and informative reports based on them.

Based on the methods of automatic referencing, it is possible to form search images of documents. According to automatically constructed annotations of large texts – search images of docu-

ments – a search can be carried out, characterized by high accuracy (naturally, due to completeness). In some cases, instead of searching the full texts of an array of large documents, it turns out to be appropriate to search an array of specially crafted annotations. Although the search images of documents often turn out to be formations that only remotely resemble the source text, which is not always perceived by a person, but due to the inclusion of the most significant keywords and phrases, they help to lead to quite adequate results when conducting a full-text search.

The unique features of the concept and technologies of Text Mining is that they can be used to extract from "raw" data non-obvious, practical and accessible for interpretation knowledge necessary for making decisions in various fields of activity, including in the field of economic competition..

On the modern market, there are a number of both Western products and production systems of post-Soviet countries that are capable of carrying out deep analysis of texts to one extent or another.

Recently, all major Western brands specializing in the development of information storages and databases, corporate management systems have expanded their product lines with Text Mining systems or modules. The availability of such modules is announced by SAP, Oracle, SAS, IBM and other companies.

The process of competitive intelligence can be viewed as building a network of objects under study and links between them. The results should provide analytical information that can be used to make decisions. Analytical information can be presented in the form of visual diagrams – semantic networks, digests, sets of storylines, relationships of key concepts, companies, individuals, technologies, etc.

The tasks of competitive intelligence have created a demand for special information technologies that provide the ability to extract and process the necessary information, which in turn has caused a flood of system proposals from software developers.

Today, public and special programs and services help to solve the problems of competitive intelligence based on information from the Internet, for example, the so-called “personalized intelligence portals” that can select information on the nar-

rowest, specific issues and topics and provide it to customers have recently gained popularity.

At present, technologies and systems of “computer competitive intelligence” have been declared, the idea of which is to automate and accelerate the processes of extracting information necessary for competitive struggle from open sources and its analytical processing.

In the conduct of competitive intelligence, new areas of science and technology are increasingly being used, which have received the names: “knowledge management” (Knowledge Management) and “knowledge discovery in databases” (Knowledge Discovery in Databases) or otherwise, Data and Text Mining – “deep analysis of data or texts.”

If knowledge management systems implement the idea of collecting and accumulating all available information, both from internal and external sources, then Data and Text Mining, as already shown, allow you to identify non-obvious patterns in data or texts – the so-called latent (hidden) knowledge. In general, these technologies are still defined as the process of discovering previously unknown but useful knowledge in raw data that is necessary for decision making. Systems of this class make it possible to analyze large arrays of documents and form subject indexes of concepts and topics covered in these documents.

A characteristic task of competitive intelligence, usually included in Text Mining systems, is finding exceptions, that is, searching for objects that stand out from the crowd with their characteristics.

Another class of important tasks solved within the framework of Text Mining technology is data modeling, situational and scenario analysis, and forecasting [Lande, Furashev, 2012].

Visualization is of great importance for processing and interpreting the results of Text Mining. Often the head of the company does not always adequately perceive the analytical information offered to him, especially if it does not fully coincide with his understanding of the situation. In this regard, the competitive intelligence service should strive to present information in a form adapted to the individual perception of the customer.

Visualization is usually used as a means of presenting the content of the entire array of documents, as well as to implement

navigation through semantic networks in the study of both individual documents and their classes.

2.4. Domain Models

An important task of competitive intelligence is to identify non-obvious patterns and relationships from the texts of web pages and identify their relationships, build matrices and relationship graphs.

Existing factual databases of structured information are not always available to the expert analyst. For the rapid determination of facts and entities, modeling of information links between them, the most promising approach is to take into account the information, knowledge that is contained in unstructured text documents, in particular, on the Internet.

Today, when almost all interested users have already accumulated a lot of experience with traditional information retrieval systems, it turned out to be obvious that the facts or concepts that are searched for using such systems are often meaningless in themselves. For example, if a user is interested in the information links of Oshchadbank with other banks or individuals, then he does not know which banks or names to indicate to him in the request, and it is physically impossible to indicate all documents containing the word "Oshchadbank". In such cases, information connections, the number of which goes beyond the statistical background, as a rule, reflect reality.

It is usually not the concepts or facts themselves that are interpreted, but the relationships between them. It is not so much the study of the concepts themselves that is important, but the study of their relationship. It is known that it is the relationship that contributes to the understanding of motivational-target features, that is, the user is not interested in the concept in itself, but in the environment, in order to immediately have an idea about the subject area, if necessary, direct a clarifying search in the right direction. Similar solutions implemented in the form of "information portraits" containing key words are used in systems such as InfoStream (infostream. ua), CyberAggregator.

The database of almost any traditional information retrieval system can be considered as a graph, the vertices of which are objects – terms, concepts, descriptors, etc., and the edges – their connections. At the same time, the basis of the search in

these cases is the search for vertices, that is, the search for objects. Search by relationships, edges, seems at first glance less efficient. Indeed, if we assume that there are N vertices in the graph, then the number of edges can theoretically be $N(N - 1) / 2$, that is, if we assume that there are only 100 thousand vertices, then there may be about 5 billion edges, which corresponds to a fairly large database, even by modern standards. However, if we use such concepts as names of people and names of companies from news documents as graph vertices, then it turns out that the corresponding incidence matrix turns out to be very sparse. Measurements showed that with the number of individual concepts extracted from 5 million news documents equal to approximately $N = 1.5$ million, the number of links was only $\nu = 4$ million.

In addition, as experiments have shown, the distribution of degrees of vertices (the degree of a vertex – the number of edges emanating from it) in such graphs is power-law, which indicates the so-called scale-free, that is, that many characteristics (in particular, the ratio of the number of vertices and ribs) should remain at the same level. Therefore, it is technically possible to use the edges of the considered graph as the basis for constructing a database of links – links between individual concepts.

Data from content monitoring systems such as InfoStream, Webscan or Yandex.News, as well as the results of monitoring specialized web services, such as databases of biographies of people, organizations, services, can be used as arrays of documentary information for such a system. employment, etc.

Information relationships between concepts are identified by processing documentary arrays and can be stored in a special database. The set of concepts used in building a database of links is formed by extracting data from a text array accessible to the user, which gives the system integrity.

In a corporate information infrastructure, the link database can be used in various ways, for example, standalone, or its capabilities can be supplemented by the capabilities of existing full-text and/or factual databases (Fig. 9). At the same time, the main result of the work is the construction of the so-called “link maps”, and as a side effect that implements the “proof mode”, the extraction of the documents themselves as sources of links can be considered.

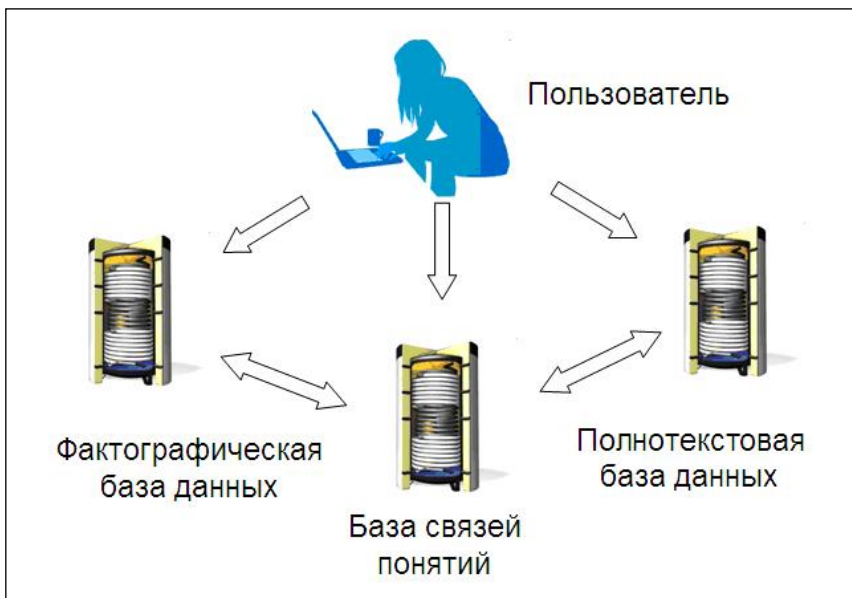


Figure 9 – Location of the concept links database in information infrastructure

When designing databases of links, promising solutions are used in the field of creating information and analytical systems, in particular, the theory and technologies of deep analysis of texts – Text Mining, including methods of extracting information (Information Extraction), technologies of databases of extra-large volumes (Big Data), the concept of "complex networks" (Complex Networks).

Within the framework of the theory of complex networks, characteristics associated with the topology of networks are studied, as well as statistical phenomena, the distribution of weights of individual vertices (which can be considered entities, concepts, facts) and edges, the effects of leakage and conduction in networks, etc.

On Fig. 10 schematically shows the possible technological steps in the formation of a database of links [Lande, Braychevsky, 2010].

With the help of the robot program, the selected web resources containing information related to the objects of research are scanned.

After that, the concepts necessary for users are extracted, for example, brand names, companies, email addresses, etc.

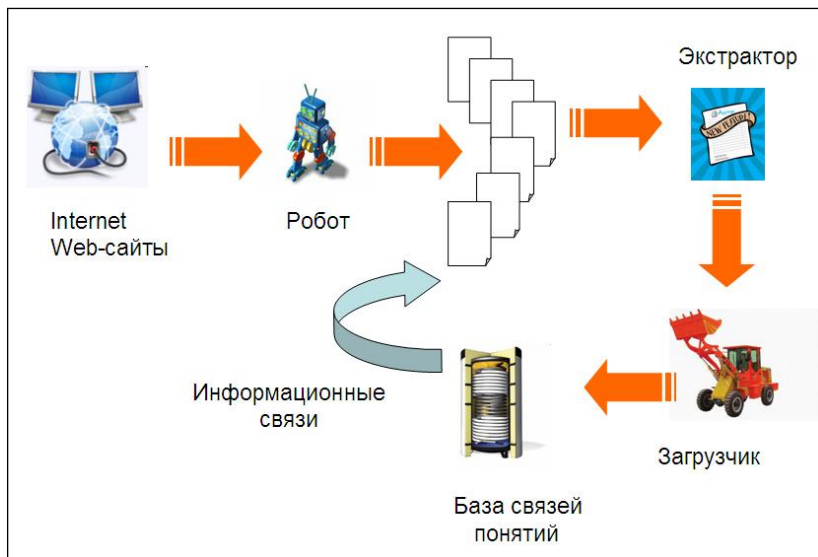


Figure 10 – Scheme of forming a database of links

The selected concepts and the corresponding relationships between them are loaded into a database of links, which also contains links to primary source documents. Concept extraction tools, as a rule, are focused on processing documents scanned from the Internet, presented in various languages.

The proposed approach to the search, of course, entails some features in the implementation of the concept link database architecture. Currently, the most popular platform for such a database is the Neo 4j graph DBMS. In addition, the architecture of the link database should be focused on such possible applications as the detection of implicit links (not explicitly revealed by the concept extraction complex), the search for individual objects, as well as the relationship with existing factual databases.

There are several systems in which this approach is partially implemented:

PolyAnalyst (www.megaputer.ru) – allows you to solve problems of forecasting, classification, grouping of objects, conduct analysis of relationships, multidimensional analysis and interactive reporting. The PolyAnalyst system (and its component – the TextAnalyst system) provides linguistic and semantic text analysis, entity detection, link visualization, document systematization, summarization and query processing in natural language;

SAP Businessobjects Text Analysis (<https://www.sap.com/sapbusinessobjects>) is a program that allows you to extract information about dozens of types of objects and events, including people, place names (place names), companies, dates, amounts of money, email addresses, and identify links between them;

Neticle Text Analysis (<https://neticle.com/textanalysisapi/>) is a technology for extracting information from unstructured texts. It allows you to identify the information contained in unstructured text and turn it into structured data that has relationships that can be analyzed.

A variant of such a system is currently implemented and used as a component of the X- SCIF competitive intelligence system of the Ukrainian company Information Corporate Service, which allows the user to receive maps of links for selected objects online and helps interpret the results. It is assumed that the user enters an object as a request. The request is sent to the database of links, from where the corresponding fragments are selected – link maps (the level of detail and time retrospective must be specified parametrically).

After identifying relevant objects and links, procedures for their automatic grouping (clustering) and visualization are performed, the results are presented to the user in the form of link maps, which are presented in the form of dynamic (most often, Java diagrams) link graphs.

In particular, in the X- SCIF competitive intelligence system, a link graph is built using Java applets and is a graphical object that contains nodes and edges. Each link graph element has a

context menu, which is an additional control element in the user interface (Fig. 11).

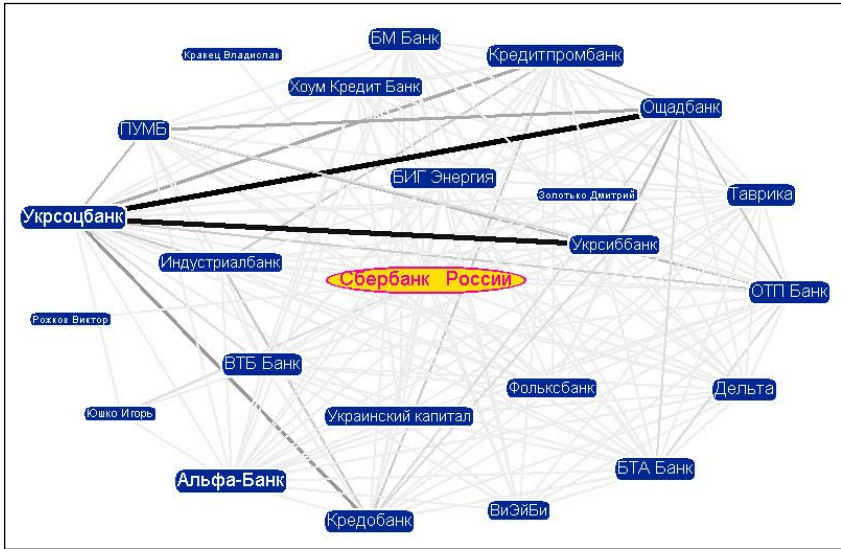


Figure 11 – Graph of information links of the concept "Sberbank of Russia"

Objects that have more connections are displayed with a larger font. Edges corresponding to more connections are shown as darker lines. The constructed network has its own controls: zooming (using the "zoom" menu or the scroll bar at the top of the screen); move the entire graph; moving an object; configuration change; highlighting the links of the selected node, etc.

On fig. 11 shows an example of using a database of links, the case when the user is interested in information links of Sberbank of Russia as of 2011. Of course, for the corresponding query, many different relationships can be identified, but there is a simple and reliable criterion for ranking the results, which consists in cutting off the statistical background. In the case under consideration, by asking the appropriate query, you can get a graph of objects (persons and companies) most related to Sberbank of Russia. And if finding the names of the bank's managers (the chairman of the board, the first deputy chairman of the board and the head of the subsidiary bank) is a fairly obvious

result, then the connections between individual banks made it possible to reveal (after referring to the primary source documents) facts that are not obvious at first glance, for example, that UkrSibbank and UkrSotsbank were partner banks.

The presented approach can be considered as the basis for the construction of the so-called "vertical" (subject-oriented) information retrieval systems, in which the issues of efficiency and screening out information noise are initially resolved. The implementation under consideration has the property of scaling in three parameters: the volume of databases, the composition of the concepts that are used, and the infrastructure environment.

Analyzing connections in a network, one can determine many non-obvious properties, for example, identify the presence of clusters, determine their composition, differences in connectivity within and between clusters, identify key elements that connect clusters to each other, etc. A serious obstacle in the analysis is the incompleteness of information about the connections between individual network nodes. At the same time, algorithms already exist today, with the help of which it becomes possible to restore the missing fragments of links with a high probability. Even without a complete description of the information network, it is possible to obtain a representative sample of "real" connections and complete the entire network using it. The presented approach implements a link between full text and factual databases.

2.5. Big Concept Data

2.5.1. Big data concept

The term Big Data appeared as a new term and logo in an editorial by Clifford Lynch, editor of Nature on September 3, 2008, who devoted an entire special issue of one of the most famous journals to the topic "what big data sets can mean for modern science". At present, this term has already taken root and has reached the peak of its use. Here the word "large" was associated not so much with some quantity, but with a qualitative assessment. Time has confirmed the validity of singling out big data as a separate phenomenon. Today, according to research by the Gartner agency, the term Big Data has already surpassed the peak of Gartner's famous Hype Cycle. On Fig. Figure 12 shows

the statistics of asking users to the Google system for the phrase “Big Data” (Google Trends service, <https://trends.google.com/>).

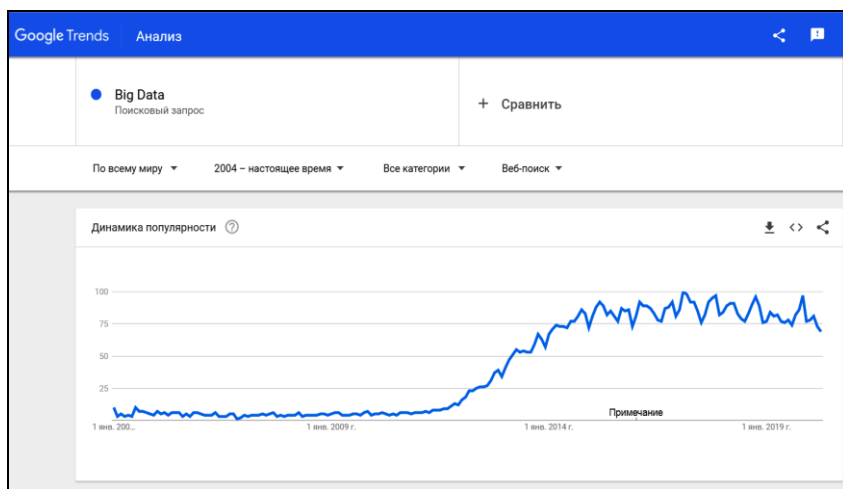


Figure 1 2 – Dynamics of requests “ Big data ”

In 2012, an article [Boyd, 2012] by Dana Boyd and Kat Crawford formulated the definition of Big Data as a cultural, technological and scientific phenomenon that includes: 1) Technology: maximizing the computational power and complexity of algorithms for collecting, analyzing, linking and comparing huge datasets. 2) Analysis: depiction of huge datasets to identify patterns for making economic, social, technical and legal claims. 3) Mythology: the general belief that huge datasets represent a higher form of knowledge and information that can generate insights that were previously impossible and with a halo of fidelity, objectivity and accuracy.

According to this definition, big data is a term that refers to the multitude of data sets so large and complex that it makes it impossible to use existing traditional database management tools and applications to process them. The problem is the collection, cleaning, storage, search, access, transfer, analysis and visualization of such sets as a holistic entity, rather than local fragments. As defining characteristics for big data, “three Vs” are noted: volume (English volume, in the sense of the size of the

physical volume), speed (English Velocity, meaning in this context the growth rate and the need for high-speed processing and obtaining results), diversity (English.variety, in the sense of being able to process different types of structured and semi-structured data at the same time). The leading characteristic here is the amount of data that must be considered in terms of applications.

Why has data volume become a problem? As computers got faster and more memory, so did the amount of data. In fact, the growth of data has even outpaced the growth of computer speed, and few algorithms scale linearly with the growth of input data. In short, data is growing faster than our ability to process it. Thus, the amount of data is growing faster than processing power. A number of consequences follow from this.

- Some methods and techniques that worked well in the past now need to be revised or replaced because they do not scale to the current amount of data.
- Algorithms cannot assume that all input data will fit in RAM.
- Data management in itself becomes a non-trivial task.
- The use of clusters or multi-core processors is becoming a necessity, not a luxury.

Modernity shows us examples of the monstrous size of the digitized data generated today. According to the giants of the IT industry (EMC, Cisco, IBM, Google), in 2012, 2 zetabytes ($2 * 1021$) or 2 thousand exabytes or 2 thousand billion gigabytes of information were generated in the world, and in 2020 this value will reach 35 zetabytes. The sources of this avalanche of data are numerous digital devices that concentrate and direct the products of the human mind into the bottomless expanses of the Internet – tweets, Facebook and VKontakte posts, queries to search engines, etc., as well as data from sensors and controllers of millions of devices, which measure temperature and humidity, the condition of roads and air conditioners, and much more, which today is united by the term “Internet of things” IOT (Internet of things). However, it is not only the problem of quantity that prevents this – the volume of data is the first “ V ”. For big data, as already noted, the second “ V ” is important – speed. The results of big data processing should be obtained in the time determined by the problem being solved with their help. This will

make it possible to turn big data analytics from a tool that answers the question “who is to blame?”, characteristic of traditional analytics systems, to a tool for obtaining answers “what to do?”. The analyst in this case turns from a pathologist into a therapist. The speed of access to data, the speed of their processing is an important criterion for the quality of technologies included in big data.

Finally, the third " V " – the diversity of data suggests that big data must be efficiently processed regardless of their structure. Here it is customary to single out three main types of data according to the degree of their structuredness.

The first level is the usual structured data, which can be represented by separable and predefined fields that contain bits that have different semantics. For example, all tables have headers in a certain field of a given length, one of the facts in another predefined field, and another of the facts in another field that determine the numeric or textual values of the semantic variables contained in the headers.

Structured data is well stored in relational databases and it is convenient to manage such data using a special SQL language – Structured Query Language. Despite their prevalence, such data is determined only in 10% of the total volume of generated data.

The second level is semi-structured data. Data of this type have structural separators, but cannot be presented in the form of a table due to the lack of some attributes for different data. An example of such data can be files in the SGML format – Standard Generalized Markup Language or BibTex in which there is no specific data storage scheme, but the semantic meaning of various data elements can be determined by analyzing the file itself. Sometimes such data is defined as self-describing. Many data stored on the Web are semi-structured data, bibliographic descriptions of publications, scientific data.

Finally, unstructured data, which, by definition, cannot fit the previously described types. This includes text in multiple languages, sound recordings, still images, video files, e-mail messages, tweets, presentations, and other business information outside of database uploads. It is estimated that 80 to 90 percent of all data in organizations is unstructured data. Often, the semi-structured data entered above is also referred to as unstruc-

tured. Sometimes the diversity scale is extended by using the whole scale from structured data to completely unstructured data. We will assume that the data variance index is zero for completely unstructured data and increases to one for well-structured relational databases. According to the participants of the World Economic Forum 2012 in Davos, those who ride the topic of big data mining will become the masters of the information space. This topic was the focus of a special report at the Big Data, Big Impact Forum. The key conclusion of the report is that digital assets are becoming no less significant economic asset than gold or currency. Research conducted by Professor E. Brynjolfsson and two colleagues in 2012 showed that big data analysis and forecasting is being adopted by corporate America. They studied 179 large companies and found that those who adopted big data mining in the last year and a half received an immediate improvement in economic performance by 5-6%. At present, the needs of society make it necessary for the emergence of big data specialists in the form of a separate profession. The name of this profession Data Scientist is a data scientist. In the United States, they assessed the needs for specialists in this profession and came to the conclusion that already in 2018 there will be a shortage of data scientists in the United States in the amount of 190,000 people! The well-known Harvard Business Review titled one of its issues as follows: “Data Scientist: The Sexiest Job Of the 21st Century” – the data scientist is the most attractive job of the 21st century.

2.5.2. Big data techniques

First, we list those functional operations on data, methods for storing and processing them. Of course, this list does not exhaust the whole variety of dynamically developing techniques, but it allows you to see what can be done with big data to achieve the goals of the researcher.

- Consolidation of data;
- Classification, clustering;
- Machine learning ;
- Visualization.

Data Consolidation

This whole set of techniques is aimed at extracting data from different sources, ensuring its quality, converting it into a single format and loading it into a data warehouse – an “analytical sandbox” (analytic sandbox) or “data lake” (data lake). Data consolidation techniques differ in the type of analytics performed by the system:

- Batch analytics (batch oriented);
- Real time analytics (real time oriented);
- Hybrid analytics (hybrid).

With batch analytics, data is periodically uploaded from various sources, the data is analyzed for the presence of faulty fragments, noise, and they are filtered. When you run real-time analytics, data is continuously produced by sources and forms a set of data streams. Analyzing these streams and getting results in a timely manner at a given pace requires that data be received asynchronously in the form of some messages and route these messages to the correct processing nodes for processing. For hybrid analytics, as a rule, data messages should not only be routed for processing, but also integrated into an analytical sandbox for further processing based on the results of data accumulation over significant time intervals. The data resulting from the consolidation must meet certain quality criteria. Data quality is a criterion that determines the completeness, accuracy, relevance and interpretability of data. Data can be of high or low quality. High quality data is complete, accurate, up-to-date data that can be interpreted. Such data provides a qualitative result: knowledge that can support the decision-making process. The set of processes that define consolidation is called ETL – Extraction-Transformation-Loading (Extraction-Transformation-Loading). In business intelligence applications, ETL processes included very complex data transformations, such as quantization, which allows to reduce the amount of data processed, normalization – the process of bringing relational tables to a canonical form or numerical data to a single scale, data encoding – the introduction of unique codes for data compression. In big data techniques, it is usually believed that it is necessary to work directly with dirty data, since it is often the nature of failures that can be the subject of analysis, and data compression is a function of the analytic algorithms themselves. The possibility of storing data in its original form should be provided by the technical

means of the analytical system. The quality of big data is often difficult to evaluate using formal algorithms, and then visualization is resorted to at an early stage of the study. In addition to assessing the quality and choosing a preprocessing method, visualization can help move on to an important stage of analytics – the choice of models, hypotheses to achieve the ultimate goal – decision making.

Visualization

The visualization technique is a powerful data mining technique. Typically, it is used to review and verify data before creating a model, and also after generating predictions. Visualization is the transformation of numerical data into some visual image in order to simplify the perception of large amounts of information. Visualizers are used for visualization. Renderers can be either a standalone application or a plugin or part of another application. The possibilities of visualizers are very wide. At present, they can present information in almost every conceivable form, as long as the analyst can formulate what he wants to see.

Text visualization

If the data are texts in natural language, then visualization using marked-up text can provide primary assistance in the analysis. The visualizer counts the frequency of mentions of a particular word, and assigns a conditional weight to the words depending on this frequency. Words of different weights have different markup when rendered, which means different representations on the screen. Some words look bigger than others. This type of visualization helps the researcher grasp the main ideas of the text very quickly.

Cluster visualization

One commonly used visualization is the cluster visualization. Clusters are groups of objects that are somewhat similar or similar in properties. Clustering algorithms, i.e. the division of a set of objects into groups, we will consider below, but here we will only show how their work can be visualized. Most visualizers support clustering algorithms and are able to divide data into clusters. Typically, contrasting colors are used to visually represent clusters for objects from different clusters.

Association Visualization

Association visualization shows the frequency with which certain items appear together in a data set, thereby determining the structure of data organization (for example, it can be about which products are often sold together). It is also possible to visualize information about the strength of the data association.

Visualization of hypotheses

Hypothesis visualization allows you to show the identified patterns that confirm the hypotheses put forward. The presentation of information in different visualizers is different. For example, if the rows of the 3D pie charts represent the features used by the classifier, then each pie chart represents the probability that the feature value or range of values is appropriate for the classification. Figure 2.10 below analyzes the wages of the US working population. The visualizer reflects the attributes that can affect salary classification. Attributes are represented by rows of 3D pie charts. The height of the pie chart (cylinder) shows the number of records in a given category ; the color indicates that the salary is greater than or less than \$50,000. There can be multiple pie charts for each attribute, for example, there are two pie charts for gender (male/female), and eight pie charts for age. Their number depends on the number of patterns identified by the visualizer.

Visualizing Decision Trees

Visualization of decision trees allows you to present hierarchically organized information in the form of a landscape and view all or part of the data set in the form of nodes and branches. The landscape can be either two-dimensional or three-dimensional. The quantitative and relational characteristics of the data are made visible using hierarchically connected nodes.

Classification

The classification technique is one of the basic big data mining techniques. It is often used when building a model of analytical systems along with another technique – clustering. Classification is the distribution of objects (observations, events) of research into previously known classes based on the similarity of features. Unlike classification, clustering distributes objects

(observations, events) according to previously unknown classes. The classification is made in accordance with the principles of supervised machine learning (Supervised Machine Learning). To carry out classification using mathematical methods, it is necessary to have a formal description of the object that can be operated using the mathematical apparatus of classification. Each object (database record) must contain information about some features of the object.

The classification process usually consists of the following steps.

1. The set of initial data (or data sample) is divided into two sets: training and test. The training set is the set that contains the data used to construct the model. The set contains the input and output (target) values of the examples. The output values are for training the model. The test set also contains the input and output values of the examples. Here the output values are used to validate the model.

2. Each dataset object belongs to one predefined class. At this stage, the training set is used, and the model is constructed on it. The resulting model is represented by classification rules, a decision tree or mathematical formulas.

3. The correctness of the model is evaluated. Known values from the test set are compared with the results of using the resulting model. The level of accuracy is calculated – the percentage of correctly classified objects in the test set.

Clustering

The clustering technique is an approach to data classification in the case when it is not known in advance to which class any of the available objects should be assigned. Clustering is carried out by automatically finding groups into which the analyzed objects should be divided. Such a process can be considered as machine learning without a teacher (Unsupervised Machine Learning). More than 100 different algorithms are known

Machine learning

The term "machine learning" is likely to come across to you more than once. Although it is often used as a synonym for artificial intelligence, in fact, machine learning is one of its elements.

At the same time, both concepts were born at the Massachusetts Institute of Technology in the late 1950s.

Machine learning (ML) is a class of artificial intelligence methods, the characteristic feature of which is not the direct solution of a problem, but learning in the process of applying solutions to many similar problems. To build such methods, the means of mathematical statistics, numerical methods, optimization methods, probability theory, graph theory, various techniques for working with data in digital form are used.

There are two types of training:

Case learning, or inductive learning, is based on discovering empirical patterns in data.

Deductive learning involves the formalization of expert knowledge and its transfer to a computer in the form of a knowledge base.

Deductive learning is usually referred to the field of expert systems, so the terms machine learning and case learning can be considered synonymous.

Many inductive learning methods have been developed as an alternative to classical statistical approaches. Many methods are closely related to information extraction (information extraction, information retrieval), data mining.

Unlike traditional software, which is great at executing instructions but not capable of improvisation, machine learning systems essentially program themselves, developing instructions on their own by summarizing known information.

A classic example is pattern recognition. Show the machine learning enough pictures of dogs labeled "dog" and cats, trees, and other objects labeled "not a dog" and it will begin to recognize dogs well over time. And for this she will not need to explain exactly how they look.

Teaching with and without a teacher

This type of machine learning is called supervised learning. This means that someone introduced the algorithm to a huge amount of training data, reviewing the results and adjusting the settings until the desired classification accuracy was achieved on data that the system had not yet "seen". It's like hitting the "not spam" button in your email program when a filter accidentally

intercepts the message you want. The more often you do this, the more accurate the filter becomes.

Typical supervised learning tasks are classification and prediction (or regression analysis). Spam and pattern recognition are classification problems, and stock price prediction is a classic example of regression.

In unsupervised learning, the system scans huge amounts of data, remembering what "normal" data looks like to be able to recognize anomalies and hidden patterns. Unsupervised learning is useful when you don't know exactly what you are looking for, in which case the system can be forced to help you.

Unsupervised learning systems can detect patterns in vast amounts of data much faster than humans. That is why banks use them to detect fraudulent transactions, marketers to identify customers with similar attributes, and security software to recognize malicious activity on the network.

Examples of unsupervised learning problems are clustering and searching for association rules. The first is used, in particular, for customer segmentation, and the mechanisms for issuing recommendations are based on the search for association rules.

Machine learning methods

The section of machine learning, on the one hand, was formed as a result of the division of the science of neural networks into network training methods and types of topologies of their architecture, on the other hand, it absorbed the methods of mathematical statistics. The machine learning methods below are based on the case of using neural networks, although there are other methods that use the concept of a training sample – for example, discriminant analysis, which operates on the generalized variance and covariance of the observed statistics, or Bayesian classifiers. Basic types of neural networks, such as perceptron and multilayer perceptron (as well as their modifications), can be trained both with a teacher and without a teacher, with reinforcement and self-organization. But some neural networks and most statistical methods can be attributed to only one of the learning methods. Therefore, if you need to classify machine learning methods depending on the method of learning, it would be incorrect to attribute neural networks to a certain type,

it would be more correct to type learning algorithms for neural networks.

- Supervised learning – for each use case, a pair of "situation, required solution" is set.
- Unsupervised learning – for each use case, only a "situation" is specified, it is required to group objects into clusters using data on pairwise similarity of objects, and / or reduce the data dimension.
- Active learning is different in that the learning algorithm has the ability to independently assign the next situation under study, on which the correct answer will become known.
- Training with partial involvement of a teacher (semi-supervised learning) – for a part of the precedents, a pair of "situation, required solution" is set, and for a part – only a "situation".
- Transductive learning is learning with the partial involvement of a teacher, when the prediction is supposed to be made only for precedents from the test set.
- Multitask learning (multi-task learning) – simultaneous learning of a group of interrelated tasks, for each of which its own pairs of "situation, required solution" are set.
- Multivariate learning (multiple-instance learning) – learning when cases can be combined into groups, each of which has a "situation" for all cases, but only for one of them (and it is not known which one) there is a pair of "situation, required solution »
- Boosting (boosting – improvement) is a procedure for sequential construction of the composition of machine learning algorithms, when each subsequent algorithm seeks to compensate for the shortcomings of the composition of all previous algorithms.
- Bayesian network.

Machine learning algorithms need data, as much data as possible from as wide a range of sources as possible. The more they "eat" this data, the "smarter" they become and the greater their potential in decision making. And the clouds provide that big data.

Big data promises to provide us with a lot of value in the digital transformation process, while the cloud offers the building

blocks for this process. Machine learning, in turn, was the first truly industrial tool to master these new values at scale. The beauty of machine learning is that its uses are almost limitless. It can be applied anywhere fast data analysis is important, and it can be revolutionary where it is important to identify trends or anomalies in large datasets, from clinical research to safety and compliance enforcement.

Limitations of machine learning

Each machine learning system creates its own link diagram, representing a kind of black box. You will not be able to figure out exactly how the classification is performed by engineering analysis, but this does not matter, the main thing is that it works.

However, a machine learning system is only as good as the training data is accurate: if you feed it “garbage” as input, then the result will be appropriate. If the training is incorrect or the size of the training sample is too small, the algorithm may produce incorrect results.

2.5.3. Big data technologies and tools

We will consider the basic technologies and tools that are most widely used in well-known projects today. This list does not exhaust all the technologies that have already been tested, let alone those in development, but it allows you to get a fairly holistic view of “what” data scientists use today and what tools you need to master in order to deploy a project using big data.

Big data technologies should provide solutions and tools to implement the techniques described above on large amounts of heterogeneous data at the required speed. This is achieved by high parallelization of computations and distributed data storage. Despite the need for significant computing power and memory, as a rule, the deployment of big data software products is carried out on clusters of computers of medium or even low class (commodity computers). This allows big data systems to scale without incurring significant costs. Recently, cloud computing services have been increasingly used to deploy big data systems. If the system is implemented in the cloud, the nodes of the computing cluster are implemented on virtual machines of the cloud infrastructure and flexibly adapt to the task, reducing the

cost of use. This serves as an additional factor that attracts many developers to build big data systems on cloud platforms.

The most popular big data technology, which is considered the de facto standard for building analytics systems that work in batch mode, is a set of solutions and software libraries, united under the name Hadoop. If big data arrives in the form of high-speed streams and system response must occur with low latency, then real-time analytics is used instead of batch analytics. Here, de facto standard approaches have not yet arisen, and from the most popular ones, we will consider a technology called Storm.

Apache Hadoop

Under the name Hadoop, the Apache community promotes a technology based on the use of a special infrastructure for parallel processing of large amounts of data. Hadoop provides an environment for functional task programming, automatic parallelization of work, shifting computational load to data. Hadoop was created by Doug Cutting, the creator of Apache Lucene, a widely used text search library. Hadoop is derived from Apache Nutch, an open source web search engine that itself was part of the Lucene project.

The Nutch project was launched in 2002. A workable crawler and search engine appeared very quickly. However, the developers realized that their architecture would not scale to billions of web pages. Help came in 2003, when an article was published describing the architecture of GFS (Google File System) – a distributed file system that was used in real Google2 projects [Ghemawat, 2003].

In 2004, an article was published in which Google introduced the MapReduce technology to the world [Dean, 2004]. In early 2005, the Nutch developers had a working Nutch-based MapReduce implementation, and by mid-year, all of Nutch's core algorithms had been adapted to use MapReduce and NDFS. Nutch's NDFS and MapReduce capabilities went well beyond search, and in February 2006 an independent sub-project of Lucene was formed called Hadoop. Around the same time, Doug Cutting joined Yahoo!, which provided the team and resources to turn Hadoop into a web-scale system (see the sidebar “Hadoop at Yahoo!” below). The results were shown in February 2008 when

Yahoo! announced that the search index it uses was generated by a 10,000-core Hadoop cluster [Yahoo, 2008].

The history of Hadoop is directly related to the development of the Google File System (2003) and then the implementation of MapReduce technology (2004). Based on these components, the Apache Nutch information search application was born in 2005, which gave way to the Apache Hadoop project the following year.

While Hadoop is most commonly associated with MapReduce and the Distributed File System (HDFS, formerly NDFS), the term often refers to a whole family of interrelated projects that are brought together by a distributed computing infrastructure and large-scale data processing. All of the underlying projects covered in this book are maintained by the Apache Software Foundation, which provides support to the open source community—including the original HTTP server from which the title is derived. As the Hadoop ecosystem expands, new projects appear, not necessarily managed by Apache, but providing additional functionality to Hadoop or building higher-level abstractions around core functionality.

42 Chapter 1: Introducing Hadoop

The following is a summary of the Hadoop projects covered in the book. Common is a set of components and interfaces for distributed file systems and general I/O (serialization, Java RPC, data structures). Avro is a serialization system for performing efficient cross-language RPC calls and long-term data storage. MapReduce is a distributed processing model and runtime that runs on large clusters of typical machines. HDFS is a distributed file system that runs on large clusters of standard machines. Pig is a data flow control language and runtime for analyzing very large datasets. Pig runs on HDFS and MapReduce clusters. Hive is a distributed data warehouse. Hive manages the data stored in HDFS and provides an SQL-based query language (which is converted by the run-time engine into MapReduce jobs) to work with that data. HBase is a distributed column-oriented database. HBase uses HDFS to organize data storage and supports both batch calculations using MapReduce and point queries (random data reading). ZooKeeper is a highly available distributed coordination service. ZooKeeper provides primitives that can be used to build distributed applications (for example, distributed locks). Sqoop is an efficient bulk data transfer tool between structured storage (such as relational databases) and HDFS. Oozie – Ha-

doop job startup and scheduling service (including MapReduce, Pig, Hive and Sqoop jobs)

Hadoop consists of four functional parts:

- Hadoop Common;
- Hadoop HDFS;
- Hadoop MapReduce;
- Hadoop YARN.

Hadoop Common is a set of libraries and utilities necessary for the normal functioning of the technology. It includes a specialized simplified command line interpreter.

When a data set outgrows the capacity of a single physical machine, it must be spread across multiple different machines. The file systems that manage the storage of data on a network are called distributed file systems. Since they operate in a networked environment, the designer must take into account all the complexities of network programming, so distributed file systems are more complicated than conventional disk file systems. For example, one of the biggest challenges is getting the file system to survive individual node failures without losing data. Hadoop comes with a distributed file system called HDFS (Hadoop Distributed Filesystem). Sometimes – in old documentation or configurations, or in informal communication – the abbreviation "DFS" is also found; it means the same. HDFS is the main Hadoop file system covered in this chapter, but Hadoop also implements a generalized file system abstraction, and we'll look at integrating Hadoop with other storage systems (such as a local file system and Amazon S3) along the way.

HDFS (Hadoop Distributed File System) is a distributed file system for storing data on many machines in large volumes. Designed to provide:

- Reliable data storage at low cost
- unreliable equipment ;
- High read-write throughput;
- Streaming access to data;
- Simplified consistency model;
- Architecture similar to Google File System.

The HDFS file system is designed to store very large files with a streaming data access scheme in clusters of conventional machines [Shvachko, 2010]. Let's consider this statement in more detail. Very large files "Very large" in this context refers to

files that are hundreds of megabytes, gigabytes, and tera bytes in size. There are now Hadoop clusters that store petabytes of data [Scaling, 2008].

Streaming Data Access HDFS is based on the concept of write-once/read-multiple access as the most efficient data processing scheme. A data set is usually generated or copied from a source, after which various analytical operations are performed on it. Each operation involves most of the dataset (or the entire dataset), so the time to read the entire dataset is more important than the delay in reading the first record. Conventional Hadoop hardware does not require expensive, highly reliable hardware. The system is designed to run on standard hardware (publicly available hardware that can be purchased from many companies)³ with a fairly high probability of failure of individual nodes in a cluster (at least for large clusters).

HDFS technology is designed so that in the event of a failure, the system continues to operate without any noticeable interruption. You should also highlight areas of application for which HDFS is currently not well suited (although this may change in the future): Fast data access Applications that require access to data with minimal latency (in the range of tens of milliseconds) do not fit well with HDFS. Recall that the HDFS system is optimized to provide high data throughput, for which you have to pay with slower access. HBase (Chapter 13) is currently better suited for providing low latency data access.

Numerous Small Files Because the name node stores file system metadata in memory, the limit on the number of files in the file system is determined by the memory size of the name node. Experience shows that each file, directory and block takes up about 150 bytes. Thus, for example, if you have a million files, each of which occupies one block, you will need at least 300 MB of memory to store information. Storing millions of files is still acceptable, but billions of files are already beyond the capabilities of today's hardware¹. Multiple Write Sources, Arbitrary File Modifications Writing to HDFS files can only be done by one source. Writing is always done to the end of the file. There is no support for multiple recording sources or modifications with an arbitrary offset in the file. (Maybe these features will be supported in the future, but they are likely to be relatively inefficient.)

The architecture of HDFS is based on storage nodes – servers of a standard architecture, on the internal disks of which data is stored. All data uses a single address space. This provides parallel input-output of information from different nodes. Thus, a high throughput of the system is guaranteed.

HDFS operates on two levels: namespaces (Namespace) and storage of data blocks (Block Storage Service). The namespace is maintained by a central name node (Namenode) that stores file system metadata and file block allocation metadata.

Numerous Datanodes directly store files. The name node is responsible for handling file system operations—opening and closing files, manipulating directories, and so on. Data nodes process data writing and reading operations. The name node and data nodes are provided with web servers that display the current status and allow browsing the contents of the file system.

HDFS is not POSIX compliant. Unix commands `ls`, `cp`, etc. do not work. Mounting HDFS on Linux OS requires special tools such as HDFS-Fuse. Files are distributed block by block between nodes. All blocks in HDFS (except the last block of the file) have the same size – from 64 to 256 MB.

To ensure server failure tolerance, each block can be duplicated on multiple nodes. The replication factor (the number of nodes on which each block should be placed) is defined in the file settings. Files can only be written to HDFS once (modification is not supported), and only one process can write to a file at a time. In this simple way, data consistency is implemented.

Hadoop MapReduce

MapReduce is a data-centric programming model. This model is simple, but not so simple that useful programs cannot be implemented in its context. Hadoop allows you to run MapReduce programs written in different languages; in this chapter, we'll look at the same program written in Java, Ruby, Python, and C++. But the most important thing is that MapReduce programs are inherently parallel, and therefore, large-scale data analysis becomes available to anyone who has enough computers at their disposal. The advantages of MapReduce are fully manifested when working with large data sets, so let's start by looking at one of these sets.

Hadoop MapReduce is the most popular software implementation of the model parallel processing of large amounts of data by dividing into independent tasks solved by the Map and Reduce functions. The MapReduce algorithm takes 3 arguments as input: the original data collection, the Map function, the Reduce function, and returns the resulting data collection.

The initial data collections are sets of records of a special type. This is a data structure of the Key, Value (KEY, VALUE) type. The user needs to set the Map and Reduce processing functions. The algorithm itself takes care of sorting the data, running processing functions, re-executing failed transactions, and much more. The resulting collection consists of the results of the analysis in an easily interpretable form. The work of the MapReduce algorithm consists of three main stages: Map, Group and Reduce. As a first step, the Map function is executed on each element of the original collection. As a rule, it takes one record of the form (KEY, VALUE) as input, and returns a number of new records (KEY1, VALUE1), (KEY2, VALUE2),..., i.e. converts the input {key:value} pair to a set of intermediate pairs. This function also plays the role of a filter – if no intermediate values need to be returned for a given pair, the function returns an empty list.

We can say that it is the responsibility of the Map function to convert the elements of the original collection into zero or more instances of {key:value} objects.

In the second stage (Group), the algorithm sorts all {key:value} pairs and creates new object instances grouped by key. The grouping operation is performed inside the MapReduce algorithm and is not specified by the user. The reduce function returns instances of the {key: reduced value} object that are included in the resulting collection.

For example, consider a simplified version of the task facing search engines. Let's say we have a database of pages on the web and we want how many times each page is linked. Let there be a first.com page with links to first.com, second.com, third.com, a second.com page with two links to first.com, and a third.com page with no links at all.

To have a uniform format of the initial data collection, let's define the form of each saved page as (KEY = URL, VALUE = TEXT). The results are easy to interpret.

Java is used as the base language for writing functions. For programming, there is a popular Hadoop plugin in Eclipse. But you can do without it: Hadoop streaming utilities allow you to use as Map and Reduce any executable file that works with the standard input / output of the operating system (for example, UNIX shell utilities, Python scripts, Ruby, etc.), there are also SWIG-compliant Hadoop pipes API in C++. In addition, Hadoop distributions include implementations of various handlers most commonly used in distributed processing.

A feature of Hadoop is to move the calculations as close to the data as possible. Therefore, user tasks are run on the node that contains the data to be processed. At the end of the Map phase, intermediate lists of data are moved for processing by the Reduce function. Note here that in addition to Hadoop, there are different implementations of MapReduce. MapReduce was originally implemented by Google. Later, other implementations of the algorithm appeared. MapReduce from Google has become an open source project – MySpace Qizmt – MySpace's Open Source Mapreduce Framework. Another known version of the algorithm is the one implemented in the MongoDB system.

Hadoop YARN (Yet Another Resource Negotiator) is a system resource management platform responsible for distributing server computing resources and scheduling user tasks.

Early versions of Hadoop MapReduce included the JobTracker job scheduler, since version 2.0 (2013) this feature has been moved to YARN. In it, the Hadoop MapReduce module is implemented on top of YARN. The programming interfaces are mostly preserved, but there is no complete backward compatibility.

YARN is sometimes referred to as a cluster operating system. This is due to the fact that the platform manages the interface between hardware resources and various applications that use computing power. The basis of YARN is a logically independent daemon – the resource scheduler (ResourceManager), which abstracts all the computing resources of the cluster and manages their provision to distributed processing applications. Reporting to him are numerous Node Managers responsible for monitoring the current status and load of individual servers.

Both MapReduce programs and any other distributed applications that support the appropriate programming interfaces

can work under the control of YARN. YARN provides the ability to run several different tasks in parallel within a server system.

A distributed application developer needs to implement a special application management class (AppMaster) that is responsible for coordinating tasks within the resources provided by the resource scheduler. The resource scheduler is responsible for creating instances of the application control class and interacting with them through the network protocol.

A number of data processing products have been built on top of Hadoop. Here is a list of just the most popular ones:

- Pig is a high-level data flow language for parallel programming;
- HBase is a distributed database that provides storage for large tables;
- Cassandra is a robust, decentralized database;
- Hive is a data warehouse with data aggregation and fast search functions;
- Mahout is a library of machine learning and knowledge extraction methods.

Hadoop is a very dynamic technology. Therefore, it is recommended to get the latest information on the Internet at <http://hadoop.apache.org/>.

Storm – stream processing system

Storm is a free technology and software implementation of a distributed real-time computing system [15]. This system allows you to build reliable processing of unlimited data streams, similar to how Hadoop does it with batch processing. Storm is used for real-time analytics, online machine learning, continuous computing, distributed ETL, and other big data streaming operations.

Storm can integrate with queuing and database technologies already in use and is language independent. The basis of Storm are Storm topologies and Storm cluster. The cluster is an object similar to the Hadoop cluster, and instead of running a MapReduce job, Storm topologies are run here. Jobs and Topologies have a key difference – the former normally complete their work, while the latter always process messages. Storm cluster has two types of nodes master node and worker nodes (Figure

2.28). The master node runs a daemon called Nimbus, which is similar to Hadoop's JobTracker. Nimbus is responsible for distributing code across cluster worker nodes, distributing tasks across machines, and starting and stopping worker processes. Each worker process executes a subset of the topology. A running topology consists of many worker processes distributed over many machines. Each worker node has a daemon called Supervisor. This module listens to all processes on its machine and starts and stops them on Nimbus's initiative. Coordination between Nimbus and all Supervisors is done through a special cluster called Zookeeper. This cluster also stores the state of all processes on its disk space, which allows you to recover from a failure separately any machine of the working cluster. To perform real-time calculations on Storm, you need to create a topology (topologies) – a graph of calculations. Each node in the topology contains processing logic and a link between nodes showing how data should be transferred between nodes.

The main abstraction in Storm is a stream. A stream is an unlimited sequence of tuples. Sources of data streams for processing are represented in the topology by an abstraction called spout, and stream handlers that can perform functions, filter streams, aggregate or combine data streams, interact with databases are called bolt.

Elastic stack

Over the past few years, various systems have appeared for storing and processing large amounts of data. Among them are Hadoop ecosystem projects, some NoSQL databases (DB), as well as search and analytical systems like Elasticsearch. Hadoop and any NoSQL database have their own benefits and uses.

Elastic Stack is a vast ecosystem of components that are used to search and process data. Main the Elastic Stack components are Kibana, Logstash, Beats, X-Pack, and Elasticsearch. The core of the Elastic Stack is the Elasticsearch search engine, which provides opportunities for storing, searching and processing data. Kibana, also known as Elastic Stack Window, is an excellent visualization tool and user interface for Elastic Stack. The Logstash and Beats components allow you to push data to the Elastic Stack. X-Pack provides powerful functionality: you can set up monitoring, add various notifications, set security

options to prepare your system for operation. Since Elasticsearch is the core of the Elastic Stack

Elasticsearch is a highly scalable real-time distributed full-text search and data analysis search engine. The utility allows you to store, search and analyze large amounts of data. Typically used as a base engine/technology to assist applications with complex search functions. Elasticsearch is the core component of the Elastic Stack.

Elasticsearch, at the heart of the Elastic Stack, plays a major role in data discovery and analysis. It is built on a unique technology – Apache Lucene. This makes Elasticsearch fundamentally different from traditional relational database or NoSQL solutions. The following are the main benefits of using Elasticsearch as a data store:

- unstructured, document-oriented;
- search capability;
- possibility of data analysis;
- support for custom libraries and REST API;
- easy management and scaling;
- work in pseudo-real time;
- high speed of work;
- resilience to errors and failures.

Overview of Elastic Stack Components

Some components are universal and can be used without Elastic Stack or other tools.

Elasticsearch

Elasticsearch stores all your data, provides search and analysis capabilities in a scalable way. We have already looked at the benefits and reasons for using Elasticsearch. You can work with Elasticsearch without any other components to equip your application with data mining and analysis tools.

To work with relational databases, you need to understand concepts such as rows, columns, tables, and schemas. Elasticsearch and other document-oriented stores work differently. The Elasticsearch system has a clear focus on documents. JSON documents are best suited for it. They are organized using different types and indexes. Next, we will look at the key concepts of Elasticsearch:

- index;
- type;
- document;
- cluster;
- node;
- shards and copies;
- markup and data types;
- reverse index.

An index is a container that stores and manages documents of the same type in Elasticsearch. An index can contain documents of the same type,

Indexes in Elasticsearch are roughly similar in structure to a database in relational databases. Continuing the analogy, the type in Elasticsearch corresponds to the table, and the document corresponds to the records in it.

Type

Types help logically group or organize documents of the same type into indexes.

Typically, documents with the most common set of fields are grouped under one type. Elasticsearch does not require a structure, allowing you to store any JSON document with any set of fields under a single type. In practice, you should avoid mixing different details in the same type, such as "customers" and "products". It makes sense to store them in different types and with different indexes.

Document

As already mentioned, JSON documents are best suited for use in Elasticsearch. A document consists of several fields and is the basic unit of information stored in Elasticsearch. For example, you might have a document that corresponds to one product, one customer, or one order item.

Documents contain several fields. In JSON documents, each field has a specific type. In the product catalog example we saw earlier, there were fields sku, title, description, price, etc. Each field and its value can be seen as a key/value pair in the document, where the key is the field name and the value is field value.

Knot

Elasticsearch is a distributed system. It consists of many processes running on different devices on the network and interacting with other processes. In Chapter 1, we downloaded, installed, and launched Elasticsearch. Thus, we launched the so-called single node of the Elasticsearch cluster.

An Elasticsearch node is a single system server that can be part of a large node cluster. It is involved in indexing, searching, and performing other operations supported by Elasticsearch. Each Elasticsearch node is assigned a unique ID and name at startup.

Each Elasticsearch node has a main configuration file that is located in the settings subdirectory. The file format is YML (full name is YAML Ain't Markup Language). You can use this file to change the default values such as hostname, ports, cluster name.

At a basic level, a node corresponds to a single running Elasticsearch process. It is responsible for managing the piece of data that corresponds to it.

Cluster

A cluster contains one or more indexes and is responsible for performing operations such as searches, indexing, and aggregations. A cluster is formed by one or more nodes. Any Elasticsearch node is always part of a cluster, even if it is a single node cluster. By default, each node attempts to join a cluster named Elasticsearch. If you run multiple nodes within the same network without changing the `cluster.name` parameter in the `config/elasticsearch.yml` file, they are automatically clustered.

The cluster consists of several nodes, each of which is responsible for storing and managing its part of the data. One cluster can store one or more indexes. An index logically groups different types of documents.

Shards and copies

Shards help distribute the index across the cluster. They distribute documents from the same index to different nodes. The amount of information that can be stored in one node is lim-

ited by the disk space, RAM, and computing capabilities of that node. Shards help to distribute the data of one index throughout the cluster and thereby optimize the resources of the cluster.

The process of dividing data into shards is called sharding. This is an integral part of Elasticsearch and is required for scalable and parallel optimization work:

- disk space for different nodes of the cluster;
- computing power for different nodes of the cluster.

Distributed systems like Elasticsearch are designed to work even when the hardware fails. To do this, replicas of shards, or copies, are provided. Each index shard can be configured to have some or no copies. Shard replicas are additional copies of the original or primary shard to provide a high level of data availability.

Markup and data types

Elasticsearch is an unstructured system, so it can store documents with any number of fields and field types. In reality, data is never completely unstructured. There is always a set of fields common to all documents of this type. In fact, types within indexes should be created based on common fields. Typically, one type of document within an index contains several common fields.

Data types

Elasticsearch supports a wide range of data types for various storage scenarios for text data, numbers, booleans, binary objects, arrays, objects, nested types, geopoints, geoshapes, and many other specialized data types such as IPv4 and IPv6 addresses. In a document, each field has an associated data type.

Logstash

The Logstash utility helps to centralize event-related data, such as information from log files (logs), various indicators (metrics), or any other data in any format. It can perform data processing before forming the sample you need. It is the key component of the Elastic Stack and is used to collect and process your data containers.

Logstash is a server side component. Its purpose is to collect data from a wide variety of input sources in a scalable fash-

ion, process the information, and send it to its destination. By default, the converted information goes to Elasticsearch, but you can choose from many other output options. Logstash's architecture is plugin-based and easily extensible. Three types of plugins are supported: input, filtering and output.

Kibana

Kibana is a visualization tool for the Elastic Stack that helps you visualize data in Elasticsearch. It is also often referred to as a window in the Elastic Stack. Kibana offers many visualization options such as histogram, map, line graphs, time series, and more. You can create visualizations with just a couple of mouse clicks and explore your data interactively. In addition, it is possible to create beautiful dashboards consisting of various visualizations, share them, and also receive high-quality reports.

Kibana also provides management and development tools. You can manage X-Pack security settings on Elastic Stack, and use developer tools to create and test REST API requests.

Kibana Console is a user-friendly editor that supports auto-completion and query formatting as you write them.

What is REST API? REST stands for Representational State Transfer. It is an architectural style for systems to interact with each other. REST has evolved along with the HTTP protocol, and almost all REST-based systems use HTTP as their protocol. HTTP supports various methods: GET, POST, PUT, DELETE, HEAD, etc. For example, GET is for getting or looking for something, POST is used for creating a new resource, PUT can be used to create or update an existing resource, and DELETE is for permanent deletion.

Elastic Cloud

Elastic Cloud is a cloud-based Elastic Stack component management service provided by Elastic (<https://www.elastic.co/>), the author and developer of Elasticsearch and other Elastic Stack components. All product components (apart from X-Pack and Elastic Cloud) are based on open source code. Elastic maintains all components of the Elastic Stack, provides training, development, and cloud services.

Besides Elastic Cloud, there are other cloud solutions available for Elasticsearch such as Amazon Web Services (AWS).

The main advantage of Elastic Cloud is that it is created and maintained by the authors of Elasticsearch and other components of the Elastic Stack.

As you can see, Elasticsearch and Elastic Stack can be used for a wide range of tasks. Elastic Stack is a platform with an advanced set of tools for building end-to-end search and analytics solutions. It is suitable for developers, architects, business analysts and system administrators. It is entirely possible to build an Elastic Stack solution with little to no coding, just configuration changes. At the same time, the Elasticsearch system is very flexible, which means that developers and programmers can build powerful applications thanks to the extensive support for programming languages and REST APIs.

Sphinxsearch

Another full-text search engine for big data is Sphinx 's earch.

Sphinx search (from *SQL Phrase Index*) is distributed under the GNU GPL or, for versions 3.0+, without source codes. A distinctive feature is the high speed of indexing and searching, as well as integration with existing DBMS (MySQL, PostgreSQL) and APIs for common web programming languages (PHP, Python, Java are officially supported ; there are community-implemented APIs for Perl, Ruby,.NET and C ++).

The official website of the system is <http://sphinxsearch.com/>.

Sphinx search system has the following features:

- High indexing speed (up to 10-15 MB/s per processor core);
- High search speed (up to 150-250 queries per second per processor core with 1,000,000 documents);
- Great scalability (the largest known cluster indexes up to 3,000,000,000 documents and supports more than 50 million queries per day);
- Distributed search support;
- Support for multiple full-text search fields in a document (up to 32 by default);
- Support for several additional attributes per document (i.e. groups, timestamps, etc.);

- Support for single-byte encodings and UTF-8;
- Support for morphological search – there are built-in modules for English, Russian and Czech languages; modules available for French, Spanish, Portuguese, Italian, Romanian, German, Dutch, Swedish, Norwegian, Danish, Finnish, Hungarian;
- Native support for existing PostgreSQL and MySQL DBMS, support for ODBC compatible databases (MS SQL, Oracle, etc.).

In 2017, the Manticore Software team forked Sphinxsearch 2.3.2, which they called Manticore Search. According to the developers, the management of the Sphinxsearch system could not cope with the maintenance of the system, namely, the detected errors were not corrected, the announced features were not implemented, and the dialogue between users and Sphinxsearch developers was hindered. Sphinx version 3 can already be perceived as a proprietary solution for a limited circle of users. In fact, the new version of the system (Manticore Search) solved many of these problems, including providing support for the code as a whole, organizing modern interaction with Sphinxsearch and Manticore users, implementing such features inherent in, in particular, Elasticsearch: replication, auto id, JSON interface, the ability to create / delete an index on the fly, the availability of a document repository, developed real-time indexes.

Neo4j

Neo4j is an open source graph database management system in the Java language with support for transactions (ACID). As of 2015, it is considered the most common graph DBMS [Robinson, 2016] . The developer is the American company Neo Technology, which has been under development since 2003.

The data is stored in its own format, specially adapted for the presentation of graph information ; this approach, in comparison with the modeling of a graph database by means of a relational DBMS, allows additional optimization in the case of data with a more complex structure. It is also stated that there are special optimizations for SSD drives, while processing the graph does not require its entire placement in the RAM of the

computing node, thus processing sufficiently large graphs is possible.

Key Applications: Social media, recommendation systems, fraud detection, mapping systems.

Graph database terminology

- graph database, graph database – a database built on graphs – nodes and connections between them
- Cypher – a language for writing queries to the Neo4j database (like SQL in MySQL)
- node, node – an object in the database, a graph node. The number of nodes is limited to 2 to the power of 35 ~ 34 billion
- node label, node label – used as a conditional "node type". For example, movie type nodes can be associated with actor type nodes. Node labels are case-sensitive, and *Cypher does not throw errors if you type the name in the wrong case.
- relation, connection – a connection between two nodes, an edge of a graph. The number of connections is limited to 2 to the power of 35 ~ 34 billion
- relation identifier, the type of relationship is in Neo4j for relationships. Maximum number of link types 32767
- properties, node properties – a set of data that can be assigned to a node. For example, if the node is a product, then in the node properties you can store the product id from the MySQL database
- node ID, node ID — unique identifier of the node. By default, this ID is displayed when viewing the result.

Saving data in Neo4j

The nodestore.db file contains a certain size of entries containing information about the node:

1. A label that indicates the entry is active;
2. Pointer to the first relation that this node contains;
3. Pointer to the first property that this node contains.

The node does not contain its own identifier. Since each entry in nodestore.db takes up the same amount of space, a node pointer can be calculated.

The relationshipstore.db file also contains entries of the same size that describe relationships, but they consist of the following elements:

1. A label that indicates the entry is active;
2. Pointer to the node that contains this relation;
3. Pointer to the node to which this relation is directed;
4. Type of relationship;
5. Pointer to the relation that comes in front (within this node);
6. Pointer to the relation that stands behind (within this node);
7. A pointer to the relation that is in front (within the Node to which this relation is directed);
8. Pointer to the relation that stands behind (within the Node in which this relation is directed);
9. A pointer to the first property of this relation.

How the data model is chosen is a directed property graph:

- Contains nodes (nodes) and connections (relationships).
- Nodes have properties. Nodes can be thought of as documents containing properties in the form of key-value pairs.
- Nodes can be labeled with one or more labels. Labels group nodes by indicating the role they play in the dataset.

Multiple labels can be assigned to the same node (because nodes can play several different roles in different domains). Links connect nodes and structure a graph. Links are named (always have the same name) and directed (always have a direction, start node, and end node). Links can also contain properties. This allows you to enter additional metadata in the algorithms graph, add additional semantics to links, and limit queries in real time.

The main transactional features are ACID support and compliance with JTA, JTS and XA specifications. The DBMS application programming interface has been implemented for many programming languages, including Java, Python, Clojure, Ruby, PHP, and a REST-style API has also been implemented. You can extend the programming interface both with the help of server-side plugins and with the help of unmanaged extensions (*unmanaged extensions*); Plugins can add new resources to the

REST interface for end users, and extensions allow you to take full control of the API, and can contain arbitrary code, so they should be used with care.

The DBMS uses Cypher, a declarative graph query language. The syntax of this language is similar to that of SQL. Operations for creating, selecting, updating, deleting data are supported. Cypher describes graphs using a pattern specification – a simple form of ASCII graphics is used, the user draws the part of the graph he is interested in using ASCII characters; vertices are taken in brackets, their labels are written after ":"; to create several nodes, they should be listed through ","; links are indicated by arrows (\rightarrow and \leftarrow), and the names of links are indicated inside square brackets after ":"; properties of nodes and links (key-value pairs) are written in curly braces.

The Cypher query language is the most common query language for graph databases, due to its use in the Neo4j DBMS. Cypher is a declarative language and allows you to create, update, and delete vertices, edges, labels, and properties, as well as manage indexes and constraints. To retrieve data from the storage, a query is used that contains a filtering template that allows you to get:

- $(n) \rightarrow(m)$ are all directed edges from vertex n to vertex m ;
- $(n:Person)$ – all vertices labeled Person;
- $(n:Person:Russian)$ — all vertices with both Person and Russian labels;
- $(n:Person \{name:\{value\}\})$ — all vertices labeled Person and filtered by an additional property;
- $(n:Person) \rightarrow(m)$ — edges between vertices n labeled Person and m ;
- $(n)-(m)$ — all undirected edges between vertices n and m .

Requests can be made in Neo4j in other ways, for example, directly through the Java API and in the Gremlin[en] language created in the TinkerPop open source project. Cypher is not only a query language, but also a data manipulation language, as it provides CRUD functions for graph storage.

Gephi

Gephi is graph (network) analysis and visualization software package. Gephi (<https://gephi.org/>) is currently the most popular network and graph visualization and analysis program ("network graphs"). Gephi provides fast layout, efficient filtering, and interactive data exploration, and is one of the best options for visualizing large-scale networks. Gephi is a multi- platform software that is distributed open source under the CDDL 1.0 and GNU General Public License v3. Mac OS X, Windows, and Linux versions of the source codes are available at <https://gephi.org/>.

Gephi is actively used in a number of academic research projects, in particular sociological ones; also quickly gained popularity among journalists. Now its user environment has expanded significantly – with this package you can deal with any topic of network analysis. Gephi was used, among other things, to visualize the global connectivity of New York Times content and study Twitter network traffic during times of social unrest; Gephi inspired the creation of LinkedIn InMaps and was used to visualize the entire Truthy network.

Gephi allows you to process a structure graph of sufficiently large volumes (up to 1 million nodes) on a personal computer due to efficient algorithms.

Gephi developers describe this program as "like Photoshop, but for data".

The program includes many different layout algorithms (concluding graphs on a plane) and allows you to customize colors, sizes and labels in graphs. Gephi is interactive software and provides a means to identify communities, as well as the ability to calculate shortest paths or relative distance from any node to a given node. Plugins from Gephi allow you to expand its functionality and add new algorithms, layouts and measurement tools. Gephi has a multi-threaded data processing scheme and thus allows you to perform several types of analysis at the same time.

The user interface of the Gephi system includes three main sections (windows) :

- "Data Lab": all initial data about the network is stored here, as well as additional calculated values;
- "Data processing": most of the user's operations take place here, in particular, manual editing of networks, testing layouts, setting filters;

- "Preview": here the form of the output of the graph is specified, as a rule, with the help of a set of tools, the graph is finalized, including from an aesthetic point of view. In the same window, a call to export the graph to PDF, PNG and SVG formats is implemented.

The program includes many different layout algorithms (concluding graphs on a plane) and allows you to customize colors, sizes and labels in graphs. Gephi is interactive software and provides a means to identify communities, as well as the ability to calculate shortest paths or relative distance from any node to a given node. Plugins from Gephi allow you to expand its functionality and add new algorithms, layouts and measurement tools. Gephi has a multi-threaded data processing scheme and thus allows you to perform several types of analysis at the same time.

These three main sections cover many tabs that allow the user to implement individual functions. Each of the main and secondary windows – sections and tabs – is discussed below.

When analyzing large and dense networks, fast layout (ordering of graph nodes) is a bottleneck, since most complex layout algorithms are demanding on processor, memory, and runtime parameters. At the same time, Gephi comes with efficient layout algorithms such as Yifan-Hu, Force-directed. In particular, the Yifan-Hu algorithm is ideal for applying after other, faster and coarser algorithms. While most of Gephi's proposed methods can be completed within a reasonable amount of time, a combination of, for example, OpenOrd and Yifan-Hu gives the best visual representations. Of course, the correct parameterization of any layout algorithm can affect both the operation and the result of rendering.

Gephi allows you to download network data in GEXF, GDF, GML, GraphML, Pajek (NET), GraphViz (DOT), CSV, UCINET (DL), Tulip (TPL), Netdraw (VNA) and Excel spreadsheet formats. In addition, Gephi allows you to export network data in JSON, CSV, Pajek (NET), GUESS (GDF), Gephi (GEFX), GML and GraphML formats. This allows Gephi to interact with other graph analysis and visualization systems.

2.6. Mathematical Foundations

This chapter focuses on the recognition of information operations based on the study of the dynamic properties of information flows in global computer networks, in particular, on the Internet.

To study information flows on the Internet, i.e. the flow of messages that are published on the pages of websites, social networks, blogs, etc., modern tools should be used. The well-known methods of generalizing information arrays (classification, phase enlargement, cluster analysis, etc.) are no longer always suitable even for an adequate quantitative reflection of the processes occurring in the information space [Lande, 2007].

A quantitative analysis of the dynamics of information flows that are generated on the Internet is becoming one of the most informative methods for studying the relevance of certain thematic areas today. This dynamic is driven by a variety of qualitative factors, many of which defy precise description. However, the general nature of the time dependence of the number of thematic publications on the Internet still allows the construction of mathematical models, their study, and forecasting. Observations of the time dependences of the volumes of network information flows convincingly indicate that the mechanisms of their generation and distribution are obviously associated with complex non-linear processes. It is to this topic that this chapter is devoted.

For the analysis of time series, which reflect the dependence of the volume of information flows on time, a variety of methods and approaches are used. It turns out that all these approaches are interrelated and, moreover, the concept of correlation plays a key role. The presentation is built around the scheme shown in Fig. 13, with particular attention paid to relationships.



Figure 13 – Relationships between approaches to time series analysis

2.6.1. Time series

A time series is a set of observed values sorted by time. Further, discrete time series will be considered, the values of which were fixed at regular intervals. We will designate such a time series x_1, x_2, \dots, x_T or briefly $\{x_t\}_{t=1}^T$ implying that the values of the series were fixed after an equal time interval h : $t_0, t_0 + h, t_0 + 2h, \dots, t_0 + (T - 1)h$.

If the values of the time series are uniquely specified by some mathematical relation (such as, for example, $x_t = A \cdot \sin(vt)$), then such a series is deterministic. If the values of a time series can only be described in terms of a probability distribution, then we are talking about a statistical time series. Such series will be considered further. Analyzing time series, we will consider them as a realization of a stochastic process.

As examples, three time series will be used below, which were obtained using the popular GoogleTrends network service. These time series show the level of interest in Donald Trump, Hillary Clinton and the "Russian hackers" from August 2016 to April 2017. Time series obtained using GoogleTrends show the dynamics of the popularity of a search query. The maximum point on the chart is equal to 100 and corresponds to the date when the query was most popular, and the remaining points on the chart are determined as a percentage of the maximum. All three time

series are shown in Fig. 14. For ease of reference to these series, we will further denote them as T (D. Trump), K (H. Clinton), X (“Russian hackers”).

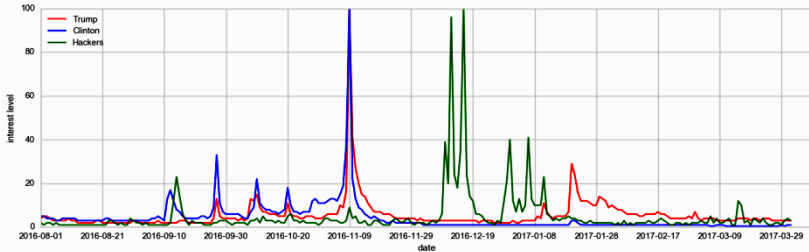


Figure 14 – Time series showing interest in Donald Trump (T), Hillary Clinton (K) and "Russian hackers" (X) from August 1, 2016 to April 1, 2017 from GoogleTrends.

In some cases it is useful to consider a smoother version of the original time series. Smoothing helps to reveal significant trends in the dynamics of the series, while hiding the noise and various features that appear at small scales. There are various smoothing methods. The simplest way to smooth is to calculate a moving average. A simple moving average is equal to the arithmetic mean of the elements of a series from an interval of a given length, namely

$$SMA_t = \frac{1}{w} \sum_{i=0}^{w-1} x_{t-i},$$

where w is the width of the smoothing interval (the number of elements over which the average is calculated), SMA_t is the value of the simple moving average at the point t . The resulting value SMA_t refers to the middle of the smoothing interval, so the smoothed series y_t can be defined as $y_t = SMA_{t+\lfloor \frac{w}{2} \rfloor}$.

When using moving average smoothing, the larger the width of the smoothing interval, the smoother the function will be. On Fig. 15 shows how the smoothed T series looks like when the value is increased w .

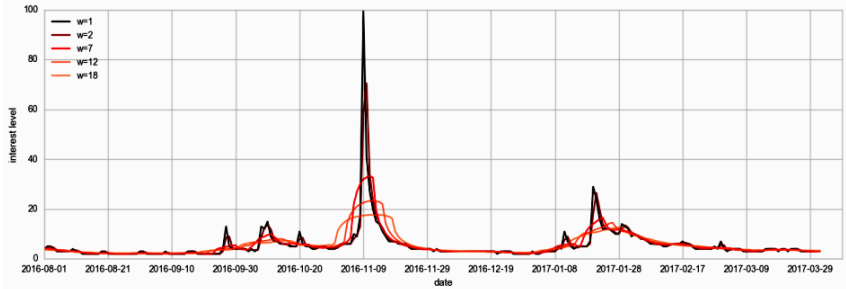


Figure 15 – Initial series T and smoothed by a simple moving average with a smoothing interval width of 2, 7, 12, 18.

The results of series smoothing can be demonstrated on a graph, in which the abscissa axis corresponds to the time axis, and the width of the smoothing interval is plotted along the ordinate axis. The graph shows the values $y_t^{(w)}$ – that is, elements of a smoothed row at a point t when using a width interval w (Fig. 16).

When calculating a simple moving average, all points that fall within the smoothing interval have the same weight. Naturally, unequal weights can be used. Thus, we come to the definition of a weighted moving average

$$WMA_t = \frac{1}{w} \sum_{i=0}^{w-1} a_i x_{t-i},$$

Where $\sum_{i=0}^{w-1} a_i = 1$.

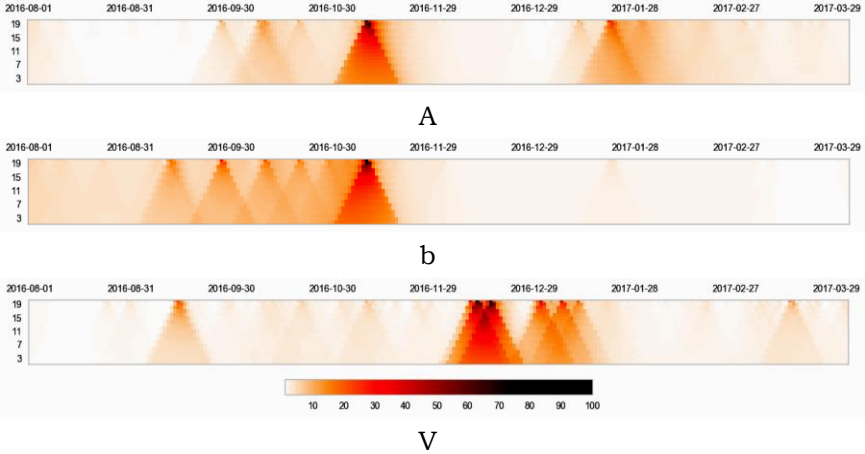


Figure 16 – Values of the time series T (a), K (b) and X (c) smoothed by a simple moving average, depending on the width of the smoothing interval. Time is plotted along the abscissa axis, and the width of the interval is plotted along the ordinate axis.

Another commonly used series smoothing method is **exponential smoothing**. The previous values of the series are taken into account with exponentially decreasing weights. We will denote the elements of the smoothed series by y_t , and immediately define $y_0 = x_0$. The next elements of the series y_t are obtained by the recursive formula

$$y_t = \alpha x_t + (1 - \alpha)y_{t-1},$$

where $0 < \alpha < 1$ is the smoothing coefficient. It is obvious that for $\alpha = 1$ the resulting series y_t coincides with the original one x_t . Thus, if the value α is close to 1, then the highest weight in the determination y_t is assigned to the corresponding x_t , and the background of the series "means little". On the other hand, if it were α equal to 0, then the entire series y_t would be smoothed to one value $y_t = y_0$. That is, when α close to 0, the history of the series is taken into account with a greater weight than the current value.

On Fig. 17 shows the series T, as well as the corresponding smoothed series for various values of the parameter α .

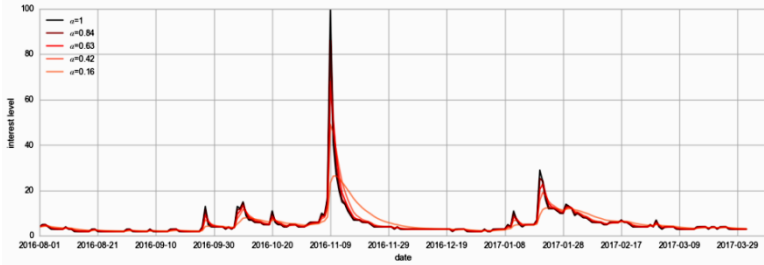


Figure 17 – The original series T and smoothed using exponential smoothing with the parameter, equal to 0.84, 0.63, 0.42, 0.16

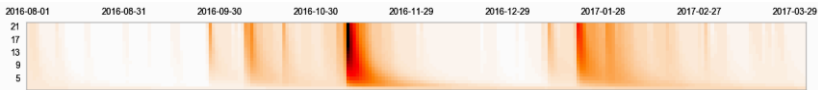
As in the case of a simple moving average, we will demonstrate the results of smoothing the series on the chart. In this case, we plot the parameter along the ordinate axis α (Fig. 18). The graph shows the values $y_t^{(\alpha)}$ is the value at the point of the initial series t smoothed with the parameter α .

As examples, we consider the time series T, K, and X, which have a weekly periodicity. This is a characteristic property of many processes in the information space. It is known that the publication of news messages often occurs with a weekly frequency, and user activity varies on weekdays and weekends.

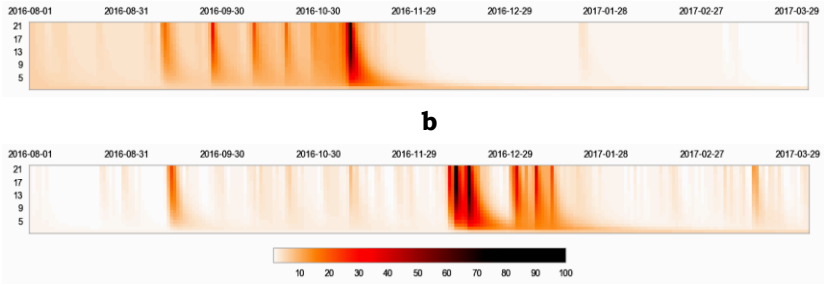
In order to exclude the periodic component from the series, we smooth them out using a simple moving average with an interval of 7 (number of days in a week) in accordance with the formula:

$$x_t^{New} = \frac{x_{t-3} + x_{t-2} + x_{t-1} + x_t + x_{t+1} + x_{t+2} + x_{t+3}}{7},$$

where x_t are the initial values of the series, x_t^{New} is the new value of the series at the moment of time t . On Fig. 19 shows the smoothed time series.



A



V

Figure 18 – Values of the exponential smoothed time series T (a), K (b) and X (c) depending on the parameter α

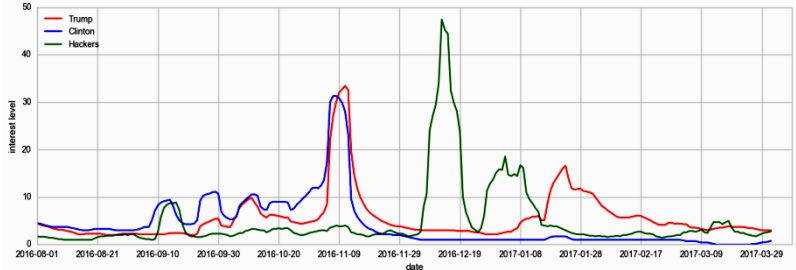


Figure 19 – Time series T, K and X, smoothed using a simple moving average with an interval of length 7

2.6.2. Correlation analysis

Many methods for studying time series are based on some assumption of statistical equilibrium or constancy. One such useful assumption is stationarity [Box 2015].

A time series is called strictly stationary or stationary in a narrow sense if its statistical properties do not change with time. Formally, if the joint distribution of random variables $x_t, x_{t+1}, \dots, x_{t+n}$ coincides with the distribution $x_{t+k}, x_{t+k+1}, \dots, x_{t+k+n}$ for any integer values of the shift k , then the time series $\{x_t\}_{t=1}^T$ is called strictly stationary. Stationary time series have a constant mathematical expectation

$$\mu = Ex_t$$

and variance

$$\sigma^2 = Var(x_t) = E(x_t - Ex_t)^2.$$

In this case, the values of μ and σ^2 can be estimated as the sample mean

$$\hat{\mu} = \bar{x} = \frac{1}{T} \sum_{t=1}^T x_t$$

and sample variance

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T (x_t - \bar{x})^2.$$

The property of stationarity is also of great importance when comparing time series. A linear relationship between two random variables is measured by covariance. For time series, a cross-covariance function is determined. By definition, cross-covariance with a time delay k between random processes $\{x_t\}_{t=1}^T$ and $\{y_t\}_{t=1}^T$ is equal to

$$\gamma_{xy}(k, t) = Cov(x_t, y_{t+k}) = E[(x_t - \mu_x)(y_{t+k} - \mu_y)].$$

From the assumption of stationarity in the narrow sense, it follows that the distribution of pairs of quantities x_t, y_{t+k} is the same for an arbitrary value of t . Therefore, the covariance between the quantities x_t and y_{t+k} does not depend on t , but depends only on the value of k , i.e. $\gamma_{xy}(k, t) = \gamma_{xy}(k), \forall t$. The set of values $\{\gamma_{xy}(k)\}$ forms a cross-covariance function.

Normalizing the cross-covariance coefficient, we obtain the cross-correlation coefficient

$$\rho_{xy}(k) = \frac{\text{Cov}(x_t, y_{t+k})}{\sigma_x \sigma_y} = \frac{\gamma_{xy}(k)}{\sigma_x \sigma_y}.$$

The cross-correlation function is a measure of similarity between two time series.

Most often, cross-covariance and cross-correlation coefficients are estimated using the formulas

$$\hat{\gamma}_{xy}(k) = \frac{1}{T} \sum_{t=1}^{T-k} (x_t - \bar{x})(y_{t+k} - \bar{y}), \quad \hat{\rho}_{xy}(k) = \frac{\hat{\gamma}_{xy}(k)}{\hat{\gamma}_{xy}(0)}.$$

Note that such estimates are valid for series that are stationary in the narrow sense, since the corresponding coefficients do not depend on time, and in the general case this may not hold. A weaker requirement than stationarity in the narrow sense is often used – stationarity in the broad sense.

The time series $\{x_t\}_{t=1}^T$ is stationary in a broad sense if its mathematical expectation does not change with time, that is, $\forall t \exists E x_t = \text{const}$ the covariance function also depends only on the difference of the arguments $\text{Cov}(x_t, x_s) = K(t - s)$.

Since the definition states that the mathematical expectation is constant and it is easy to see that the variance also does not change with time $\text{Var}(x_t) = \text{Cov}(x_t, x_t) = K(0) = \text{const}$, then in this case, as for strictly stationary series, estimates (1) and (2) are valid.

On Fig. 20 shows an illustration of a correlation calculation.

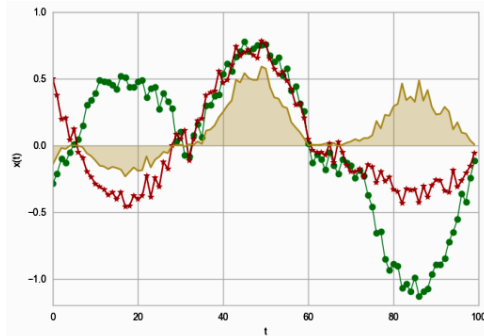


Figure 20 – Illustration for the definition of correlation. Two time series shown

Two centered time series are considered. To calculate the correlation coefficient, you need to multiply the corresponding elements of the series and calculate their average value. The result of multiplication in Fig. 3.8 is shown as a line. The area of the shaded area under the line, taking into account the sign, is equal to the coefficient of covariance between the two rows.

As an example, we present an estimate of the cross-correlation functions for the series T, K, X. Figure 21a shows the correlation function for the T and K series. The time delay (lag) is plotted along the abscissa, and the estimate of the correlation coefficient is plotted along the ordinate. The function reaches its maximum value (approximately equal to 0.8) at a time delay of 0. That is, the two time series associated with interest in Donald Trump and Hillary Clinton are strongly correlated. On Fig. Figure 21b shows the correlation function for the T and X series. The maximum value (approximately equal to 0.7) is reached by the function at a time delay of 34 days. This is in line with the fact that since December 13, 2016 (34 days after the US election on November 8), the number of news reports about “Russian hackers” has increased dramatically.

autocorrelation

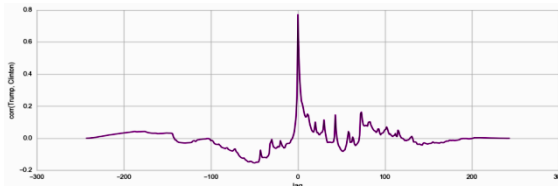
You can calculate the covariance not for two different series, but for one series. This covariance is called autocovariance with time delay or lag k

$$\gamma_k = \text{Cov}(x_t, x_{t+k}) = E[(x_t - \mu)(x_{t+k} - \mu)].$$

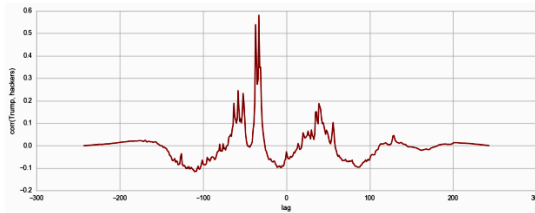
The set of values γ_k , $k = 0,1,2, \dots$ is called the autocovariance function, and their normalized value ρ_k , $k = 0,1,2, \dots$ is called the autocorrelation function

$$\rho_k = \frac{E[(x_t - \mu)(x_{t+k} - \mu)]}{\sqrt{E(x_t - \mu)^2 E(x_{t+k} - \mu)^2}} = \frac{Cov(x_t, x_{t+k})}{Var(x_t)} = \frac{\gamma_k}{\gamma_0}$$

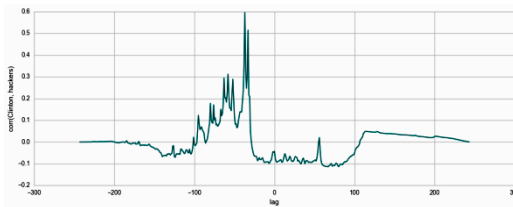
The autocorrelation function describes the relationship between the values of a random process at different points in time (Fig. 22). The figure shows the time series and the same series shifted 10 values to the right. The shaded area shows the contribution to the value of the autocorrelation coefficient with a lag of 10.



A



b



V

Figure 21 – Correlation functions for pairs of series T and K (a), T and X (b), K and X (c)

On Fig. 23 shows a autocorrelation functions for the series T (a), K (b), X (c). The time delay (lag) is plotted along the abscissa axis, and the autocorrelation coefficient is plotted along the ordinate axis. The shaded area shows the standard deviation for estimating the autocorrelation coefficient.

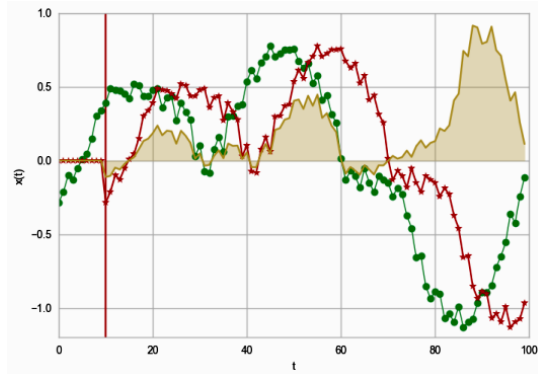


Figure 22 – Illustration for the definition of autocorrelation

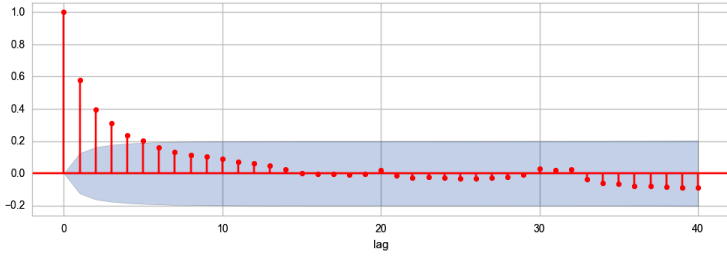
Most often, autocovariance and autocorrelation coefficients are estimated using the formulas

$$\hat{\gamma}_k = \frac{1}{T} \sum_{t=1}^{T-k} (x_t - \bar{x})(x_{t+k} - \bar{x}), \quad \hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}.$$

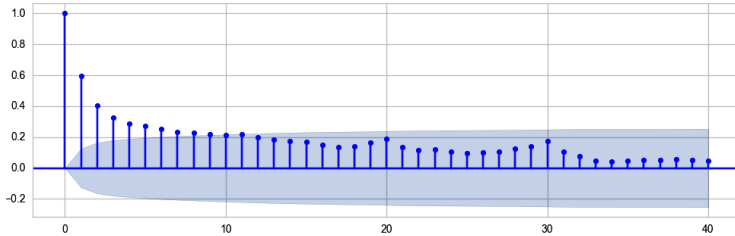
the coefficients equal ρ_k zero starting from some value k ? To answer this question, you need to compare the value of the estimate $\hat{\rho}_k$ with its standard deviation. If we accept the assumption that $\rho_k = 0$, then the standard deviation of the estimate $\hat{\rho}_k$

$$se(\hat{\rho}_k) \cong \frac{1}{\sqrt{T}}.$$

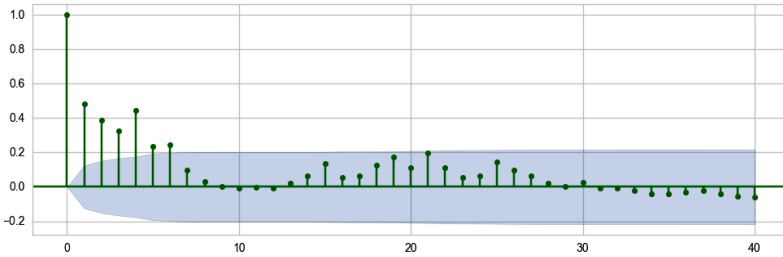
In practice, the rule of thumb is often used, according to which autocorrelation coefficients are estimated for a time delay of no more than $T/4$.



A



b



V

Figure 23 – Autocorrelation functions for the series T (a), K (b), X (c)

The definition of the autocorrelation function was introduced for stationary time series, but its value can be estimated for an arbitrary time series. For non-stationary time series, such an autocorrelation function decreases very slowly.

2.6.3. Fourier analysis

Classical Fourier analysis provides the ability to explore a function in the time and frequency domain. The essence of the transition to the frequency domain is that the function is decomposed into components, which are harmonic oscillations with

different frequencies. In this case, each frequency corresponds to a coefficient that displays the amplitude of the oscillation at a given frequency. If we represent a function graphically in the time domain, we get information about how the function changes over time. If we depict a function in the frequency domain, we will obtain information about the frequencies at which it contains oscillations. For this, the direct and inverse Fourier transforms are used.

$$\hat{x}(\nu) = \int_{-\infty}^{\infty} x(t)e^{-i2\pi\nu t} dt,$$

$$x(t) = \int_{-\infty}^{\infty} \hat{x}(\nu)e^{i2\pi\nu t} d\nu.$$

On Fig. 24a shows an example of a function that is actually the sum of three sinusoids with different periods. Looking only at the graph of a function in the time domain, it is quite difficult to understand that it consists of three harmonic oscillations and determine their periods. Figure 24b shows the Fourier transform for this function. From the graph in the frequency domain, it is clearly seen that the function contains oscillations at three different frequencies.

Today, the Fourier transform and spectra find a variety of applications in machine learning systems. Often Fourier spectra are used as training parameters. For example, in [Rodrigues 2014], a time series forecasting model is proposed in which the Fourier spectrum, along with some other parameters, is fed to the input of a neural network.

Fourier transforms and spectra are often used in speech recognition. In [Alam 2014], special features formed on the basis of the Fourier transform are used in a speech recognition system with different learning conditions.

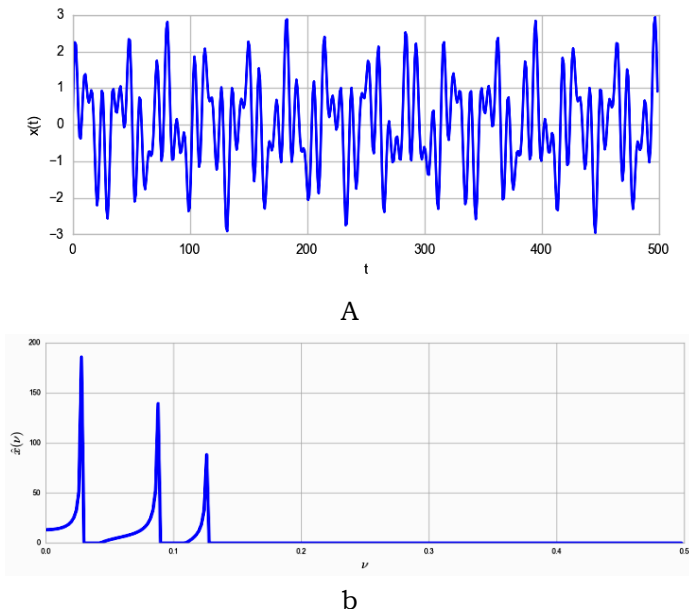


Figure 24 – A function that is the sum of three sinusoids with different periods (a) and the estimated Fourier spectrum for this function (b)

Fourier spectra are also used as training parameters for neural networks in automatic detection of certain events in speech or noise [Sazonov 2010, Wang 2014]. Another task in the field of recognition is to determine the emotional coloring of speech. [Wang 2015] proposes a recognition model based on certain Fourier parameters and demonstrates the effectiveness of using such parameters to identify various emotional states in voice signals.

Kernel-based machine learning algorithms, such as support vector machines, often use random Fourier features to approximate high-dimensional kernels. This approach was proposed in [Rahimi 2008] and is based on Bochner's theorem from harmonic analysis, which guarantees that for some properties of the kernel, its Fourier transform will be a probability distribution.

The Fourier transform can be thought of as determining the correlation between the original signal and harmonic functions

with different oscillation frequencies. On Fig. 25 shows an illustration similar to Fig. 20 and Fig. 22. The shaded area shows the contribution to the value of the Fourier transform or amplitude that corresponds to a given oscillation frequency.

Despite its advantages and numerous applications, the Fourier transform is a poor method for investigating functions that evolve over time. For such functions, some way of estimating the spectrum is needed not over the entire length of the time series, but over its various parts. An example of such an approach is the windowed Gabor transform

$$G(\nu, \tau, s) = \int_{-\infty}^{\infty} x(t) e^{-\frac{(t-\tau)^2}{s^2}} e^{-i2\pi\nu t} dt.$$

The time window $e^{-\frac{(t-\tau)^2}{s^2}}$ selects a segment of the time series centered at a point τ and has a width that is determined by the parameter s , which allows you to select a part of the series under study.

When using the Gabor transform, the problem of choosing the window width arises. To make the window function dependent on frequency so that the window becomes wider for low frequencies and narrower for high frequencies, the following class of transformations, namely the wavelet transform, allows. The main advantage of the wavelet transform is that the piece selected from the time series is analyzed with the degree of detail that corresponds to its scale.

2.6.4. Wavelet analysis

The wavelet transform has a correlation nature. In this case, the correlation of the original function with the wavelet function on different scales is considered. In order for such a procedure to always be performed and the correlation coefficients to be informative, the wavelet must have certain mathematical properties.

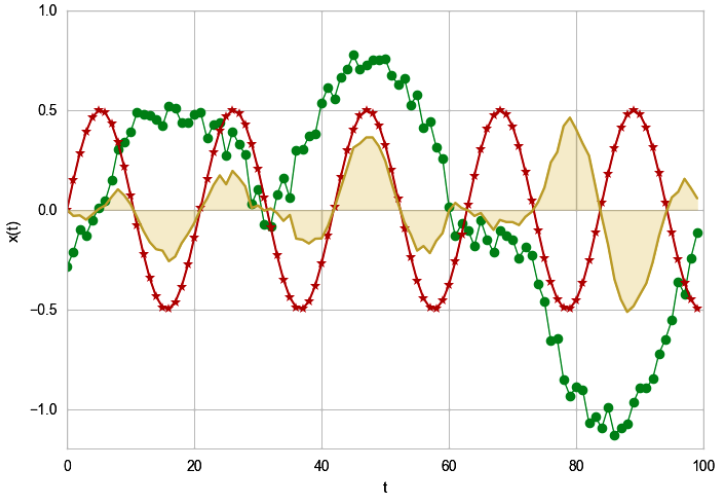


Figure 25 – Illustration for the definition of the Fourier transform as a calculation of the correlation between the original signal and the harmonic oscillation.

Literally, the word wavelet translates as "small wave" or "burst", and as the name implies, the wavelet is well localized in time. From a mathematical point of view, a wavelet is a function $\psi(t)$, which satisfies the following properties:

1. Function $\psi(t)$ square integrable ($\psi \in L^2(\mathbb{R})$) or, in other words, has a finite energy

$$E = \int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty.$$

2. Denote $\hat{\psi}(\lambda)$ the Fourier transform of the function $\psi(t)$, then

$$\int_0^{\infty} \frac{|\hat{\psi}(\lambda)|^2}{\lambda} d\lambda < \infty.$$

On Fig. 26 shows examples of wavelets that are often used in practice.

Continuous Wavelet Transform

The wavelet $\psi(t)$ whose properties have been described above is often referred to as the mother or base wavelet. Based

on the mother wavelet, a family of functions is built using stretching/compression and parallel translation. This is necessary to explore different areas of the original signal and with varying degrees of detail.

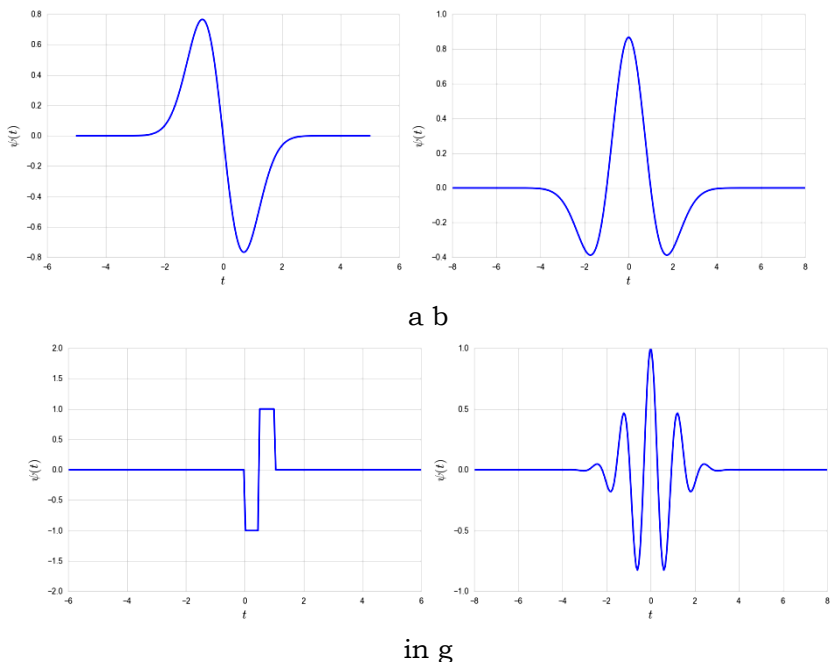


Figure 26 – Examples of wavelets that are often used in applications: (a) Gaussian wave (the first derivative of a Gaussian function), (b) Mexican hat, (c) Haar wavelet, (d) Morlet wavelet (real part).

and shift (location) l parameters s , then the transformed version of the mother wavelet will be as follows

$$\psi_{s,l}(t) = \frac{1}{\sqrt{|s|}} \psi\left(\frac{t-l}{s}\right).$$

The continuous wavelet transform of a function $x(t) \in L^2(\mathbb{R})$ is the expression

$$W(s, l) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t-l}{s} \right) dt = \int_{-\infty}^{\infty} x(t) \psi_{s,l}^*(t) dt,$$

where $l, s \in \mathbb{R}$, $s \neq 0$; ψ^* is the complex conjugate function of ψ , the quantities $\{W(s, l)\}_{l, s \in \mathbb{R}}$ are called wavelet transform coefficients.

From the formula in the definition of a continuous wavelet transform, it is directly seen that the essence of such a transform is to calculate the correlation coefficients of a special form. On Fig. 27 shows how the Mexican hat wavelet is superimposed on the original series and the correlation between part of the series and the “template”, which is a wavelet, is determined. The shaded area shows the contribution to the value of the wavelet transform for the shift and scale data.

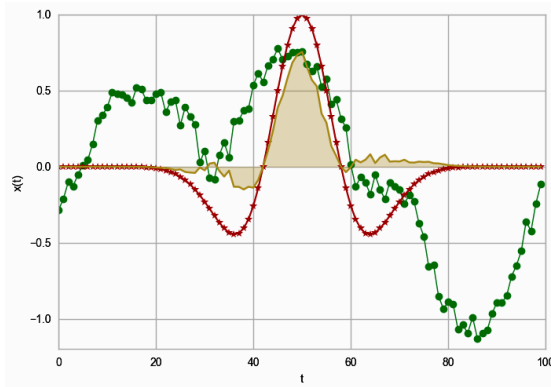


Figure 27 – Illustration for the calculation of the wavelet transform as the calculation of the correlation between the original signal and the wavelet function

It is worth noting that the continuous wavelet transform is a reversible operation. The inverse wavelet transform is performed as follows:

$$x(t) = \frac{1}{C_g} \int_{-\infty}^{\infty} \int_0^{\infty} W_x(s, l) \psi_{s,l}(t) \frac{ds dl}{s^2}.$$

Let's illustrate the results of the wavelet transform with a few simple examples (Fig. 15). The first example is the sum of two

oscillatory processes. Below the signal plot are the wavelet transform coefficients that were obtained using the Mexican hat wavelet. The time changes along the horizontal axis (shift parameter l), while the scale changes along the vertical axis (parameter s). On the graph of the wavelet coefficients, two periodic processes can be seen. On Fig. 28, you can see how periodic processes with different amplitudes and frequencies, as well as individual peaks in the signal, affect the value of the wavelet coefficients.

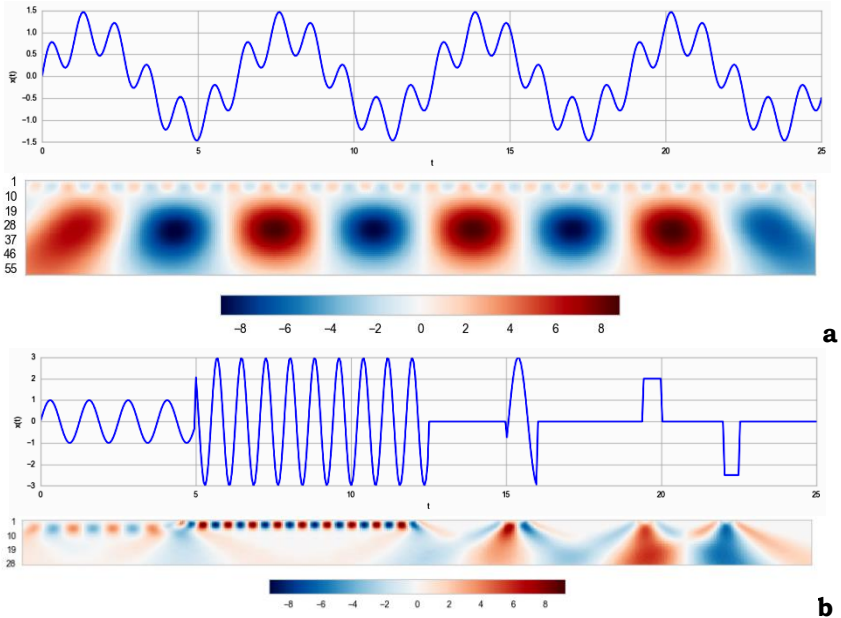


Figure 28 – Examples of wavelet transform. The sum of two sinusoids and its wavelet transform (a). A function composed of oscillatory processes of different frequencies and amplitudes, as well as individual peaks (b).

Consider the wavelet transform coefficients for the time series T, K, X. 29 a, b shows the results of the wavelet transform for the T series using the Mexican hat wavelet (a) and the Gaussian wave (b). On Fig. 29 c, d show the results of the wavelet transform for the K and X series using the Mexican hat wavelet.

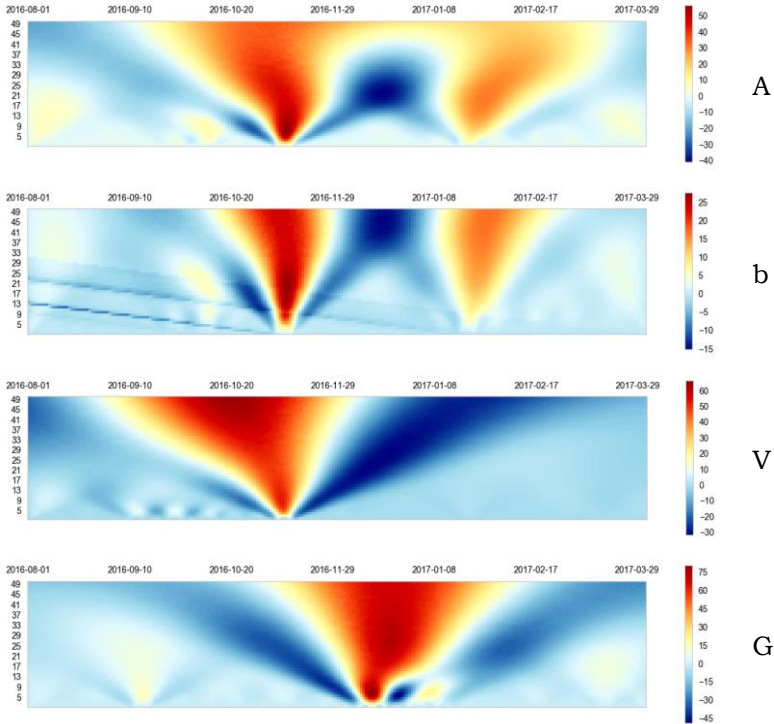


Figure 29 – Wavelet transform for the T series using the Mexican hat wavelet (a); the real part of the wavelet transform for the series T using the Morlet wavelet (b); wavelet transform for series K (c) and series X (d) using Mexican hat wavelet

Once again, we note that the continuous wavelet transform, like the Fourier transform, can be considered in terms of correlation. The Fourier transform is the correlation between the original time series and the wave $\varphi(t) = e^{-i2\pi\nu t}$. The wave covers the entire time axis and is characterized only by the frequency ν , so the Fourier transform depends only on the frequency. The wavelet transform is the correlation between the original time series and the wavelet $\psi(t)$. Thus, the wavelet transform depends on the position of the wavelet on the time axis and its scale, which are determined by the parameters $land$ srespectively.

Signal energy

Total signal energy $x(t)$ is by definition equal to

$$E = \int_{-\infty}^{\infty} |x(t)|^2 dt.$$

Using the wavelet transform coefficients, you can determine the signal energy that corresponds to a certain shift and scale

$$E(s, l) = |W(s, l)|^2.$$

The values $E(s, l)$ can be plotted in the same way as the wavelet transform coefficients. Such a graph is usually called a scalogram. It is also possible to determine the relative contribution of energy, which corresponds to a certain scale, to the total energy, or in other words, the distribution of energy depending on the scale

$$E(s) = \frac{1}{C_g} \int_{-\infty}^{\infty} |W(s, l)|^2 dl.$$

Note that on the scalogram, in contrast to the graph with wavelet transform coefficients, all values are positive, and the areas with the highest energy stand out most clearly (Fig. 30).

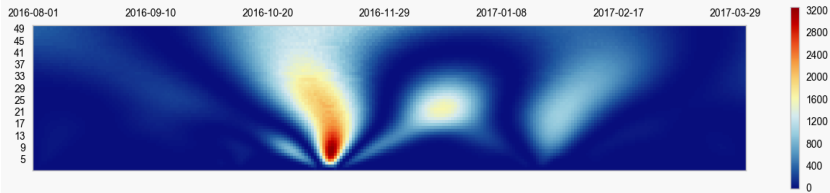


Figure 30 – Scale diagram for the time series T

Time series comparison using wavelet transform

Previously, a way was shown to identify the relationship between two time series using cross-correlation. Now we will consider some ways to compare time series using wavelet transform coefficients. These methods can also be used to reveal some type of relationship or relationship between time series. Metrics for comparing wavelet transform coefficients, as well as examples of their application to real practical problems, are described in detail in [Addison 2017].

Consider two time series x_t and y_t , and denote the wavelet transform coefficients of these series $W_x(s, l)$ and $W_y(s, l)$. Let's start with the simplest comparison method – take the difference between the moduli of the corresponding coefficients

$$DiffMOD_{x,y}(s, l) = |W_x(s, l)| - |W_y(s, l)|.$$

On Fig. Figure 31 shows the values $DiffMOD_{x,y}(s, l)$ for two pairs of series – on top for the T and K series, and on the bottom for the T and X series. In this simple way, you can select areas in which the wavelet transform coefficients are similar, which means that there are similar sections in the original time series.

Another simple way of comparison is the ratio of the absolute values of the coefficients of the wavelet transform

$$RatioMOD_{x,y}(s, l) = \frac{|W_x(s, l)|}{|W_y(s, l)|}.$$

This metric should be used with caution, as $W_y(s, l)$ take on zero or near-zero values.

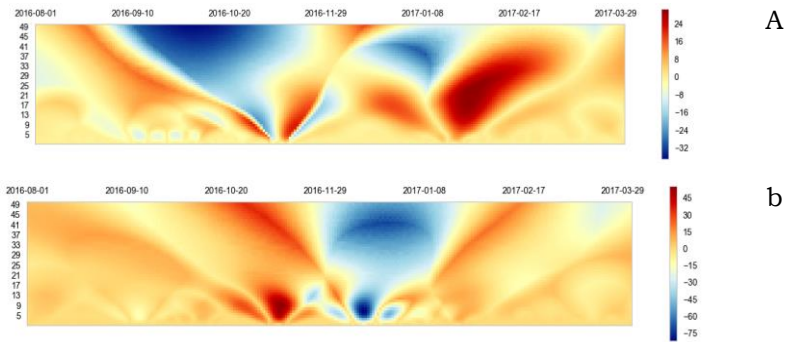


Figure 31 – $DiffMOD_{x,y}(s, l)$ for rows T and K (a), T and X (b)

Additional information can be obtained by using a complex wavelet (for example, the Morlet wavelet). Then, in addition to the absolute value of the wavelet coefficients, a phase appears. The complex coefficient can always be represented as

$$W(s, l) = |W(s, l)|e^{i\phi(s, l)},$$

where, as you know, the modulus of a number is equal to

$$|W(s, l)| = \sqrt{\text{Re}(W(s, l))^2 + \text{Im}(W(s, l))^2},$$

a phase

$$\phi(s, l) = \tan^{-1} \left[\frac{\text{Im}(W(s, l))}{\text{Re}(W(s, l))} \right].$$

Therefore, one can also compare the phases of the coefficients

$$\Delta\phi_{x,y}(s, l) = \phi_x(s, l) - \phi_y(s, l)$$

Cross-wavelet transform is used to highlight areas of equal energy between signals in the transform region, as well as to determine the relative phase

$$CrWT_{x,y}(s, l) = W_x^*(s, l)W_y(s, l).$$

The figures usually display the value $|CrWT_{x,y}(s, l)|$, by analogy with the scalogram. In this case, if the time series x is identical to the series y , then we will get a scalogram for the series x .

Of particular interest is the calculation of the cross-wavelet transform when a complex wavelet is used (for example, the Morlet wavelet). Then

$$\begin{aligned} CrWT_{x,y}(s, l) &= W_x^*(s, l)W_y(s, l) = |W_x(s, l)|e^{-i\phi_x(s, l)}|W_y(s, l)|e^{i\phi_y(s, l)} = \\ &= |W_x(s, l)||W_y(s, l)|e^{i(\phi_y(s, l) - \phi_x(s, l))}. \end{aligned}$$

Thus, by calculating the cross-wavelet transform, one can extract the value of the phase difference between the wavelet transform coefficients for two time series.

On Fig. 32 shows the values $CrWT_{x,y}(s, l)$ for the T and K rows. For the T and K rows, the area corresponding to the peak of interest during the election is highlighted, which means that this is a high energy area for both rows.

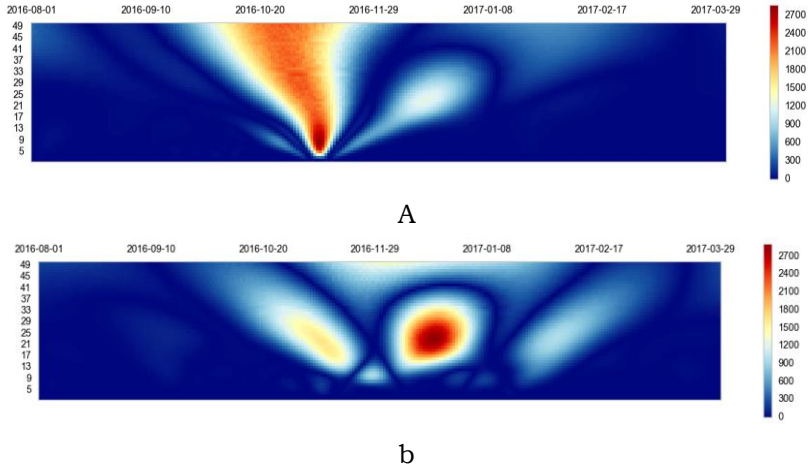


Figure 32 – Cross-wavelet transform using the Mexican hat wavelet for the T and K series (a) and the T and X series (b)

If we integrate the coefficients of the wavelet transform over time, we get a wavelet cross-correlation measure (Wavelet Cross-Correlation Measure), which depends on the scale

$$W_{x,y}(s) = \frac{|\int W_x^*(s, l)W_y(s, l)dl|}{\sqrt{\int |W_x(s, l)|^2 dl \int |W_y(s, l)|^2 dl}} = \frac{|\int CrWT_{x,y}(s, l)dl|}{\sqrt{\int |W_x(s, l)|^2 dl \int |W_y(s, l)|^2 dl}}$$

Such a measure helps to detect the correlation between signals that contain oscillations with different amplitude or phase, but, nevertheless, are correlated with each other (Fig. 33).

It is also possible to expand the definition of wavelet cross-correlation measure by introducing a dependence on the shift between the series

$$W_{x,y}(s, k) = \frac{|\int W_x^*(s, l)W_y(s, l - k)dl|}{\sqrt{\int |W_x(s, l)|^2 dl \int |W_y(s, l)|^2 dl}}$$

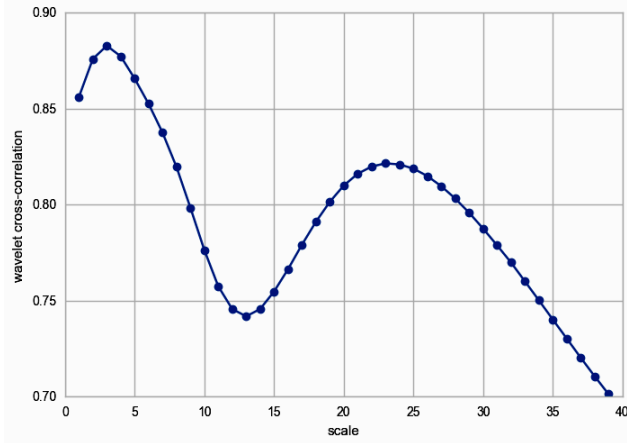


Figure 33 – Dependence of the wavelet-cross-correlation measure on the scale for the time series T and K

To investigate the relationship between time series components locally in the transformation plane, we can define a quadratic wavelet coherence estimate as follows:

$$WCH_{x,y}^2(s, l) = \frac{|\langle W_x^*(s, l) W_y(s, l) \rangle|^2}{\langle |W_x(s, l)|^2 \rangle \langle |W_y(s, l)|^2 \rangle},$$

where $\langle \cdot \rangle$ denotes a local smoothing operation, both in the time scale and in the scale, while smoothing is performed on the components of the transformation.

Cross-wavelet analysis methods are used to study the properties of several time series that are non-trivially dependent on each other. For example, in geophysics, the problem arises of identifying causal relationships or correlations between meteorological or other environmental phenomena that occur at a great distance from each other. In [Maraun, 2004], the features of applying the cross-wavelet transform and estimating the coherence of wavelets for two-dimensional time series of this type are analyzed. Also, wavelet analysis methods, including cross-wavelet transform, were used in [Adamowski, 2008] to study meteorological time series and river flow data. From time series of two types, components were isolated, which were further used in the flood forecasting model. The paper shows that the use of cross-

wavelet analysis is useful when there is a relatively stable phase shift between the streaming and meteorological time series. Using cross-wavelet transform, the phase difference between the streaming and meteorological data was determined, which improved the quality of the flood forecasting model.

Another example is [Labat, 2010], where cross-wavelet analysis was carried out for climate indices and freshwater discharge rates in Africa. In this case, cross-wavelet transform and coherence estimation were used to visualize and analyze periodic fluctuations in data 2-8 years long. In [Kelly, 2003], it is confirmed that methods based on cross-wavelet analysis can be an effective tool in the search for quasi-periodicity of a time series.

Cross-wavelet analysis methods have also found application in medicine. [Li, 2007] describes the use of cross-wavelet transform, coherence estimation, and some other wavelet-based methods to study the interaction dynamics between oscillations generated by two anatomically different groups of neurons. The results of the study can be used to analyze and quantify the temporal interaction between neural oscillators, as well as to study the mechanisms of epilepsy.

[Aguiar-Conraria, 2008] uses cross-wavelet analysis tools to show that the relationship between monetary policy variables and macroeconomic variables has changed over time, and these changes are not uniform across frequencies.

Data obtained using cross-wavelet transform can also be used as input for classification algorithms. In [Dey, 2010], cross-wavelet transform coefficients were fed to the input of an artificial neural network and a Fuzzy classifier.

Discrete wavelet transform and approximation of functions using series

The discrete wavelet transform is defined in such a way that it is possible to completely reconstruct the original signal using infinite sums of discrete wavelet coefficients. This approach also leads to fast computation of the wavelet transform and its inverse [Addison, 2017].

Let $x(t)$ belong to the space 2π -of periodic square-integrable functions. Then $x(t)$ can be represented as a Fourier series

$$x(t) = \sum_{n=-\infty}^{\infty} c_n e^{int},$$

where the coefficients c_n look like:

$$c_n = \frac{1}{2\pi} \int_0^{2\pi} x(t) e^{-int} dt.$$

The feature set $w_n(t) = e^{int}$ is an orthonormal basis in space $L^2(0, 2\pi)$, built using a scaling transformation $w_n(t) = w(nt)$ from the base function $w(t) = e^{int}$.

Let now $x \in L^2(\mathbb{R})$. The basis function in space $L^2(\mathbb{R})$ must be a function that decays fast enough to 0 on $\pm\infty$. Therefore, wavelets are used to build the basis – well-localized soliton-like functions. In order to cover the entire real axis with wavelets, translation along the axis is used. For simplicity, integer shifts k and sinusoidal frequency analogs can be used, as powers of two $\psi_{jk} = \psi(2^j t - k)$. A wavelet $\psi \in L^2(\mathbb{R})$ is called orthogonal if the family of functions $\{\psi_{jk}\}$ forms an orthonormal basis in $L^2(\mathbb{R})$.

2.6.5. Pattern correlation

With the help of a continuous wavelet transform, sections of the series under study are revealed, which are most similar in shape to a wavelet (Fig. 34). The idea is to compare parts of a series with some pattern at different scales (Fig. 35). In this case, the wavelet as a function must have certain mathematical properties, in particular, it must rapidly decrease to zero at infinity. In some cases, it is useful to use a pattern that does not match the requirements of the wavelet. To do this, instead of the wavelet transform, we will calculate the correlation between a part of the time series and some pattern p

$$C(l, k) = \frac{\sum_{i=1}^k (x_{l+i} - \bar{x})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^k (x_{l+i} - \bar{x})^2 \sum_{i=1}^k (p_i - \bar{p})^2}}$$

Received coefficient $C(l, k)$ depends on the values x_{l+1}, \dots, x_{l+k} . That is, the parameter l corresponds to the template shift, and the parameter k corresponds to the number of points in the template and in the considered segment of the series. The parameter

k in this case is analogous to the scale s , which was used in the wavelet transform.

If the entire time series was always used when calculating the wavelet transform coefficient, then in this case, to calculate $C(l, k)$ row points and a length pattern k are used k .

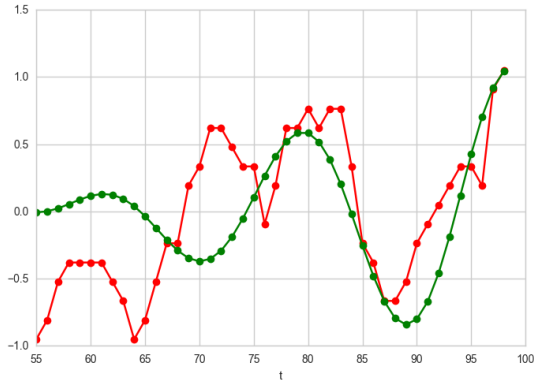


Figure 34 – Segment of the time series with a superimposed template

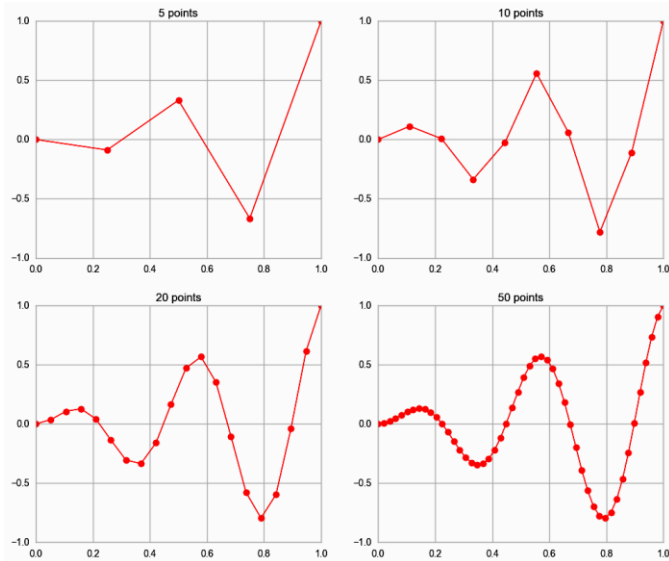
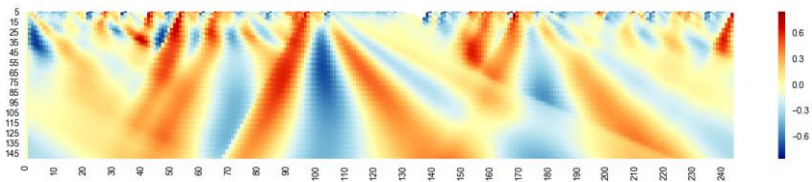
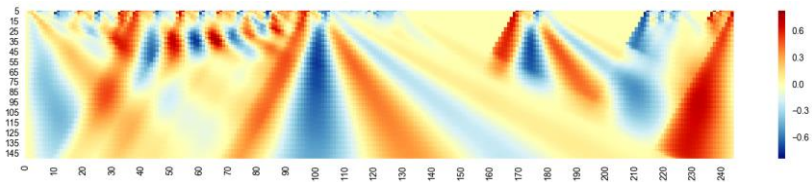


Figure 35 – Snake pattern with different number of dots

Obtained correlation coefficients $C(l, k)$ represent on a graph that looks like a scalogram (Fig. 36).



A



b

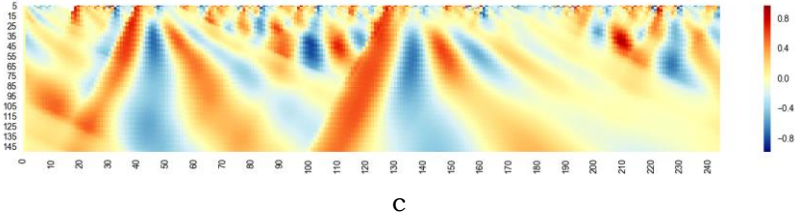


Figure 36 – Correlation coefficients $C(l, k)$ calculated for the series T (a), K (b) and X (c) using the template shown in Fig. 3.22

2.6.6. fractal analysis

The term fractal was introduced and popularized by Benoit Mandelbrot. Most often, fractals are called geometric objects that have a strongly jagged shape and have the property of self-similarity.

A strictly and generally accepted definition of a fractal does not currently exist, although Benoit Maldebrot used several tentative definitions. One of them, introduced in [Mandelbrot, 1982], reads as follows:

A fractal is a set whose Hausdorff-Besikovich dimension is strictly greater than its topological dimension.

A rigorous definition of the Hausdorff-Besikovich or fractal dimension will be introduced later. The essence of such a definition is to single out a class of strongly jagged objects, for which the topological dimension is not enough to describe. For example, there are curves whose topological dimension is always 1, but which are curved in such a complex way that they fill a plane or space. So the Peano curves pass through any point of the unit square. Another example is the trajectory of a Brownian particle, which is not smooth at any point.

The first definition, although strict, excludes many physical fractals and is therefore not used. The following definitions of a fractal have been proposed:

A fractal is a structure consisting of parts that are in some sense similar to the whole.

The second definition emphasizes that the hallmark of a fractal is self-similarity. Let us give a rigorous definition of a self-

similar set, which is used in mathematics. This will require a few preliminary definitions.

Let the set $E \subset \mathbb{R}^d$ be closed. Then the mapping $S: E \rightarrow E$ is called a mapping of similarity (similarity) on E if $\exists t: 0 < t < 1: |S(x) - S(y)| = t|x - y|, \forall x, y \in E$.

That is, the similarity mapping turns a set E into a geometrically similar set.

Consider a set of similarity mappings S_1, \dots, S_m . A set $F \subseteq E$ is invariant under transformations S_i if

$$F = \bigcup_{i=1}^m S_i(F).$$

A set that is invariant under a set of similarity mappings is called self-similar.

The definition of a self-similar set can also be understood intuitively. Indeed, by definition, a set is self-similar if it can be "assembled" from pieces that are similar to the whole set. Then the simplest example of a self-similar set is the segment $[0,1]$. Take, for example, $S_1 = \frac{x}{2}$ and $S_2 = \frac{1}{2} + \frac{x}{2}$. Then $[0,1] = S_1([0,1]) + S_2([0,1])$.

Obviously, simple self-similarity is not enough to call an object a fractal. Indeed, we will not consider a straight line segment or a piece of paper in a box as fractals. We will consider fractal objects that have the property of self-similarity, as well as a complex structure.

Let us give one basic elementary example of a fractal set, which will be convenient to use to demonstrate the main ideas in the future. This is the set of Cantor or Cantor dust. The classical process of constructing a Cantor set begins with a unit segment $C_0 = [0,1]$. Let us remove the middle third $C_1 = \left[0, \frac{1}{3}\right] \cup \left[\frac{2}{3}, 1\right]$ from C_0 , leaving the set C_1 . The set C_1 consists of two segments, from each of which we now remove the middle third, we get the set C_2 . Continuing to repeat this procedure, we obtain a sequence of sets $\{C_i\}_{i=1}^{\infty}$. The Cantor set is the intersection

$$C = \bigcap_{i=1}^{\infty} C_i.$$

Note that the set C is self-similar. Take similarity maps $S_1(x) = \frac{x}{3}$ and $S_2(x) = \frac{x}{3} + \frac{2}{3}$, then $C = S_1(C) \cup S_2(C)$. On the other hand, it is known that the Lebesgue measure of a Cantor set is equal to 0, just like for a point, or any countable set. But it is obvious that the structure of the Cantor set is much more complex, and here the idea already arises that a special measure is needed to describe such a set, which will be the fractal dimension of the set.

Fractal dimension

Let us return to the definition of the Hausdorff-Besikovich dimension. To do this, we need the notion of a set cover.

Let U be a non-empty set in \mathbb{R}^d . The diameter of a set U_i , by definition, equal to

$$|U| = \sup\{|x - y| : x, y \in U\}.$$

If $F \subset \bigcup_{i=1}^{\infty} U_i$ and $0 < |U_i| \leq \delta$ for any i , then the set of sets $\{U_i\}$ is called δ -a cover for the set U .

Let F be a subset of some closed set in \mathbb{R}^d . For arbitrary $s \geq 0$ and $\delta > 0$ define

$$\mathcal{H}_\delta^s(F) = \inf \left\{ \sum_i |U_i|^s : \{U_i\} - \delta - \text{покрытие для } F \right\},$$

where the infimum is taken over all possible δ -covers of the set F . By definition, the s -dimensional Hausdorff measure

$$\mathcal{H}^s(F) = \lim_{\delta \rightarrow 0} \mathcal{H}_\delta^s(F).$$

Such a limit exists for any set $F \subset \mathbb{R}^d$, with the proviso that it is often equal to zero or infinity.

The Hausdorff-Besicovich dimension of a set F is defined as

$$D_H(F) = \inf\{s : \mathcal{H}^s(F) = 0\} = \sup\{s : \mathcal{H}^s(F) = \infty\}.$$

Or what is the same

$$\mathcal{H}^s(F) = \begin{cases} 0, & s < D_H(F); \\ \infty, & s > D_H(F). \end{cases}$$

For example, let's calculate the dimension of the Cantor set C . The construction procedure was described above, and according to it, at the n th step there are 2^n segments of length $1/3^n$ each, and then the set only decreases. Therefore, as the diameter of the coating, δ you can take the value $1/3^n$ and use 2^n the sets in the coating. A-priory

$$\mathcal{H}^s(C) = \lim_{\delta \rightarrow 0} \mathcal{H}_\delta^s(C),$$

and now we can go from the limit on $\delta \rightarrow 0$, to the limit $n \rightarrow \infty$

$$\lim_{\delta \rightarrow 0} \mathcal{H}_\delta^s(C) = \lim_{n \rightarrow \infty} \sum_{i=1}^{2^n} \left(\frac{1}{3^n}\right)^s.$$

It remains to determine the value of such a limit

$$\lim_{n \rightarrow \infty} \sum_{i=1}^{2^n} \left(\frac{1}{3^n}\right)^s = \lim_{n \rightarrow \infty} \left(\frac{2}{3}\right)^n = \begin{cases} 0, & \frac{2}{3} < 1, \\ 1, & \frac{2}{3} = 1, \\ \infty, & \frac{2}{3} > 1. \end{cases}$$

Therefore, the limit $\mathcal{H}_\delta^s(C)$ at $\delta \rightarrow 0$ is not equal to zero or infinity at $\frac{2}{3^s} = 1$, therefore

$$D_H(C) = \frac{\ln 2}{\ln 3} \approx 0,63.$$

Statistically self-similar processes and the Hurst exponent

Many objects in the world around us are statistically self-similar (a classic example is coastlines), which means that parts of such objects have the same statistical characteristics when scaled. When studying the evolution of information flows, the structure of document arrays on the Internet and the study of processes in the information space, self-similar structures often arise, and in particular time series.

Let us define a self-similar process.

A real-valued process $\{x(t), t \in \mathbb{R}\}$ is self-similar with the Hurst exponent $H > 0$ if, for all, $\alpha > 0$ the finite-dimensional distribu-

tions $\{x(\alpha t), t \in \mathbb{R}\}$ are identical to the finite-dimensional distributions $\{\alpha^H x(t), t \in \mathbb{R}\}$, which can be briefly written

$$\{x(\alpha t), t \in \mathbb{R}\} =^d \{\alpha^H x(t), t \in \mathbb{R}\}.$$

That is, by definition, for a self-similar process, changing the time scale is equivalent to changing the scale of process values. This means that implementations of such processes look the same at different scales. In this case, it is natural that the process is not an exact copy of itself on different scales, only statistical properties are preserved.

exponent is a measure of persistence—the propensity of a process to trend. The value $H = 0.5$ corresponds to the uncorrelated behavior of the values of the series, as in the Brownian motion. Values in the range $0.5 < H < 1$ mean that the dynamics of the process directed in a certain direction in the past, most likely, will entail the continuation of movement in the same direction. If $H > 0.5$, then it is predicted that the process will change direction.

Let us describe some properties of self-similar processes that are important for applications. First, for such processes, the autocovariance function decays hyperbolically and has the form

$$\rho_k \approx k^{(2H-2)} L(t) \text{ при } k \rightarrow \infty,$$

where $L(t)$ is a function slowly varying at infinity, that is, such that

$$\forall x > 0: \lim_{t \rightarrow \infty} \frac{L(tx)}{L(t)} = 1.$$

Therefore, for self-similar processes, the series of covariance coefficients diverges

$$\sum_{k=1}^{\infty} \rho_k = \infty.$$

Such an infinite sum indicates a long-term dependence in the series.

Second, the variance of the sample mean decreases more slowly than the reciprocal of the sample size

$$\sigma^2(x_t^{(m)}) \sim m^{2H-2},$$

where the sequence $\{x_t^{(m)}\}$ obtained by dividing the original sequence $\{x_t\}$ into non-overlapping blocks of length m and taking the average in each of the blocks.

Hurst exponent estimation methods

The method for estimating the Hurst exponent, proposed by himself, is called the normalized range method or R/S analysis. For a time series, $\{x_t\}_{t=1}^T$ the standard deviation S is determined by the formula

$$S = \sqrt{\frac{1}{T} \sum_{t=1}^T (x_t - \bar{x})^2}, \quad \bar{x} = \frac{1}{T} \sum_{t=1}^T x_t,$$

and the range of the series

$$R = \max_{1 \leq t \leq T} x^{(t)} - \min_{1 \leq t \leq T} x^{(t)}, \quad x^{(t)} = \sum_{i=1}^t (x_i - \bar{x}).$$

The ratio R/S is the normalized range. Hurst found that for many observed time series, the normalized range is well described by the empirical relation

$$\frac{R}{S} = \left(\frac{T}{2}\right)^H.$$

The values of the Hurst exponent can be estimated by calculating the values of statistics R/S depending on T and plotting such a dependence on a double logarithmic scale. The estimate of the Hurst exponent will be the estimate of the slope of the straight line, which best approximates the dependence $\log R/S$ on $\log T$.

We use the method R/S to calculate the Hurst exponent for the series T, K, and X. Figure 37 shows the results of estimation for the T and K series. The obtained values of the Hurst exponent – 0.62 and 0.68, respectively – indicate the propensity of these processes to trends, although not very high.

In the case of series K in Fig. 37 shows that the dependence $\log R/S$ on $\log t$ is poorly approximated by a linear dependence, since the graph has a strong break. If we build the dependence of

the Hurst exponent on time (Fig. 38), then we can determine the moment of time, starting from which the value of the exponent begins to decrease. By marking this point in time on the X time series graph, you can see that this is the moment of a sharp increase in the values of the series, until which the values of the series had significantly less variance.

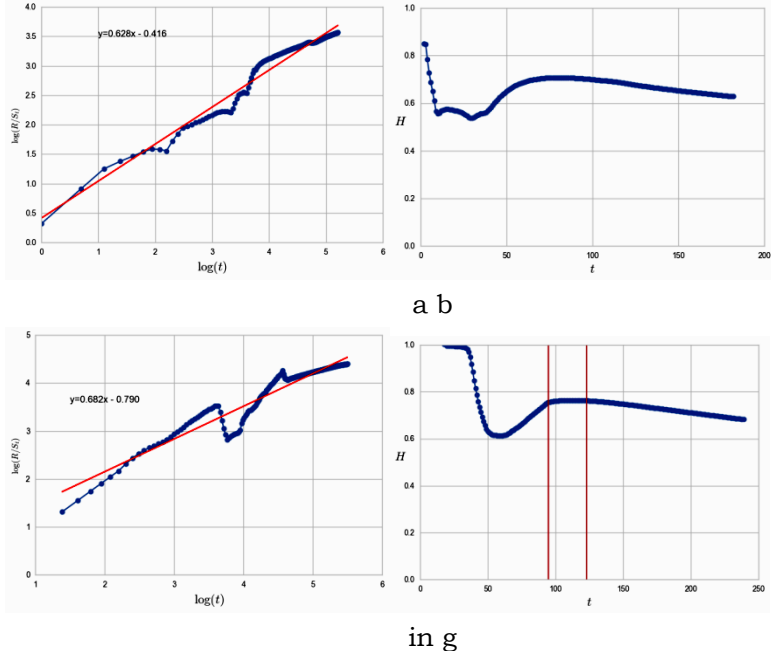
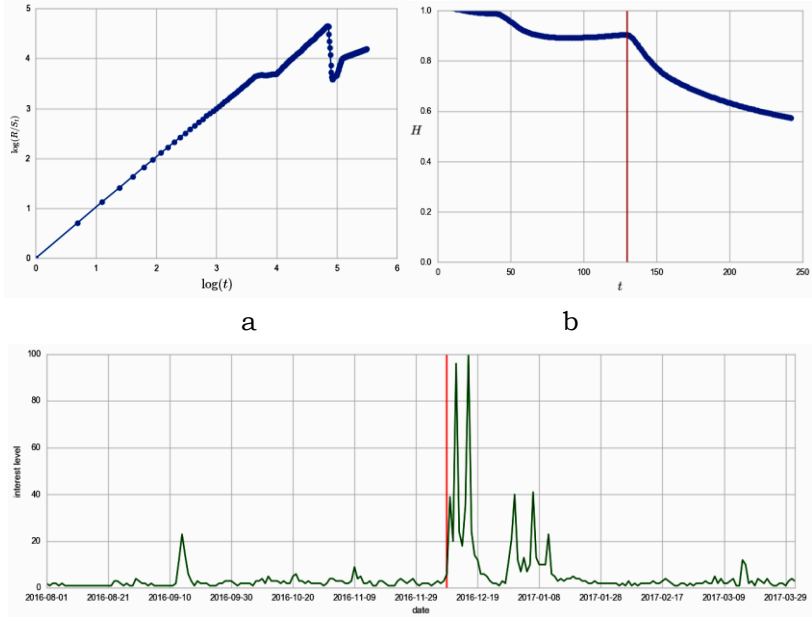


Figure 37 – Evaluation of the Hurst exponent for the T and K series. Dependence of statistics R/S for the T series (a) and the K series (c) on time in a logarithmic scale. Time dependence of the Hurst exponent for the T series (b) and the K series (d)

The behavior of the X series since the beginning of December 2016 (the beginning of the largest peak in the values of the series) can be considered separately. The estimate of the Hurst exponent for the second part of the series is shown in Fig. 39 It should be taken into account that in this example the time series becomes too short, since R/S series with at least 200 elements are used for analysis. However, the sharp changes in the dependence of the Hurst exponent on time, which have the form of

a “step”, indicate that the process under study consists of various processes that it makes sense to consider separately.

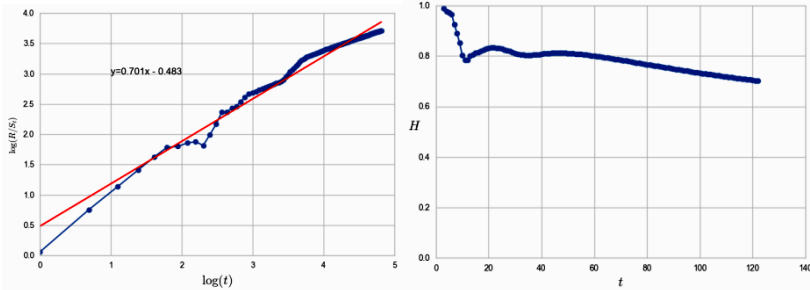


V

Figure 38 – Evaluation of the Hurst exponent for series X. Dependence of statistics R/S on time in a logarithmic scale (a). Time dependence of the Hurst exponent (b). The vertical line marks the moment of time after which the value of the Hurst exponent begins to decrease. Time series X with a marked point in time after which the Hurst exponent changes (c)

ΔL -method

Scalegrams obtained using a continuous wavelet transform are used to visualize the features of the time series. In [Lande, 2009], another visualization method is proposed, which also helps to reveal trends, periodicities, and local features of the series. The proposed approach is much easier to implement than wavelet analysis.



a b

Figure 39 – Estimation of the Hurst exponent for the second part of the series X. Dependence of statistics R/S on time in a logarithmic scale (a). Time dependence of the Hurst exponent (b).

The method, which the authors called Δ - the method, is based on the DFA (Detrended Fluctuation Analysis) method, which will also be discussed below. The essence of the approach is to determine and display the absolute deviation of the points of a series of accumulated values from the corresponding values of the linear approximation.

Let's describe Δ the method in more detail. To begin with, let's fix a certain window width s (the scale on which the row is viewed). Consider a point x_l and choose a width window for it s so that the point l is in the center of this window (or offset by 1 if seven). Let us construct a linear approximation by the points of the window and denote $L_{l,j,s}$ the value of the local approximation at the point j for the segment centered at l . Next, we calculate the absolute deviation x_j (Fig. 40) from the approximation line $\Delta_{l,j,s} = |x_j - L_{l,j,s}|$.

The method assumes the calculation of values $\Delta_{l,j,s}$ for all points $l = 1, \dots, T$ and windows with a width of $s = 1, \dots, [T/4]$. For a fixed window width, the standard deviation is calculated

$$E(l, s) = \sqrt{\frac{1}{s} \sum_j |x_j - L_{l,j,s}|^2} = \sqrt{\frac{1}{s} \sum_j \Delta_{l,j,s}^2}.$$

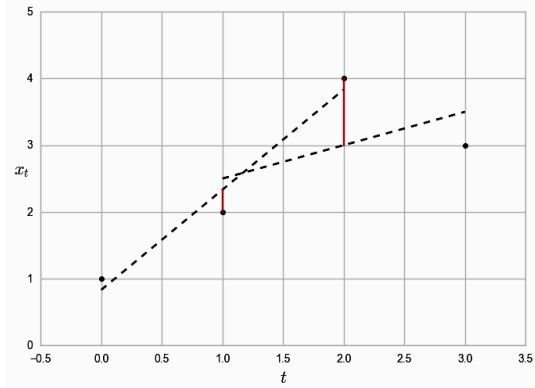


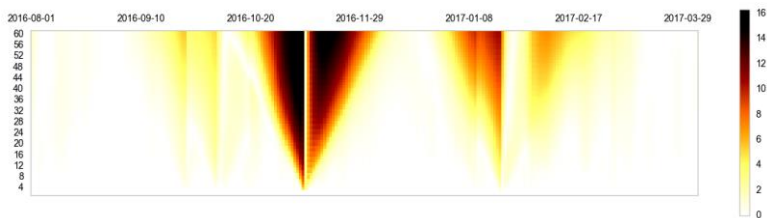
Figure 40 – Four points of the time series with a linear fit for two windows with a width of three. The deviation of the center point of the window from the corresponding linear approximation is also shown $\Delta_{t,j,s}$.

Further, the obtained values are shown on a diagram similar to a scalogram. Examples of such diagrams are shown in Fig. 41.

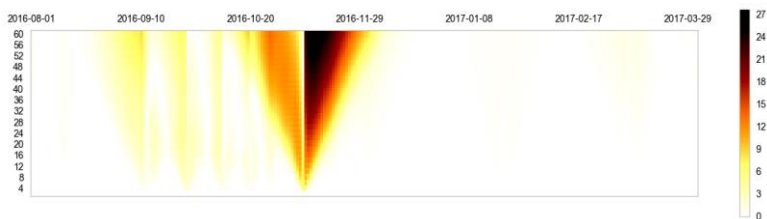
The proposed method of visualization of absolute deviations ΔL , as well as the method of wavelet transforms, allows you to identify single and irregular "bursts", sharp changes in the values of quantitative indicators in different periods of time, as well as harmonic components in a series.

2.6.7. Multifractal analysis

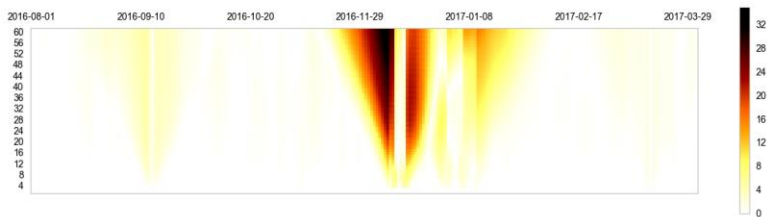
To describe self-similar objects that arise in nature, one fractal dimension is often not enough, since in many cases such objects are not homogeneous. The most general description of the nature of such objects is given by the theory of multifractals, according to which an object is characterized by an infinite hierarchy of dimensions, which makes it possible to distinguish homogeneous objects from heterogeneous ones.



A



b



V

Figure 41 – Coefficients obtained using ΔL the -method for the series T (a), K (b) and X (c).

A multifractal set (signal) can be understood as a kind of union of various homogeneous fractal subsets (signals), each of which has its own value of fractal dimension. The values of such fractal dimensions are displayed in the multifractal spectrum, the formal definition of which will be introduced later. It is important that the multifractal spectrum can be used as a similarity measure. Such an approach can be used, for example, to form representative samples from arrays of documents, as an addition to traditional methods based on identifying the content similarity of documents. Practical applications of this approach: presenting the user with visible search results that reflect the entire spec-

trum of the documentary array or selecting subsets of documents for further research [Lande, 2009a; Lande, 2009b].

In order to consider in detail the idea of a multifractal set and a multifractal spectrum, we need a few additional concepts and definitions. We will use as an example the generalized Cantor set, as well as a measure on this set. The multifractal spectrum in this example appears in a fairly natural and simple way, so it helps to get an intuitive understanding. Also, for the analysis of multifractality, the Hölder exponent is of key importance, the definition of which will be introduced a little later.

We start by constructing a generalized Cantor set and defining a measure on it. The classical Cantor set can be generalized in several ways. For example, you can use the compression function $F_i(x) = rx + (1 - r)i$, where $i = 0, 1$ and $r \in (0, \frac{1}{2})$, instead of $F_i(x) = \frac{1}{3}x + \frac{2}{3}i$. Let's denote such a set $C(r)$. On the generalized Cantor set, one can maintain a uniformly distributed measure, or one can define p – a measure with the same contraction function, but with probabilities p and $1 - p$, where $p \in [0, 1]$.

A more interesting generalization is the Cantor set with variable separation factors. Let there be a sequence $\{r_j\}$, $r_j \in (0, \frac{1}{2})$. The set will be constructed using the following iterative procedure. Let $C_0 = [0, 1]$. Remove from the middle of the segment $[0, 1]$ an open segment of length $1 - 2r_1$. There will be two closed segments of length r_1 . The union of these two segments is denoted by C_1 . At j -th iteration, the set C_j will consist of the union 2^j of closed segments of length $r_1 \cdot r_2 \cdot \dots \cdot r_j$. Thus, we get a set

$$C(\{r_j\}) = \bigcap_{j=1}^{\infty} C_j.$$

Closed segments that appear in the iterative construction of the Cantor set can be encoded using finite words from the alphabet $\{0, 1\}$. At the first step, we get the left segment I_0 and the right segment I_1 . At n -th step, the name of the segment w has a length n , and at the next step, the segment I_w will be divided into segments I_{w0} and I_{w1} .

On the set, $C(\{r_j\})$ you can introduce p -mepy, which will have the following property

$$\mu(I_{w_0}) = p\mu(I_w), \quad \mu(I_{w_1}) = (1 - p)\mu(I_w).$$

We obtain an even more generalized measure for $C(\{r_j\})$, if we use a sequence of weights $\{p_j\}, p_j \in [0,1]$ and define the measure using the following rule $\mu(I_{w_1w_2\dots w_n}) = p_{w_11} \cdot p_{w_22} \cdot \dots \cdot p_{w_nn}$, where $p_{0j} = p_j$, a $p_{1j} = 1 - p_j$.

In order to draw some conclusions about the structure of the measure, we need the Hölder exponent. Let us first consider the definition and meaning of the Hölder exponent for functions and measures, and then apply them to the just constructed generalized Cantor set and measure on it.

Hölder exponent and multifractal analysis for measures

A characteristic of the smoothness of a function is the Hölder exponent, which contains information about the behavior of the function in a neighborhood of a point. The smaller the value of the Hölder exponent, the less smooth the function is.

Let be x a bounded function on \mathbb{R} and $t_0 \in \mathbb{R}$, then the local Hölder exponent of a function x at a point t_0 is defined as

$$h_x(t_0) = \sup_{\Delta t \rightarrow 0} \{\alpha \geq 0: |x(t + \Delta t) - x(t)| = O(\Delta t^\alpha)\}.$$

In other words, the local Hölder exponent characterizes the behavior of a function in a neighborhood of a point as follows

$$|x(t + \Delta t) - x(t)| \sim \Delta t^{h_x(t)}.$$

The last relation is worth comparing with a similar relation for monofractal processes

$$|x(t + \Delta t) - x(t)| \sim \Delta t^H,$$

where H is the Hurst exponent. That is, for multifractal processes, the local Hölder exponent $h_x(t)$ is essentially a "local Hurst exponent", which can vary depending on t .

In order to measure the regularity of a measure in a neighborhood of a point, the Hölder exponent is also introduced.

Hölder exponent or local measure dimension h_μ on \mathbb{R} defined as follows

$$h_\mu(x) = \lim_{r \rightarrow 0^+} \frac{\log[\mu(B(x, r))]}{\log r},$$

where $B(x, r)$ is a ball centered at point x and radius r .

Let us return to the examples with the Cantor set. For a uniformly distributed measure on the classical Cantor set:

$$h_\mu(x) = \frac{\log 2}{\log 3}, x \in C.$$

If we consider a set $C(\{r_i\})$ with a measure μ defined by weights p and $1 - p$, then the local dimension of the measure will differ at different points. For example,

$$\begin{aligned} h_\mu(0) &= \lim_{j \rightarrow \infty} \frac{\log p^j}{\log \prod_{i=1}^j r_i} = \frac{\log p}{\lim_{j \rightarrow \infty} \frac{1}{j} \sum_{i=1}^j \log r_i} = \frac{\log p}{\log r_0}, h_\mu(1) \\ &= \frac{\log(1 - p)}{\log r_0}, \end{aligned}$$

where the notation was introduced $\lim_{j \rightarrow \infty} \frac{1}{j} \sum_{i=1}^j \log r_i = \log r_0$. If we assume that $\sup_i r_i < 1/2$, and assume without loss of generality that $p < 1 - p$, then we can show that

$$h_\mu(x) \in \left[\frac{\log p}{\log r_0}, \frac{\log(1 - p)}{\log r_0} \right], x \in C(\{r_i\}).$$

(proof [Aldroubi 2016]). Thus, the possible values of the local dimension are known for p – the Cantor measure μ . In order to describe some properties of a measure μ , it is necessary to consider sets of the level of values of the local dimension, namely, sets of the form

$$E_h = \{x \in \mathbb{R}: h_\mu(x) = h\}.$$

Next, we can compare the sizes of the sets E_h at different values h . In many cases of practical importance, it will be necessary to use the fractal dimension to compare such sets. Thus, we come to the definition of a multifractal spectrum.

Multifractal spectrum measures μ on \mathbb{R} is called mapping $d_\mu(h) = D_H(E_h)$.

That is, using the multifractal spectrum, it is displayed which values of the Hölder exponent are present in an inhomogeneous object (measure, set, signal), and in what relation to each other. Each value of the Hölder exponent corresponds to the fractal dimension of the set of points where the value of the Hölder exponent is equal to the given one (Fig. 42).

For p – the Cantor measure, μ it can be shown that (proof [Al-droubi 2016])

$$d_\mu(h) = \inf_{q \in \mathbb{R}} \left(qh - \frac{\log(p^q + (1-p)^q)}{\log r_0} \right).$$

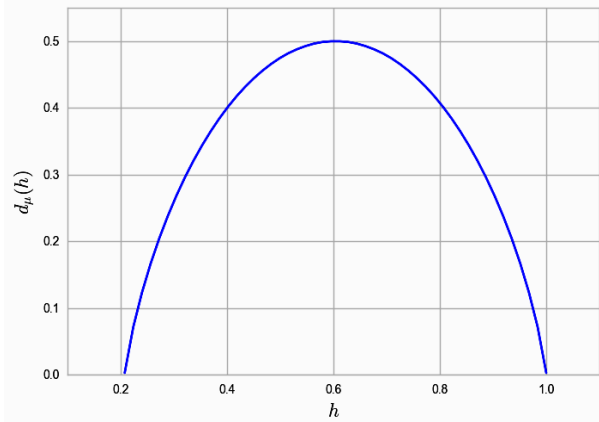


Figure 42 – Multifractal spectrum for p -Cantor measure.

Approach to Multifractal Spectrum Estimation

The theoretical approach to determining the multifractal spectrum for a measure was described above. For practical purposes, the direct calculation of the Hölder exponent at each point and the calculation of the fractal dimensions of the level sets of this exponent is not feasible. The key to the practical approach is the following definition of the structure function of a measure μ

$$Z(q, j) = \frac{1}{2^j} \sum_i \mu \left(B \left(\frac{i}{2^j}, \frac{1}{2^{j+1}} \right) \right)^q,$$

where the sum is taken only over such segments where the measure is not equal to 0. We also define a scaling function

$$\tau(q) = \liminf_{j \rightarrow +\infty} \left(\frac{\log \left(Z_\mu(q, j) \right)}{\log 2^{-j}} \right).$$

It is known that in order to cover the set one E_h needs approximately $2^{-d_\mu(h)j}$ balls and $h_\mu(x) = \lim_{j \rightarrow \infty} \frac{\log[\mu(B(x, 2^{-j}))]}{\log 2^{-j}}$, it follows from the definition that $\mu(B(x, 2^{-j})) \sim 2^{-h_\mu(x)j}$, so the scale function can be estimated as follows

$$Z(q, j) \sim 2^{-j} \sum 2^{-h_\mu(q)j} \sim 2^{-j} 2^{-d_\mu(h)j} 2^{-hqj} = 2^{-(1+d_\mu(h)+hq)j}.$$

On the other hand, the scaling function is defined in such a way that $Z(q, j) \sim 2^{-\tau(q)j}$, therefore

$$2^{-(1+d_\mu(h)+hq)j} = 2^{-\tau(q)j}.$$

And at $j \rightarrow \infty$

$$d_\mu(h) = \inf_{q \in \mathbb{R}} (1 - \tau(q) + hq).$$

Thus, we have obtained expressions for the multifractal spectrum in terms of the scale function. This approach allows one to numerically determine the multifractal spectrum for time series. First, the structure function is determined, with the help of it the scale function is determined, and then, through the Legendre transformation, the transition to the multifractal spectrum takes place.

Multifractal processes

A stochastic process is called multifractal if it has stationary increments and satisfies the equality

$$\mathbb{E}[|x(t)|^q] = c(q)t^{\tau(q)+1},$$

for some positive q , where $\tau(q)$ is the scale function.

If the scale function $\tau(q)$ depends linearly on q , then the process is called monofractal. If the process $x(t)$ self-similar with the Hurst exponent H , then $\tau(q) = Hq - 1$.

DFA method and its application to multifractal spectrum estimation

In [Peng, 1994], Detrended Fluctuation Analysis (DFA) is proposed for determining long-term correlations in noisy and non-stationary time series. The key feature of the DFA method is that it is based on the theory of random walks. The time series is not analyzed in its original form, instead, the series is centered and the transition to the accumulated sums is performed

$$y_t = \sum_{k=1}^t x_k.$$

In such a case, it can be considered y_t as a position of a random walk after t the steps. Further, the DFA method involves the analysis of the standard deviation of the values of the series from the trend on different non-intersecting pieces of the series.

For the DFA method, many modifications have been proposed, as well as applications for various practical problems. An overview of such methods is given, for example, in [Kantelhardt 2009]. An important step was the development of an approach to the numerical estimation of the multifractal spectrum based on the DFA method. This method is called Multifractal Detrended Fluctuation Analysis (MF-DFA) and was proposed in [Kantelhardt 2002]. The effectiveness of the MF-DFA method was analyzed for various model time series (Brownian motion, fractional Brownian motion, binomial cascades) [Oswiecimka 2012]. The method is also actively used to analyze real time series, often economic ones [Suarez-Garcia 2013].

A detailed step-by-step description of the MF-DFA algorithm is given in [Thompson 2016]. Let's describe all these steps.

Step 1. Bringing the time series to an aggregated form. Distinguish between aggregated and disaggregated datasets. An example of aggregated data is the number of new messages on the Internet on a certain topic per day. The corresponding disaggregated data is the increase in the number of messages compared to the previous day. If the original time series $\{z_t\}_{t=1}^{T+1}$ is aggregated, then you need to switch to disaggregated $\{y_t = z_{t+1} - z_t\}_{t=1}^T$. The time series that will be used in the algorithm when centering and calculating the accumulated amounts is as follows:

$$x_t = \sum_{k=1}^t (y_k - \bar{y}), \quad t = 1, \dots, T.$$

This stage of processing the series is necessary for the correct operation of the method, since it is based on the theory of random walks.

Step 2. Set the set $\mathcal{S} = \{3, [N/4]\}$. For each value, $s \in \mathcal{S}$ we divide the time series $\{x_k\}_{k=1}^N$ into $N_s = \lfloor \frac{N}{s} \rfloor$ non-overlapping parts of length s . If N is not divisible by s , then you need to repeat the procedure, starting from the other side of the time series, and the result will be $2N_s$ parts.

Step 3. For each value, $j = 1, \dots, N_s$ – a part of the time series consists of values $\{x_{(j-1)s+i}\}_{i=1}^s$, and similarly for $j = N_s + 1, \dots, 2N_s$ – $\{x_{N-(j-N_s)s+i}\}_{i=1}^s$. For each j – a part of the time series, you need to determine the trend $X_j(i)$. In many cases, it is sufficient to use a linear approximation of the series values obtained using the least squares method. For sufficiently long series, a polynomial approximation of the degree m is used (an algorithm for selecting the optimal parameter m will be given below). Now let's calculate the standard deviation of the part of the time series from the trend

$$F^2(j, s) = \frac{1}{s} \sum_{i=1}^s [x_{(j-1)s+i} - X_j(i)]^2, \quad \text{при } j = 1, \dots, N_s,$$

and similarly for $j = N_s + 1, \dots, 2N_s$.

Step 4. Denote \mathcal{Q} the set of moment order values. The set \mathcal{Q} must contain 0, positive and negative values. Usually, a set that is symmetric with respect to 0 is chosen. For fixed values of $s \in \mathcal{S}$ and $q \in \mathcal{Q}$ we calculate the norm l_q for the vector consisting of the estimated variances $\{F^2(j, s): j = 1, \dots, 2N_s\}$

$$F_q(s) = \left(\frac{1}{2N_s} \sum_{j=1}^{2N_s} [F^2(j, s)]^{q/2} \right)^{1/q}, \quad q \in \mathcal{Q} \setminus \{0\},$$

$$F_0(s) = \exp \left\{ \frac{1}{4N_s} \sum_{j=1}^{2N_s} \ln[F^2(j, s)] \right\}.$$

Step 5. For each $q \in Q$, you need to perform a linear approximation of the dependence $\ln[F_q(s)]$ on $\ln(s)$. In this case, the slope of the resulting linear function is an estimate for $h(q)$. The scale function $\tau(q)$ is obtained from the expression $\tau(q) = qh(q) - 1$.

Step 6. Estimate the derivative of the obtained estimate of the function $\tau(q)$

$$\alpha_0 = \left. \frac{d\tau(q)}{dq} \right|_{q=q_0}, \quad q_0 \in Q,$$

and as a result we obtain an estimate of the multifractal spectrum

$$f(\alpha) = q_0 \alpha_0 - \tau(q_0).$$

Examples of scale functions and multifractal spectra for the T, K, and X series obtained using the MF-DFA method are shown in Fig. 43.

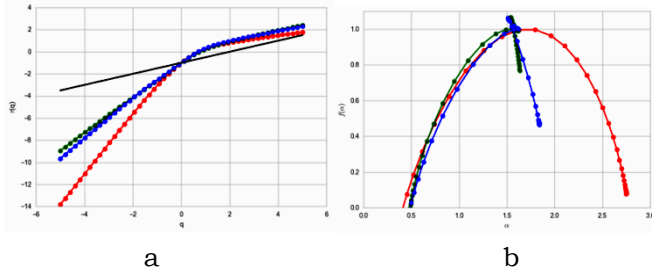


Figure 43 – Scale functions (a) and multifractal spectra (b) for the T, K and X series obtained using the MF-DFA method

For the correct operation of the above algorithm, it is necessary to select in advance the set of S lengths of the segments into which the series is divided. Let's find out what the set should be S . On the one hand, if we use the values $s \geq N/4$, then we will divide the series into a small number of parts, and the estimate $F_q(s)$ will be calculated from a small number of variance estimates $F^2(j, s)$. On the other hand, if you use polynomial regression, then at $s \leq 10$ step 3, few points will be used for polynomial regression. So a reasonable limit is: $10 \leq s \leq N/4$. In this case,

in the case of linear regression, small values can also be used $s = 3, 4, \dots$

For long time series that contain more than two thousand values, it is recommended to use restrictions

$$s_{min} = \max\{10, N/100\}, \quad s_{max} = \min\{20s_{min}, N/10\},$$

and choose the step so that the set \mathcal{S} contains no more than 100 values. Also, the degree m must be chosen in such a way that the regression polynomial $X_j(i)$ adequately displayed the trend in each piece of the time series.

In the case of short time series such as T, K and X, it is sufficient to use linear regression.

2.6.8. network models

Recently, a separate scientific direction has emerged – the analysis of social networks (SNA, Social Networks Analysis), which is based, on the one hand, on sociology, and on the other hand, on the theory of complex networks (Complex Networks) [Newman, 2003]. Within the framework of the theory of complex networks, network characteristics are studied not only from the point of view of network topology, but also statistical phenomena, the distribution of weights of individual nodes and edges, the effects of percolation and conduction. Despite the fact that various networks (electrical, transport, information) fall into the consideration of the theory of complex networks, the studies of social networks made the greatest contribution to the development of this theory [Lande et al., 2009]. In the theory of complex networks, there are three main areas:

- study of statistical properties that characterize the behavior of networks;
- creation of network models;
- predicting the behavior of networks when changing structural properties.

Network settings

In applied research, such characteristics typical of network analysis as network size, network density, degree of centrality, etc., are most often used. In the analysis of complex networks, as in graph theory, the following are investigated:

- parameters of individual nodes;
- network parameters in general;
- network substructures.

For individual nodes, the following parameters are distinguished:

- the input degree of a node is the number of graph edges that enter the node;
- the output degree of a node is the number of graph edges that exit the node;
- the distance from this node to each of the others;
- the average distance from a given node to others;
- eccentricity – the largest of the geodesic distances (minimum distances between nodes) from a given node to others;
- mediation (betweenness), showing how many shortest paths pass through a given node;
- centrality – the total number of connections of a given node in relation to others.

To analyze the network as a whole, parameters such as:

- number of nodes;
- number of ribs;
- geodetic distance between nodes;
- average distance from one node to others;
- density – the ratio of the number of edges in the network to the possible maximum number of edges for a given number of nodes;
- the number of symmetric, transitive and cyclic triads;
- network diameter – the largest geodetic distance in the network, etc.

There are several topical problems of the mathematical study of social networks, among which the following main ones can be distinguished:

An important characteristic of the network is the node degree distribution function $P(k)$, which is defined as the probability that a node i has a degree $k_i = k$. Networks characterized by different $P(k)$, exhibit different behavior, $P(k)$ in some cases may be

Poisson distributions ($P(k) = e^{-m} m^k / k!$), where m —, exponential ($P(k) = e^{-k/m}$) or power ($P(k) \sim 1/k^\gamma$, $k \neq 0$, $\gamma > 0$).

Networks with a power-law distribution of node degrees are called scale-free. It is scale-free distributions that are often observed in real social networks. With a power-law distribution, the existence of nodes with a very high degree is possible, which is practically not observed in networks with a Poisson distribution.

The distance between nodes is defined as the number of steps that must be taken to get from one node to another along existing edges. Naturally, nodes can be connected directly or indirectly. The shortest path d_{ij} between nodes i and j is the shortest distance between them. For the entire network, you can introduce the concept of the average path, as the average over all pairs of nodes of the shortest distance between them:

$$l = \frac{2}{n(n-1)} \sum_{i > j} d_{ij},$$

where n is the number of nodes, d_{ij} — the shortest distance between nodes, i and j .

Hungarian mathematicians P. Erdős and A. Rényi showed that the average distance between two vertices in a random graph grows as a logarithm of the number of its nodes [Erdős, 1960].

Some networks may be disconnected, i.e. there are nodes, the distance between which is infinite. Accordingly, the average path may also be equal to infinity. To take into account such cases, the concept of the average inverse path (it is also called "network efficiency") between nodes is introduced, calculated by the formula:

$$il = \frac{2}{n(n-1)} \sum_{i > j} \frac{1}{d_{ij}}.$$

Networks are also characterized by such a parameter as the diameter or the maximum shortest path, which is equal to the maximum value of all d_{ij} .

D.Watts and S.Strogatz in 1998 defined such a network parameter as the clustering coefficient [Watts, 1998], which characterizes the level of connectivity of nodes in the network, the tendency to form groups of interconnected nodes, so called cliques. In addition, for a particular node, the clustering coefficient indicates how many nearest neighbors of a given node are also nearest neighbors to each other.

The clustering coefficient can be determined both for each node and for the entire network. For a network, the clustering coefficient is defined as the sum of the corresponding coefficients for individual nodes normalized by the number of nodes.

The clustering coefficient for an individual network node is determined as follows. Let a node come out k with edges that connect it with k other nodes, nearest neighbors. If we assume that all nearest neighbors are connected directly to each other, then the number of edges between them would be $1/2 \cdot k(k-1)$. Those. this is the number that corresponds to the maximum possible number of edges that could connect the nearest neighbors of the selected node. The ratio of the real number of edges that connect the nearest neighbors of a given node to the maximum possible (such that all the nearest neighbors of a given node would be connected directly to each other) is called the node clustering coefficient $i-C(i)$. Naturally, this value does not exceed unity.

Intermediation (betweenness) is a parameter showing how many shortest paths pass through the node. This characteristic reflects the role of this node in establishing links in the network. The nodes with the most mediation play a major role in establishing links between other nodes in the network. Node m mediation b_m is determined by the formula:

$$b_m = \sum_{i \neq j} \frac{B(i, m, j)}{B(i, j)},$$

where $B(i, j)$ is the total number of shortest paths between nodes i and j , $B(i, m, j)$ is the number of shortest paths between nodes i and j passing through node m .

One can speak of a “community structure” when there are groups of nodes that have a high density of edges between them-

selves, while the density of edges between individual groups is low. The traditional method for identifying the structure of communities is cluster analysis. There are dozens of acceptable methods for this, which are based on different measures of distances between nodes, weighted path indices between nodes, and so on. In particular, for large social networks, the presence of a community structure turned out to be an integral property.

The properties of real social networks also include the so-called "weak" ties. An analogue of weak social ties are, for example, relationships with distant acquaintances and colleagues. In some cases, these ties are more effective than "strong" ties. For example, a group of researchers from the UK, the US and Hungary came up with a conceptual conclusion in the field of mobile communications that "weak" social ties between individuals are the most important for the existence of a social network [Bjorneborn, 2004].

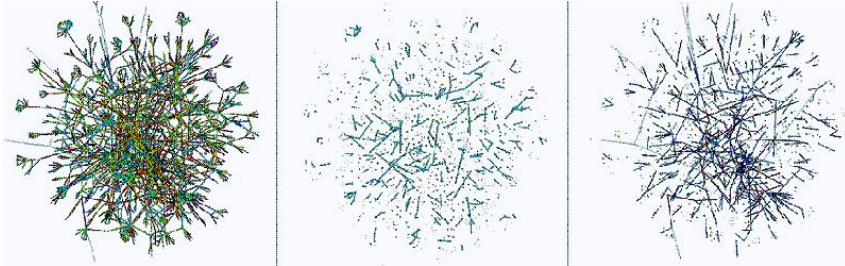
For the study, calls were analyzed from 4.6 million mobile subscribers, which is about 20% of the population of one European country. This was the first time in the world practice when it was possible to obtain and analyze such a large sample of data related to interpersonal communication.

In a social network with 4.6 million nodes, 7 million social connections were identified, i.e. mutual calls from one subscriber to another and back, if return calls were made within 18 weeks. The frequency and duration of conversations were used to determine the strength of each social connection.

It was revealed that it is weak social ties (one or two callbacks over 18 weeks) that tie together a large social network. If these connections are ignored, the network will fall apart into separate fragments. If strong connections are not taken into account, then the connectivity of the network will be broken (Fig. 44). It turned out that it is weak ties that are the phenomenon that binds the network into a single whole.

Despite the huge size of some social networks, many of them have a relatively short path between any two nodes – the geodesic distance. In 1967 psychologist S. Milgram, as a result of large-scale experiments, calculated that there is a chain of acquaintances, on average, six links long, between almost any two US citizens [Milgram, 1967].

D. Watts and S. Strogatz discovered a phenomenon that is characteristic of many real networks, called the effect of small worlds (Small Worlds) [Watts, 1998]. In the study of this phenomenon, they proposed a procedure for constructing a visual model of the network, which is inherent in this phenomenon.



1 2 3

Fig. 44 – Network structure:

- 1) a complete map of the social communications network;
- 2) a social network from which weak ties have been removed;
- 3) a network from which strong ties are removed: the structure retains connectivity

The three states of this network are shown in Fig. 45: a regular network – each node of which is connected to four neighboring ones, the same network in which some "near" connections are randomly replaced by "far" ones (in this case, the phenomenon of "small worlds" occurs) and a random network in which the number of similar substitutions has exceeded a certain threshold.

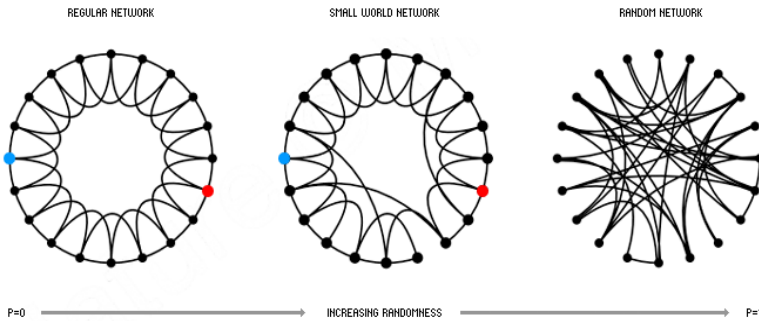


Figure 45 – Watts-Strogatz Model

In reality, it turned out that it is precisely those networks whose nodes have simultaneously a certain number of local and random “distant” connections that simultaneously demonstrate the effect of a small world and a high level of clustering.

On Fig. 46 shows graphs of changes in the average length of the path and the clustering coefficient of the artificial network by D. Watts and S. Strogatz on the probability of establishing "distant connections" (on a semi-logarithmic scale).

For example, WWW is a network for which the phenomenon of small worlds is also confirmed. An analysis of the topology of the web by S.Zhou and R.J.Mondragon of the University of London showed that nodes with a high degree of outbound hyperlinks have more links to each other than to nodes with a low degree, while the latter have more connections with nodes with a large degree than among themselves. This phenomenon has been called the "rich-club phenomenon". The study found that 27% of all connections are between just the top 5% nodes, 60% are connections between the other 95% nodes with the top 5%, and only 13% are connections between nodes that are not in the top 5%.

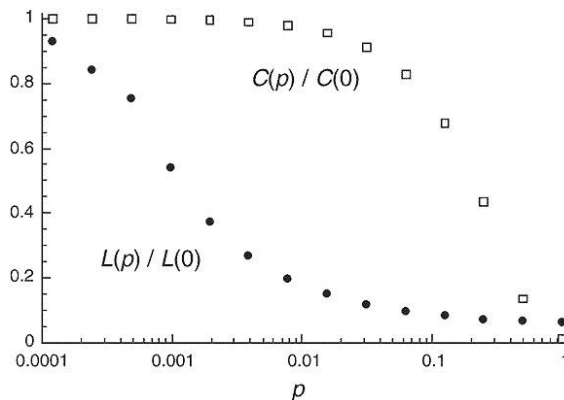


Figure 46 – Dynamics of changes in the length of the path and the clustering coefficient in the Watts-Strogatz model in a semi-logarithmic scale (axis OX – the probability of replacing short-range bonds with distant ones)

These studies give reason to believe that the dependence of the WWW on large nodes is much more significant than previously assumed; it is even more susceptible to malicious attacks. Related to the concept of "small worlds" is also a practical approach called "network mobilization", which is implemented over the structure of "small worlds". In particular, the speed of information dissemination due to the effect of "small worlds" in real networks increases by orders of magnitude compared to random networks, because most pairs of nodes in real networks are connected by short paths.

Practice has proven [Rothenberg, 2002] that terrorist networks are most often not only scale-free, but also exhibit the properties of small worlds, i.e. that the presence of tightly connected clusters (groups of tightly connected nodes) ensures local communication even in cases of successful attacks, when hubs (the largest intermediaries) fail.

In the study of "small worlds", an interesting approach was determined, logically related to the concept of percolation (flow) [Broadbent, 1957], [Snarsky, 2007]. It turns out that many of the questions that arise when analyzing network security on the Internet are directly related to this theory. The simplest formulation of the problem of percolation theory, cleared of all physical and mathematical layers, is as follows: "Given a network, a random part of the edges of which conducts a signal, and the other part does not. The main question is what is the minimum concentration of conducting connections at which there is still a path through the entire network? The tasks that are solved in the framework of the theory of percolation and network analysis include such as determining the limiting level of conductivity, changing the length of the path and its trajectory when approaching the limiting level of conductivity, the number of nodes that need to be disabled in order to disrupt the connectivity of the network.

Security experts have recently increasingly associated the effect of "small worlds" with the networks of terrorist organizations, the so-called overlay networks, i.e. networks built on top of the Internet.

Analyzing links in a network, one can learn about its important properties, for example, identify the presence of clusters,

determine their composition, differences in connectivity within and between clusters, identify key elements that connect clusters to each other, etc. At the same time, a serious obstacle in the analysis is incomplete information about the links between individual network nodes.

Recently, a group of researchers from the Santa Fe Institute presented an algorithm that makes it possible to automatically obtain information about the hierarchical structure of such networks [Clauset, 2008]. A new method of restoring ties can be used by both special services and competitive intelligence units of companies. So, for example, knowing only about half of the connections between terrorists, it will be possible with a high probability to restore the missing links of the entire chain.

Even without a complete description of the system, you can get a representative sample of connections and try to complete the entire network using it. Analysis of the resulting graph allows you to identify potentially important connections that could not be found in a real network. For example, having information about only half of the contacts of network participants with each other, it is possible with a probability of 0.8 to predict those connections about which nothing was known at first. Obviously, this method can be very useful in identifying hidden network groups, and thus put the matter of ensuring state and commercial security on a qualitatively new level.

To analyze complex networks of concepts mentioned in individual documents from information flows, methods of in-depth text analysis can be used, or rather, content monitoring and extraction of concepts such as persons, companies, toponyms (geographical names), etc.

One of the areas of analysis of social networks is visualization, which is important because it often allows you to draw important conclusions about the nature of the interaction of subjects-nodes without resorting to precise methods of analysis. When displaying a network model, it may be useful to:

- placement of network nodes in two dimensions;
- spatial ordering of objects in one dimension in accordance with some quantitative properties;

- the use of methods common to all network diagrams to display the quantitative and qualitative properties of objects and relationships.

Network signs of information operations

As an extension of the above multi-agent information dissemination model, we can consider a model that takes into account the structure of the network being formed [Pugachev, 2015]. Within the framework of this model, each agent – the source of information – does not have a "potential", but a certain rating (which corresponds to the size of the corresponding node in the diagrams). The links in the network under consideration are facts of reprinting or "retelling".

The model is based on the assumption that when carrying out information operations, the most rated sources reprint information from the least rated ones, or clusters of low-rated publications reprint the same news.

On Fig. 47 shows an example of typical information operations identified within this model.

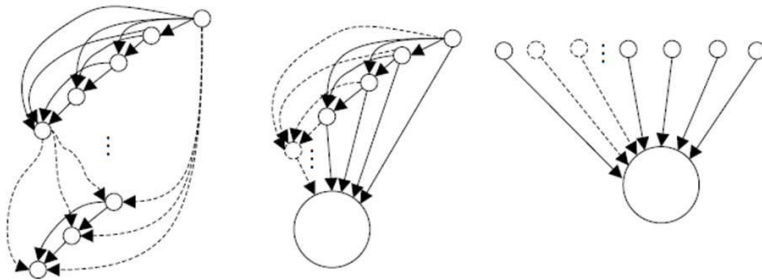


Figure 47 – Examples of information dissemination networks with signs of information operations

As part of the formalization of the same model, several dozen parameters of the topology of information distribution networks are selected, such as diameter, density, clustering, mediation, etc., which are compared with some reference values.

The advantages of this model include its formal rigor and compliance with the recently actively developing direction of Complex Networks, which allows us to expect its further devel-

opment. The disadvantages should, apparently, include a low correlation with the content side of recognizable information operations, as well as a certain computational complexity in identifying fuzzy informational duplicates of documents.

Technological stages of the study of the mutual influence of information sources

For an effective study of the mutual influence of information sources from the Internet (web resources, social media), a sequence of steps, stages of information processing is proposed, each of which in itself provides an analytical product. The set of such stages, which are based on the use of the necessary and available tools, special techniques, can be considered as a procedure for carrying out actions aimed at obtaining analytical materials, including the construction and analysis of a network of their mutual influence.

When conducting information and analytical research based on content monitoring, such tasks include:

- Finding relevant publications on a given topic.
- Identification of mutual contextual references and reprints in documents presented by various information sources.
- Building a network of influence, analysis and visualization of the relationship of information sources, including ranking the nodes of the constructed network according to the degree of influence.
- In identifying possible information operations and building a scenario for counteracting information operations in a network environment.

Obtaining a representative array of publications

To obtain a representative array of publications on a selected topic, it is necessary to choose a content monitoring system that provides a flow of information messages on a specific topic. The topic can be expressed by a query in the language of the information retrieval system.

The authors chose the InfoStream system as a content monitoring system, which currently covers 10,000 sources of infor-

mation in Russian and Ukrainian. More than 100,000 documents enter the system's databases daily. The InfoStream system provides search, as well as viewing the list and full texts of relevant documents.

In the one shown in Fig. 48 example shows a fragment of the system interface through which a request was processed related to the discussion in January 2016 of the issue of the resignation of the Prime Minister of Ukraine A. Yatsenyuk.

As a result, a thematic information array was formed, which covers over 3 thousand documents.

Definition of contextual links

The basis for building a network of influence of information sources is contextual links and reprints in the thematic information flow. Contextual links are identified by identifying patterns in the documents of the selected information array and signs of exact reprints, determined by plagiarism detection methods. In turn, the templates themselves are periodically determined / supplemented by experts in an automated mode by analyzing the context of the document flow of the content monitoring system using Text Mining methods.

Building a network of influence of information sources

The contextual references and reprints found in the texts make it possible to form a citation matrix, transposing which, in accordance with the above hypothesis, can form an influence matrix. This matrix corresponds to the source influence network, an example of visualization of which for the thematic information array considered above using the Gephi system is shown in Fig. 49.

Активная база данных: Система интеграции интернет-ресурсов

Главная Помощь Кабинет Источники Статистика Новости проекта

Вход Выход

InfoStream Online

Яценюк отставка

Период: Другой Убрать дубли Морфология Постранично

От: 201601 До: 201601

Найти Динамика Дайджест

Очистить События Сожеты

Язык запросов Примеры

Яценюк отставка

Найдено документов - 3196, страница 1 из 214

Статистика слов:

ЯЦЕНЮК - 77404, ОТСТАВКА - 46651

Добавить канал

1. БПП не против Яценюка на посту премьера

Коррелирует дат 2016.01.31 23:46

Фото: АР Премьер-министр Украины Арсений Яценюк В БПП назвали условия сохранения за Яценюком поста премьер-министра. Лидер фракции Блок Петра Порошенко Юрий Луценко заявил, что фракция выступает за сохранение поста премьера за Арсением Яценюком, если под его руководством будет сформирован новый Кабинет министров.

Дубли - Похожие документы - Оригинал

2. Фракция БПП не будет требовать отставки Яценюка

ІСТУ Факти 2016.01.31 23:42

Кабин не состоялся как коллектив - Луценко Фракция БПП готова продолжать работать с нынешним премьер-министром Арсением Яценюком для сохранения политической стабильности в Украине, заявил глава фракции Юрий Луценко, передает УНІАН. - В первую очередь необходимо говорить о корректировке коалиционного соглашения и программы действий правительства.

Дубли - Похожие документы - Оригинал

3. Луценко объяснил, почему для БПП Яценюк как премьер является приемлемым

Буквы 2016.01.31 23:36

Автор статьи #Б/К/В/Ы Председатель фракции "Блок Петра Порошенко" в Верховной Раде Юрий Луценко прогнозирует досрочные выборы в парламент в случае, если премьер-министр Арсений Яценюк и, соответственно, весь Кабинет Министров будет отправлен в отставку. Об этом Луценко сообщил в эфире

Информационный портрет

Уточнить запрос

Рубрики (29)
Языки (2)
Страны источников (17)
Источники (50)
Размер (3)
Цифровая насыщенность (2)
Тональность (2)
География (50)
Персоны (50)
Компании (50)
Слова (50)
Классификатор-навигатор
ОТСТАВКА

Figure 48 – Fragment of the interface of the content monitoring system

Study of the network of influence of information sources

The constructed network of influence of information sources can be investigated using generally accepted tools (for example, using the Gephi system, the following parameters of the constructed network were obtained, such as the number of nodes: 141, edges 196, graph density: 0.01, average clustering coefficient: 0.026, average path length: 1.26, etc.).

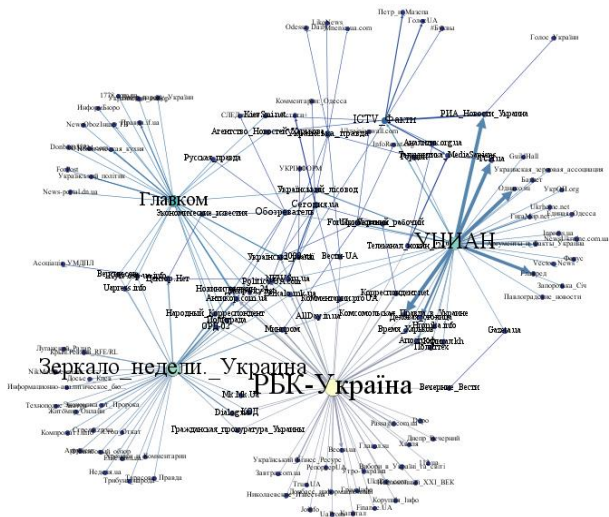


Figure 49 – Fragment of the network of links of sources on the selected topic

For meaningful analysis, the weight of network nodes is of great importance. The following list of the most important nodes was obtained according to the criterion of output power:

No.	Web resource	output degree
1	RBC-Ukraine	50
2	Mirror of the week	38
3	UNIAN	35
4	commander-in-chief	28
5	ICTV Facts	10
6	Today.ua	7
7	Ukrainian truth _	7
8	Reviewer	6
9	Forbes-Ukraine	4
10	Censor. No	3

A promising approach to ranking sources by the level of influence is the HITS algorithm proposed by J. Kleinberg [Kleinberg, 2006].

The HITS algorithm selects from the network the best "authors" (the nodes to which the links will be entered) and "intermediaries" (the nodes from which the inclusion links come).

In accordance with the HITS algorithm, for each network node, its significance as an author $a(v_j)$ and an intermediary (hub) is recursively calculated $h(v_j)$ using the formulas:

$$a(v_i) = \sum_{j \rightarrow i} h(v_j);$$
$$h(v_i) = \sum_{i \rightarrow j} a(v_j).$$

In these formulas, the sum is over all nodes that refer to (or are referred to) by the given node.

Paraphrasing the notation given in [Kleinberg, 2006], namely, replacing "authorship" with "exposure" and "mediation" with "influence", it is possible to calculate the corresponding characteristics of the influence network nodes with little computational effort.

Also, to identify information impacts, the definition of "hidden" links is of great importance. The methodology for determining hidden connections, hidden influences, in particular, is given in [Snarskii, 2016].

Definition of possible information operations

The network of information impact of information sources allows you to quickly identify possible information operations in accordance with the approaches proposed in [Potemkin, 2015]. It is assumed that there is little likelihood of an information operation if information about an incident first originates in an influential information source and then reprinted (with or without links) by less influential sources (Figure 50). The reverse phenomena, when more influential publications reprint information in less influential, albeit numerous, may be a sign of an information operation, an attack (Fig. 51). It is these patterns that

were observed in the network analysis of real thematic information flows (Fig. 52).

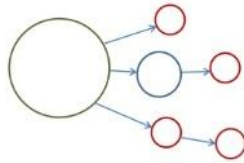


Figure 50 – Typical information dissemination scenario

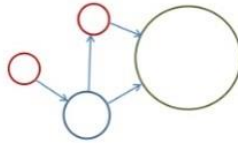


Figure 51 – Information dissemination scenario inherent in an information operation

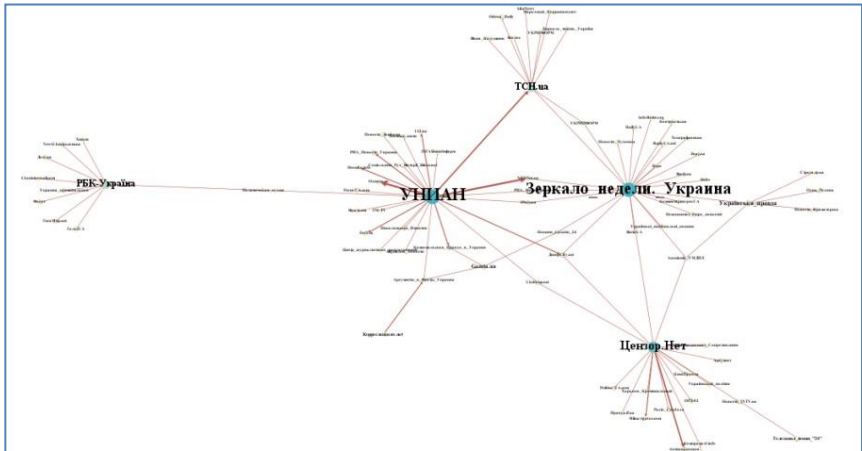


Figure 52 – Fragment of a network of links of sources on special topics

The advantages of this model include its formal rigor and compliance with the recently actively developing direction of Complex Networks, which allows us to expect further development. The disadvantages should, apparently, include a small

correlation with the content side of recognizable information operations, as well as a certain computational complexity in detecting fuzzy informational duplicates of documents.

2.7. Implemented Competitive Intelligence Technologies

The competitive intelligence system should allow the management, analytical, marketing departments of the company not only to quickly respond to changes in the market situation, but also to assess risks and opportunities, predict them and make decisions on further development paths, ensure the transition from traditional intuitive decision-making based on insufficient information to management based on reliable forecasts and knowledge.

One of the main general requirements for a competitive intelligence system should be the compliance of the information processing cycle in such a system with the classical information intelligence cycle. Those. The system must independently or with the participation of the operator provide:

- choice of subjects and areas of reconnaissance (target designation);
- selection of information sources (websites, blogs, forums, etc.);
- automatic search and download of information in the specified areas of monitoring and specified sources according to the planned schedule (planning and data collection);
- processing the collected data and turning it into information;
- content analysis and synthesis of information – its transformation into knowledge;
- timely delivery of information to end users.

Since for the purposes of competitive intelligence it is necessary to analyze data from all available sources of information in which this information can be presented in various forms and formats, an extremely important requirement for the system is to provide it with a single information space of interconnected objects and facts, regardless of the type of their sources or content.

Two other requirements relate to maintaining the connection of objects and facts with relevant data and sources of information (argumentation) and providing a historical-spatial model of the system data bank, which implies that all objects have attributes of time, place and data source, as well as the impossibility of their irretrievable removal from the system over time.

The main objects of accounting and monitoring in competitive intelligence systems, as a rule, are:

- sources of information (official websites, Internet publications, personal websites of organizations or individuals, Internet representations of print media, news agencies, television and radio channels, open databases, accounting objects, etc.);
- geographic regions;
- markets and lines of business;
- structures (enterprises, organizations, etc.);
- persons (competitors, contractors, partners, employees, candidates, etc.);
- normative-legislative base and the facts of its violation;
- political and economic situation;
- criminal situation;
- other specialized topics.

Of course, the competitive intelligence system, which uses the Internet as one of the sources of information, must be adjusted to the specifics of the company's activities. It should include an appropriate classification, flexible search mechanisms, prompt delivery of data, as well as a qualitative assessment of information. One of the most important tasks of information analysis is to determine its reliability, i.e. the task of analyzing and filtering noise and false information. Without such assessments, there is always a risk of making wrong decisions. After analyzing the reliability of information, assessments of its accuracy and importance should follow. The main criterion for the reliability of data in practice is the confirmation of information by other reliable sources.

Even a superficial analysis of the basic requirements for competitive intelligence systems on the Web shows that traditional Internet search engines cannot be considered full-fledged competitive intelligence tools on the Internet.

Competitive intelligence information systems can also be conditionally classified by the presence of automatic and expert fact extraction modules in them. The correlation between automatically retrieved by the system and manually (with the help of experts) facts, events, accounting objects in different systems is different. Facts extracted automatically by the system are called A-facts, facts extracted by experts are called E-facts [Kiselev, 2005].

The existing competitive intelligence systems on the market differ both in their completeness and compliance with the full intelligence cycle, and in their toolkit and, accordingly, in their pFigure In addition, the systems can be designed to be used as a toolkit exclusively by the company's internal competitive intelligence unit, or it can involve outsourcing some of the tasks to specialized competitive intelligence structures. The choice of systems, approaches and methods of competitive intelligence remains with the consumer, and in each case is individual. Yes, this is understandable, you can't compare the needs and tasks performed by an intelligence analyst and an employee, for example, of the marketing department of a small business.

Currently, there are a number of systems in the world that partially implement the solutions of the above tasks of monitoring subjects, extracting facts, building connections, but some of them do not stand up to criticism in terms of functionality, some are too expensive. Let us briefly dwell on the possibilities of some such systems implemented at the present time.

RCO system (www.rco.ru) – the main purpose – the identification of factual information from unstructured texts of large volumes (search for facts in Big data). It has a wide range of algorithms and technologies for intelligent text processing, presented in natural language. In particular, RCO technologies allow solving the problems of identifying named objects, relationships and facts from unstructured data arrays. RCO Fact Extractor is a personal application for Windows, which is designed for analytical processing of text in Russian and revealing facts related to given objects – individuals and organizations. The main area of application of the program is tasks from the areas of competitive intelligence, the fight against corruption, which require high-precision information retrieval.

RCO Zoom (Fig. 62) is a specialized search and analytical system that combines the functionality of traditional search engines with the functions of real-time content analysis and transactional document storage. RCO Zoom has the tools to conduct effective operational search and analytical research of information.

The RCO Zoom system allows you to work with huge arrays of textual information in real time (database size – up to hundreds of gigabytes, search and processing time – seconds). Display tool – information portrait makes it possible to obtain keywords, formulate and test hypotheses, separate objects, highlight statistical invariants in the first approximation. The system can be used as a highly reliable document storage. The system can work with documents in different languages. It is integrated with the RCO Fact Extractor SDK library. The interface for the Python language allows you to implement all kinds of add-ons to solve completely different tasks in addition to storage and search: from finding informational duplicates of documents to their classification and clustering.

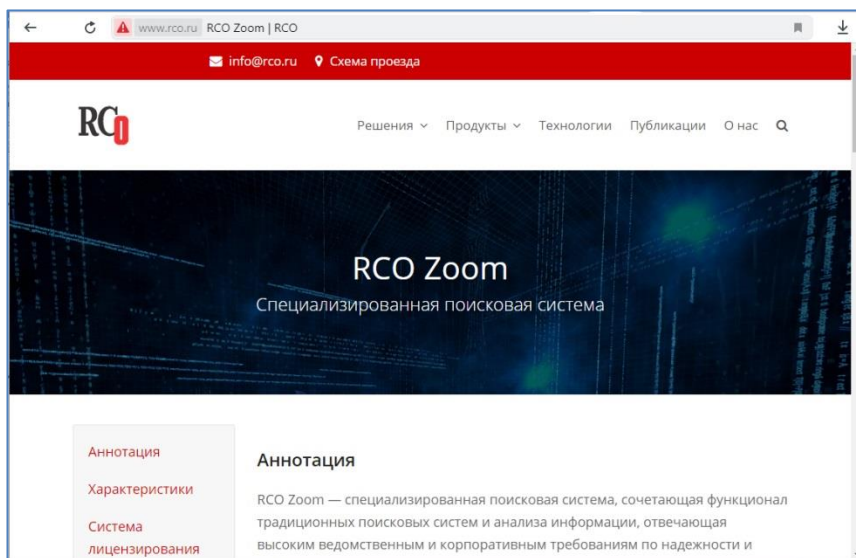


Figure 62 – A fragment of a website with a description of the possibilities

"Medialogy " (www.mlg.ru) is a service that provides online access to the media database with the ability to automatically monitor the media and express analysis of received messages in real time. The system on which the service is built consists of two main parts: database: 56,000 media sources, 900 million social network accounts (as of 06/01/2020); automated analytical module. Medialogia database is updated and replenished on a daily basis. With the help of "Medialogy" you can carry out operational monitoring of the company's media, its top managers, brands, competitors, etc. The filter capabilities allow you to set up monitoring and evaluate the tone of the press, magazines, TV and online publications for almost any informational task.

PolyAnalyst (www.megaputer.ru) is the main product of Megaputer Intelligence. PolyAnalyst is a data analytics software platform that provides a framework for text mining, data mining, machine learning, and predictive analytics. The PolyAnalyst GUI contains various nodes that can be linked into a flowchart to perform an analysis. The software provides nodes for data import, data preparation, data visualization, data analysis, and data export. PolyAnalyst's text analytics features include nodes for text clustering, sentiment analysis, fact, keyword and entity extraction, and taxonomy and ontology generation. Polyanalyst also contains nodes for analyzing structured data and executing Python and R code. As of 2020, the software supports text analysis in 16 languages. PolyAnalyst is commonly used to create custom business tools. It uses a client-server model and is licensed under a software-as-a-service model.

IBM Integrated Analytics System (<https://www.ibm.com/en/products/integrated-analytics-system>) is a comprehensive hybrid data analytics solution that enables massively parallel processing. The solution includes a hardware platform and a software database query system to support the analysis of various types of data. The provided turnkey solutions are customized and tested. The system provides access to data, query and analysis of data in the data warehouse and Hadoop in real time using machine learning tools, provides the ability to develop and improve machine learning models directly on the data storage platform.

mode BI (<https://modusbi.ru/>) – A business intelligence platform that allows you to collect and visualize data from various sources, generate reports and create forecasts for making effective management decisions and monitoring the most important business information. The Modus system collects data from heterogeneous sources, cleans it and prepares it for analysis at a rate of 50 million lines per hour. Modus ETL and Data Quality Management solution. Allows you to collect data from a variety of sources, provides processes for verification, normalization and subsequent formation of a single corporate data warehouse.

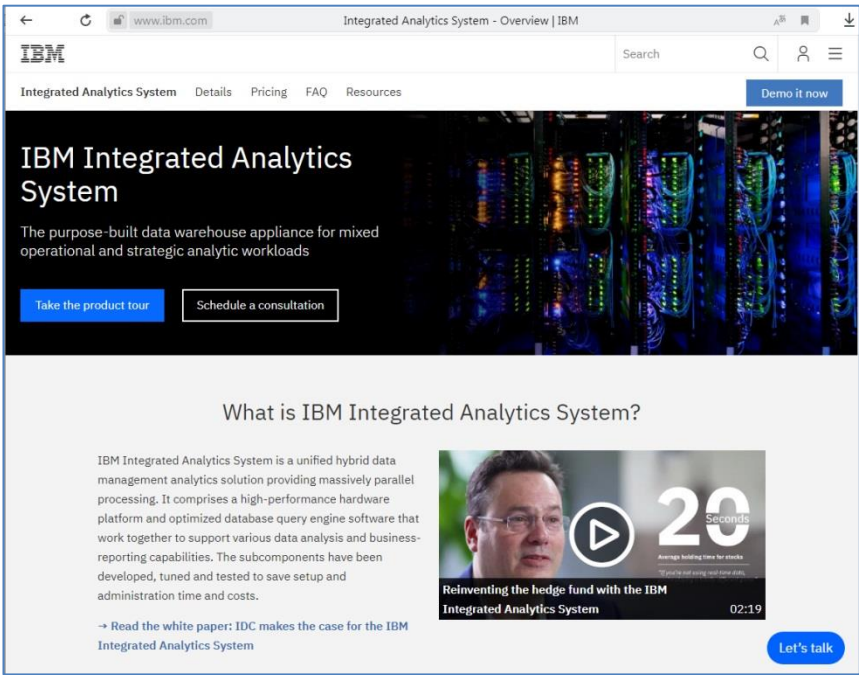


Fig. 63 – Fragment web resource IBM Integrated Analytics System

Oracle Analytics cloud (<https://www.oracle.com/business-analytics/analytics-cloud.html>) and an integrated set of analytics tools designed to provide a better vision and understanding of the business to a wide range of users and allows any user in the organization to get quick web access to up-to-date

information. In Oracle system Analytics Cloud built-in machine learning tools can complement the original data and offer options for their intellectual enrichment. Machine learning tools include automatic explain elements that make it easier to analyze data and create predictive analytics. Embedded machine learning allows you to generate data-driven predictions.

SAP BusinessObjects (<https://www.sap.com/index.html#business-process-intelligence>) – A flexible, scalable business intelligence (BI) system that allows you to discover and share data for effective decision making. SAP BusinessObjects Business Intelligence is a centralized package for real-time reporting, visualization, and data sharing. The system transforms the raw data into useful and accessible analytical information, available anytime and anywhere.

Rocket Folio/NXT (<https://www.rocketsoftware.com/products/rocket-folionxt>) – programs that allow you to identify information elements such as entities, relationships and events in unstructured texts, as well as detect implicit relationships and events in texts. Rocket Folio, an app-based platform, provides reliable search, content management and publishing. Rocket NXT, a server-side search and publishing platform, provides an integrated solution for both structured and unstructured data – from any device, anywhere, anytime.

Recently, all the major Western brands specializing in the development of storages and databases, corporate management systems, have expanded their product lines with Business Intelligence (BI) modules or, literally, business intelligence. Oracle, SAS, SAP, IBM and other brands claim the availability of such modules.

At the request of a group of analysts at Harvard University, Russian developers from the Inforus company created the **Avalanche information and analytical system (IAS)** (www.tora-centre.ru/avl3.htm), designed to monitor changes taking place on the Internet. It collects information from web pages according to a given algorithm and adds this information to its own database.

Avalanche's technology is based on three components: a stand-alone intelligent search robot, smart folder creation, and an embedded database that allows them to be converted into a "personal encyclopedia". When working with the Avalanche IAS,

a model of the area required by the user is formed in the form of a set of "smart folders", each of which "knows" what should fall into it, and ensures that there is no duplication. The filling of "smart" folders is carried out by a specialized search robot, which is launched from a computer in accordance with the settings required by the user. The robot can also start automatically at a certain time set for it. Avalanche provides "fine" settings that allow for more detailed monitoring.

" **Semantic Archive**" (www.anbr.ru) is an analytical tool that allows you to automate the entire technological chain of solving analytical and intelligence tasks, from collecting the necessary information, its intellectual analysis, and ending with convenient reporting. The platform makes it possible to analyze and apply heterogeneous information for the timely adoption of optimal management and business decisions. "Semantic Archive" has a modular structure, which makes it easy to select and configure the desired system configuration.

A flexible ontological data model allows you to work with different topics and areas of activity. IAS "Semantic Archive" allows you to store information imported from various relational databases, enter information from any other sources: the Internet, the media, databases, online libraries and systems (Spark, Integrum, etc.), any document, experts' own information.

The created repository serves analysts to search for information, add their own confidential data, identify the relationship between objects and events, receive analytical reports, visualization: diagrams, graphs and maps.

X-Files dossier management system is a product of the I-Teco company (<https://www.i-teco.ru/>), designed to solve the problem of extracting reliable facts from various sources, filling them in dossiers on monitoring objects and their subsequent analytical processing. X-Files is a dossier management system designed to analyze facts from various sources. The system helps to make decisions in the presence of a large amount of information "gaps" in the formation of a holistic image of the object. It is used to support decision-making processes in the presence of a large amount of "raw" content, which is typical for the activities of public authorities, law enforcement agencies, and large commercial companies.

The X-Files system involves the implementation of three principles:

1) a single information space of interconnected facts or hypotheses, regardless of the type of their content (content of information sources);

2) connection of facts or hypotheses with relevant sources of information (argumentation of facts and hypotheses);

3) historical-spatial information model of the database of facts and hypotheses. This means that there are attributes of time and place for each fact, as well as the impossibility of their permanent removal from the system.

Xfiles implements a semantic network that reflects only the relationships between objects.

IBM security i 2 Analyst's Notebook

(<https://www.ibm.com/products/i2-analysts-notebook>) is a visual data structure design system for storing data about various individuals and organizations.

The database provides for the possibility of storing certain events that occur with them and the existing relationships. IBM system security i 2 Analyst's Notebook allows you to quickly and efficiently analyze a system of interrelated objects and the dynamics of successive events, displaying the results of the study in the form of easy-to-understand diagrams and diagrams. The solution provides features such as connected network visualization, social network analysis, and geospatial or temporal views to help you discover hidden connections and patterns in your data. Information is displayed on the diagram as objects, to which, if necessary, you can add additional attributes and data cards with comments. Objects in the diagram can be represented not only as icons, but also as photos, files, audio recordings, video recordings, etc. The program allows you to create diagrams using queries to relational databases, as well as importing data from external files. Using the functions available in Analyst's Notebook, you can combine diagram elements, look for relationships between them, use the element search system, trace the "path" that connects objects, and so on.

Analyst's Notebook provides a range of useful visualization formats, each of which makes sense of information and shows relationships between objects in a different way. Analyst's Notebook is equipped with an editor that allows you to formulate a

query in a graphical form to search for objects and identify their relationships, create templates of events of interest. The Analyst's Notebook system can be integrated into applications already running on the user's site. The system provides:

- search for common elements and relationships hidden in the data;
- ease of interpretation of complex information;
- graphic display of results;
- creating dynamic charts;
- distribution of diagrams in printed and electronic form.

Speaking about products leading in the field of Business Intelligence, it should be noted that this term, as a rule, refers to a set of tools for analyzing statistical digital data and other corporate reports and their visualization, in contrast to Competitive Intelligence (competitive intelligence), which is a much broader area of information and analytical activities.

On the Ukrainian market in the segment of information and analytical systems of competitive intelligence, such systems as X-SCIF, Encyclopedia of Business Information of Ukraine, Iceberg BI, Strabis-VEB, etc. are presented.

I would like to note that not all of these systems have full functionality and corresponding modules that ensure the implementation of the entire range of competitive intelligence tasks.

one of the most full-featured domestic systems, in which information processing corresponds to the classical information intelligence cycle.

Let's consider how the stages of the reconnaissance cycle are implemented with the help of this system, for which we will dwell on the description of the capabilities of the X-SCIF system a little more.

Online instrumental corporate system for monitoring, aggregation and analysis of information X-SCIF (hereinafter referred to as X-SCIF ICS) is a software and hardware complex designed to solve the problems of automated collection, processing, creation of an integrated data bank and analysis of various information.

The X-SCIF system provides:

- monitoring information from user- specified websites (web pages) on the Internet (Intranet) on specified topics;

- search for new sources of information on the Internet on topics specified by the user and their subsequent monitoring;
- creating and saving complex queries on given topics, in the form of a cataloged list or rubric, for subsequent automatic monitoring, search or content analysis;
- bringing the selected information to a single format and loading it into the repository;
- filtering, classifying, clustering, categorizing and announcing the loaded full-text information;
- automatic extraction (extraction) from the received information of entities (objects and facts);
- creation, on the basis of the unformalized full-text and formalized factual information loaded into the system, of an integrated data bank (repository) of objects, facts, events and documents interconnected by various types and motives of links, taking into account the attributes of reliability, relevance, as well as the weight coefficients of such connections;
- end-to-end search for information on user requests or topics, covering both search in an internal integrated data bank of previously extracted and accumulated information, and online metasearch on the Internet (search engines, websites, blogs, social networks) and other connected external data sources (official data-banks of government agencies, BKI, etc.);
- analytical processing of information (allows you to analyze joint mention and identify implicit links between objects, identify objects and group information by plot, build chains and graphs of links, analyze information activity, emotional coloring of documents, intersection of specified headings or topics, automatically create an information portrait of those selected on request documents, highlighting the objects, sources, regions, etc. mentioned in them, calculate the information favored index and much more);
- generation of output forms according to user-specified parameters (allows you to automatically create an electronic dossier, link diagrams, digests, reviews, com-

parative charts, information notes and aggregated reports);

- prompt delivery of query results through various channels (the system includes a virtual office, with its own remote crypto-protected document storage and mail system, which allows you to provide both "online" secure access via an encrypted channel to documents stored in the cloud, and "offline" receipt of the resulting documents by e-mail).

Structurally, the X-SCIF ICS consists of several subsystems focused on the corresponding needs of corporate customers, namely:

- X-Stream is a website monitoring subsystem created on the basis of InfoStream technology, as well as a full-text data bank (archive) of non-formalized information (articles, messages, etc.), which has been automatically updated since 1996 and is the most complete among the existing electronic archives in Ukraine;
- X-Files is an integrated data bank for the accumulation of various formalized reference and factual information, extracted and aggregated from all monitoring information sources available to the system, as well as a system of end-to-end search through internal and external sources (websites, blogs, online databases, social networks, etc.).
- X-Office is a virtual office system that provides secure access to corporate resources from anywhere in the world without installing additional software. The system includes a "cloud" file storage of documents and secure corporate web mail. Additionally, a VoIP-telephony server can be integrated into the virtual office for conducting confidential negotiations.
- X-Scoring is a pre-scoring system that allows you to automatically verify data and preliminary check the reliability of counterparties (individuals and legal entities).

Let us dwell on the consideration of each of the subsystems in more detail.

subsystem, built on ElVisti 's InfoStream technology, is designed to monitor information on the Internet according to user-

defined parameters, search for information on user requests or topics, prompt delivery of search results, and thus minimize efforts and save time spent on searching and processing the necessary information. The X-Stream subsystem provides the user with access to information on topics of interest to him simultaneously from a large number of websites, including those favorites that he is used to viewing daily. Currently, automatic monitoring of more than 7,000 sources is carried out, the flow of information exceeds 80,000 documents per day. Territorial coverage – Russian – English- and Ukrainian -language editions of Ukraine, Russia and other countries of near and far abroad. If necessary, any website available on the Internet can be covered. Information from the system is never deleted, but transferred to the archive. The archive of publications has been maintained continuously since 1996 and currently contains over 85 million documents.

The difference between this subsystem and competing products is its volume and the possibility of individual customization. It is focused not only on the rapid delivery of general news feeds, of which there are many in the web space, but also on monitoring according to parameters individually set by the user or archival search.

Viewing information is carried out through a single unified interface. The user can not only receive search results in real time, but also generate digests, information dossiers, build story-lines, analyze the relationship of headings, information activity, information links and joint mention of objects, etc.

Below are examples of displaying search results (Fig. 64), viewing individual material (Fig. 65) and the result of analytical processing of search results – a digest (Fig. 66).

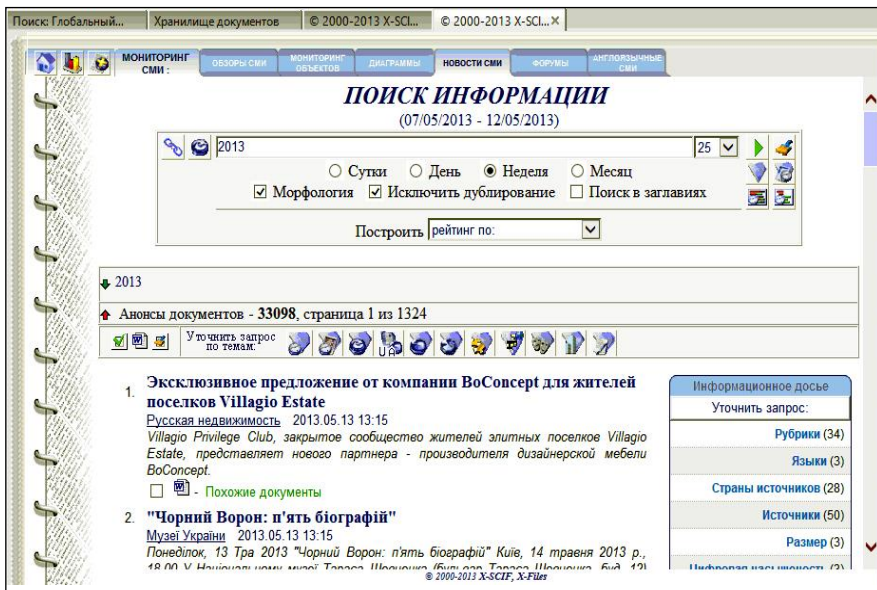


Figure 64 – Displaying search results

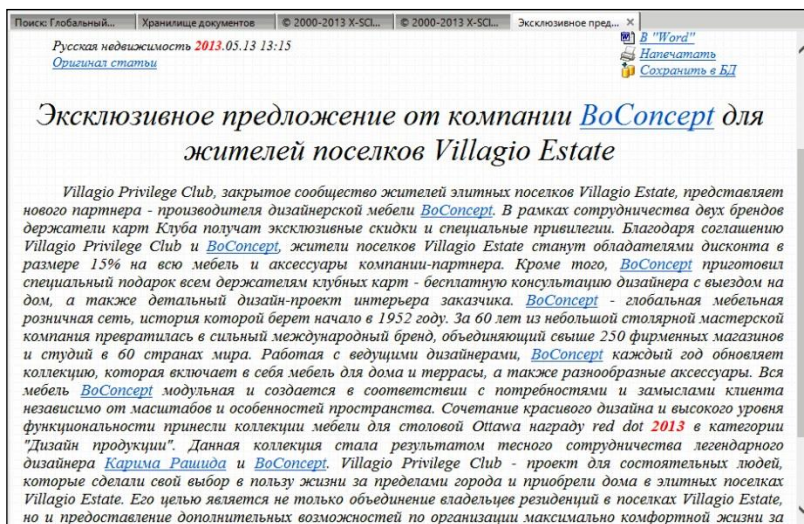


Figure 65 – View article

МОНИТОРИНГ СММ : ОБЗОРЫ СММ МОНИТОРИНГ ОБЪЕКТОВ ДИСТРИБУТЫ НОВОСТИ СММ АССОЦИИ АНГЛОЯЗЫЧНЫЕ СММ

ДАЙДЖЕСТ ЭЛЕКТРОННОЙ ПРЕССЫ

(27/08/2013 - 01/09/2013)

["Азаров Николай"](["Николай Азаров"])(["Азаров Никола"])(["Никола Азар"] 25

Сутки День Неделя Месяц

Морфология Исключить дублирование Поиск в заглавиях

Построить рейтинг по: [v]

["Азаров Николай"] ("Николай Азаров") ("Азаров Никола") ("Микола Азаров") ("Azarov Nikolay") |

Уточнить запрос по темам

Содержание

1. "Насильно мир не будешь..."
2. Азаров розповів, як звести ризики Митного союзу "практично до нуля"
3. Азаров поздравил нотариусов
4. Рыбак: вопрос Тимошенко - юридический
5. Украина через 7-10 лет может увеличить собственную добычу углеводородов до уровня, который позволит отказаться от их импорта
6. Азаров отмечает важность подписания с ЕС соглашения об Общем авиационном пространстве
7. Азаров заверил, что резких скачков цен на лекарства не будет
8. Угода про асоціацію з ЄС збільшить американські інвестиції в Україну, - Посол США
9. ПРИВІТАННЯ ПРЕМ'ЄР-МІНІСТРА УКРАЇНИ МИКОЛИ АЗАРОВА З НАГОДИ ДНЯ НОТАРІАТУ
10. ГАИшники на дорогах необходимы Украине - Азаров

Figure 66 – Digest

Using the X-Stream subsystem allows you to:

- promptly receive the necessary information as it appears on the Internet, analyze events, respond to them in a timely manner;
- to form their own information channels, which are conditioned by requests in the information retrieval language, to form archives for further processing and retrospective analysis;
- analyze the flow of information coming in real time;
- timely identify development trends and the state of markets for goods or services;
- track information about the activities of individual organizations, parties, movements, their PR activity;
- assess possible spheres of influence of conflict or crisis situations, carry out information control of probable sources of risks;
- find and check potential partners and customers.

The next structural element of the X-SCIF ICS is the X-Files subsystem (not to be confused with the well-known Russian system). This subsystem is intended for the accumulation and storage of formalized information obtained from all available sources, the implementation of an end-to-end search and further analytical processing of the information found.

Information obtained from various sources is processed, formalized, brought to a single form and recorded in an integrated data bank that structurally covers objects and relationships between them. Its structure, developed taking into account the practical needs of analysts, includes more than 40 types of accounting objects and more than 1000 motives for links between them.

The search for the necessary information is carried out by means of a global search, which is performed on all available data banks, and also provides for the automatic receipt of information from online information providers.

On fig. 67 shows the query input interface for global search. The subsystem allows you to present search results in various forms, the most convenient for solving the current problem.

One of the most common forms of presentation of selected data is the information dossier (Fig. 68). This form allows you to display information about the accounting object of the integrated data bank in a form that presents all the details of this accounting object, as well as all records related to it in other data banks. The information dossier output format is shown in fig. 69.

Глобальный поиск

Поисковые значения

Код предприятия/учредителя	<input type="text"/>	
Наименование предприятия, органы	<input type="text"/>	
Телефон/факс	<input type="text"/>	
МФО	<input type="text"/>	
Счет	<input type="text"/>	
Адрес	<input type="text"/>	
Фамилия	<input type="text"/>	
Имя	<input type="text"/>	
Отчество	<input type="text"/>	
Дата рождения	>= <input type="text"/> <= <input type="text"/>	
Свидетельство налогоплательщика	<input type="text"/>	
Идентификационный номер	<input type="text"/>	
№ паспорта	<input type="text"/>	
Другое удостоверение	<input type="text"/>	
Гос. номер авто	<input type="text"/>	
Номер кузова	<input type="text"/>	
Электронный адрес	<input type="text"/>	
Контекстный поиск	<input type="text"/>	
Поиск в Интернет и архивах	<input type="text"/>	

Figure 67 – Interface to global search query water

1. Информационная справка - Сообщения СМИ

Ключевые реквизиты

Вид сообщения	СООБЩЕНИЕ О БАНКРОТСТВЕ
Название	"Голос України" N 87 (3337) від 14 травня 2004
Дата	14.05.2004
Анонс	18.03.2004 р. господарським судом Харківської області (61022, м. Харків, Держпром, (під.) порушено провадження по справі N Б-19/22-04 про визнання банкрутом Сільськогосподарського

Информация

Текст сообщения
 18.03.2004 р. господарським судом Харківської області (61022, м. Харків, Держпром, 8 під.) порушено провадження по справі N Б-19/22-04 про визнання банкрутом Сільськогосподарського товариства з обмеженою відповідальністю "Агрокомбінат Богодухівський" (62103, м. Богодухів, Харківської обл., вул. Залізнична, 14, код ЄДРПОУ 22660116, п/р 26008301747 у Першому ХФ АКБ "Базис", МФО 331598). Розпорядником майна призначено арбітражного керуючого Панасюка І.В. (лицензія АА N 047594 від 03.07.01, адреса: м. Харків, вул. Полтавський шлях, буд. 154, кв. 84). Заяви кредиторів приймаються протягом місяця з дня опублікування.

Источники Голос України

Дополнительная информация

Особенности информации: БОЛЬШАЯ ЦИФРОВАЯ НАСЩЕННОСТЬ
 Дата ввода: 17.07.2012
 Дата редактирования: 17.07.2012

Связанная информация

СМИ о ИП

Мотив связи	По сообщениям СМИ
Дополнительные реквизиты	
Дата возникновения связи	14.05.2004
Достоверность	ДОСТОВЕРНАЯ ИНФОРМАЦИЯ
Ключевые реквизиты	
Код ОКПО	22660116
Наименование объекта	Сільськогосподарське Товариство з Обмеженою Відповідальністю "Агрокомбінат "Богодухівський"
Правовая форма	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ
Адрес (текст)	ХАРКІВСЬКА ОБЛ., БОГОДУХІВСЬКИЙ Р-Н, м.БОГОДУХІВ ВУЛ. ЗАЛІЗНИЧНА БУД. 14
Страна	УКРАЇНА

Figure 68 – Information dossier

(заключение на кандидата/сотрудника)

Дата: 13.05.2013 № _____ на вх. № _____


Место работы, должность

Организация: ██████████ ТОВАРИСТВО З ОБМЕЖЕНОЮ ВІДПОВІДАЛЬНІСТЮ
 ██████████
 Должность: Охранник

Фотография, краткая характеристика, результаты проверки

Род занятий, специализация: СПЕЦІАЛІСТЫ, ОХРАНИК, ТЕЛОХРАНИТЕЛЬ
 Краткая характеристика: Здесь текст характеристики

Семейное положение: ЖЕНАТ
 Хобби: Рыбалка



Установочные данные:
 (дата и место рождения, гражданство, адрес регистрации, адрес проживания и др.)

Дата рождения: ██████████
 Место рождения: ТУРКМЕНИСТАН, м.Красноводськ
 Гражданство: УКРАИНА
 Адреса регистрации, проживания:
 УКРАЇНА, м. КИЇВ, БРАТИСЛАВСЬКА, ██████████
 Телефон:
 38044 ██████████
 ██████████
 Адрес регистрации:

Figure 69 – Information dossier output format

Another example of a data presentation form is a graphic dossier. The selected objects, together with their connections, are displayed in the form of a graph in which the vertices are accounting objects, and the edges are the connections between the corresponding objects (Fig. 70). This form of presentation makes it possible to carry out analytical studies of both explicit and implicit relationships of accounting objects, present them on the screen in the form of graphs at various scales, print diagrams of these graphs, etc. Also in this mode, the user has access to all the tools for editing, entering and deleting information, providing intuitive and fast editing of formalized data.

For analytical processing of large volumes of the same type of information, the system provides a mechanism of aggregated forms. It allows, on the basis of initial information, which is difficult to directly analyze, to build aggregated forms, graphs, and to calculate integral characteristics.

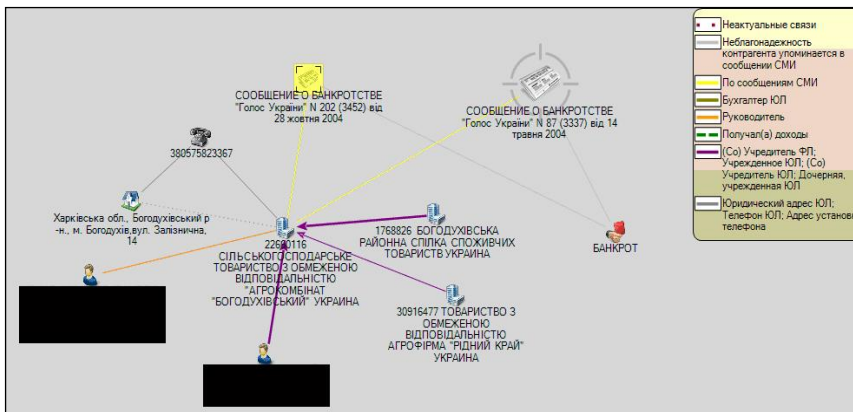


Figure 70 – Visualization of graphs of connections of the accounting object

One of the key features of the X-Files subsystem is a module for automated input and recognition of full-text documents. It allows you to create accounting objects (persons, companies, phones, addresses, e-mail addresses, etc.) without the participation of an operator and establish links between them based on informal documents (texts, questionnaires, cards, etc.).

The X-Office subsystem is designed to provide remote user interaction and ensure effective collaboration. It, in turn, includes the following subsystems:

- "Corporate Webmail". Provides work with corporate mail from anywhere via an encrypted communication channel without the need for configuration and "traces" on the computer;
- "File storage of documents" (Fig. 71). Represents a remote secure file storage that is accessible from anywhere only to members of a closed group. Provides access to personal and corporate documents with the ability to collaborate with multiple users (Fig. 72). The document storage provides the possibility of end-to-end search through the content of documents. Differentiation of access to the text of the document is made according to the access profile or with the permission of the author of the document;

- "Negotiation room" (Fig. 73). Provides system users with the ability to communicate within a closed group in text, voice and video mode using a secure communication protocol. Also available is the ability to make calls to landline and mobile phones outside the group with the inability to determine the outgoing subscriber.

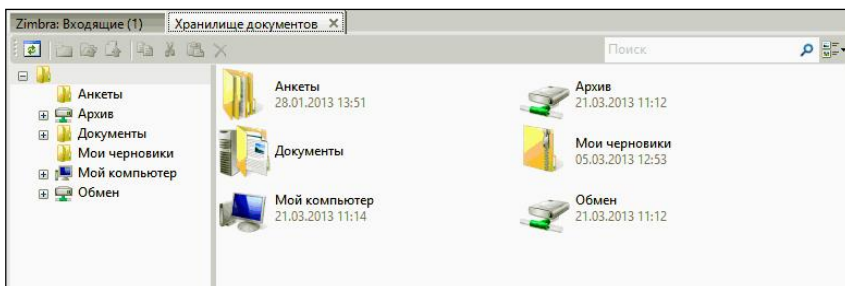


Figure 71 – Storage interface corporate documents

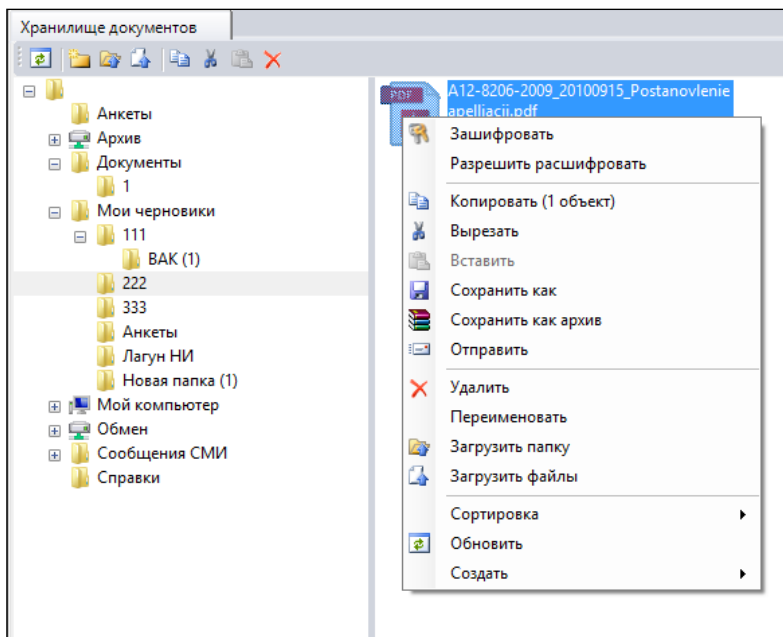


Figure 72 – Operations available in file storage

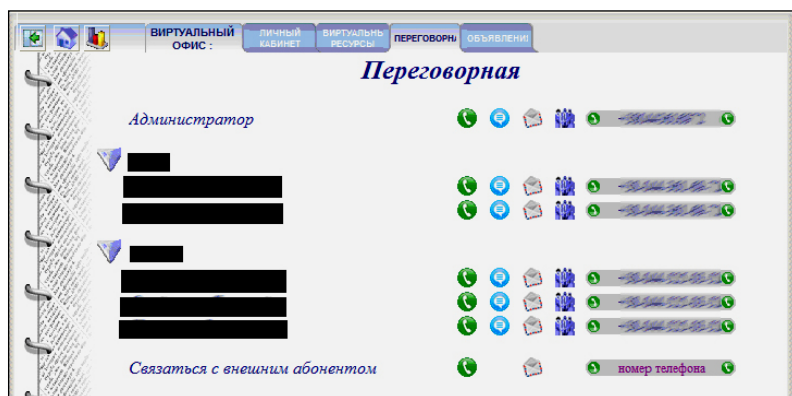


Figure 73 – Interface of the subsystem "Negotiation room"

The X-Office subsystem has a number of features that give its users significant advantages over similar software and hardware systems. Authorization in the subsystem occurs by a combination of a fingerprint and a password, and encrypted communication channels and only trusted certificates are used for communications, which prevents traffic analysis at the ISP level or at any other point of interception. Upon completion of work with the subsystem, no traces of its operation remain on the user's computer.

The file storage of corporate documents, which is part of the X-Office subsystem, provides the user with the ability to work with remote cloud storage as simply as with a folder on the local drive of his computer. The subsystem makes work with various sources (local disk, corporate documents, file storages) transparent for the user. Editing office documents is possible without installing and configuring additional software. Differentiation of access to various files is carried out on the basis of access templates set by the administrator. If there is highly secret information, the user has the ability to additionally restrict access to it by means of encryption directly from the program interface.

The X-Scoring subsystem is available to check the reliability of counterparties in automatic mode. The subsystem is implemented on the basis of XML Web service technology, which pro-

vides transparent integration with most existing systems on the customer's side. Despite the large amount of constantly updated information, on the basis of which a decision is made about the reliability of the counterparty, the system gives an answer in less than 3 seconds.

The decision-making algorithm can be flexibly adjusted to the needs of a particular customer. A typical algorithm for conducting a check and making a decision on the reliability of a counterparty consists of the following steps:

- assessment of the economic solvency of the client according to the information provided by him;
- automatic verification of the data bank, aimed at checking the compliance of the information provided by the borrower and identifying possible fraud attempts on the part of unscrupulous borrowers;
- detailed verification of the counterparty, whose application has passed all the previous stages.

It should be noted that full-featured competitive intelligence systems are not always available and even necessary, due to their cost characteristics or other reasons. At the same time, certain tasks of competitive intelligence can be partially solved by quite affordable means. The use of new approaches, as well as open, accessible and relatively inexpensive information sources, allows today to effectively support the adoption of managerial decisions in many, including strategic, areas of business.

The competitive intelligence technologies of tomorrow are in many cases already implemented today in the form of military information technologies. So, for example, to carry out intelligence at the state level in the United States back in 2005 within the framework of National Intelligence (National Intelligence), a special structure was created – the Center for Open Sources (Open Source Center). At present, all US open source intelligence centers are integrated into this single information system. B 2006. the information resources of this system are called Intelink-U [Kondratiev, 2010].

Despite the fact that the materials in this system are extracted from publicly available open sources, it is not intended for everyone. Information from the system is distributed over restricted access networks.

Intelink – U systems include numerous databases, including :

- the CIRC database containing over 10 million scientific and technical articles, including information on patents, standards, military weapons and military equipment;
- the DTED database containing a wide variety of maps from the National Geospatial-Intelligence Administration;
- materials of centers and points of information service of foreign broadcasting FBIS;
- database of periodicals IC ROSE;
- information portals of research and educational institutions;
- Jane's Information Group online directories;
- resources of the non-state information and analytical agency STRATFOR's, which provides, among other things, regular updates on the areas of deployment of aircraft carrier and expeditionary strike groups of the US Navy.

The ways of filling such information resources also deserve attention. For example, to collect information from the sites of the World Wide Web, its systematization, translation and archiving in the World Information Library (World Basic Information Library – WBIL), the operation of which is entrusted to the Office of the Study of the Armed Forces (AF) of Foreign States FMSO of the Command of Training and Scientific Research for the construction of the ground forces (SV) (Training and Doctrine Command – TRADOC), the personnel of the reserve of the SV and other branches of the Armed Forces are involved.

To meet the needs of the US Department of Defense and the intelligence community, the HARMONY database has been created, containing bibliographic references to all available sources of information (meta-information) about foreign countries. Database HARMONY is characterized by ease of use, the ability to quickly find the necessary documents, and the rapid exchange of data within US government structures.

The World Basic Information Library (WBIL) is a special program of the U.S. Intelligence Community administered by the Foreign Military Studies Office (FMSO) of the Training and Doc-

trine Command. Command, TRADOC). Personnel with access to the database can collect information from the Internet, organize it and archive it in the WBIL library using the Pathfinder analytical toolkit. The Pathfinder system allows you to analyze 500 thousand documents from various databases in a few minutes.

The combination of these technologies allows US military intelligence officers to access huge amounts of data and meet the needs for intelligence information.

Let us consider some projects of the US National Security Agency (NSA) focused on data mining, analytics and forecasting [Chernykh, 2013].

The project of total integration of information flows involved the creation of software tools that provide a solution to a very complex problem that has not been solved so far in the civilian sector – the integration of all information flows in a single repository, on the one hand, and their separation according to special criteria, on the other hand. By now, according to experts, the problem has been completely solved. In the military sphere, the Prosecutor's Management Information System database management system (PROMIS) developed by Inslaw Inc. under the direction of Bill Hamilton (B. Hamilton). For more than 30 years, the program, which initially has 570,000 lines of code, has been continuously improved by the NSA's own developers.

The PROMIS program is capable of simultaneously integrating an unlimited amount of information received from an unlimited number of programs and contained in any number of databases, regardless of their types, languages in which the original programs are written, architectures of operating systems and platforms from which information is retrieved. Apparently, there are no analogues of PROMIS in the world.

Another important area is real-time knowledge mining. Today, according to available information, machines can extract knowledge about about 40 million entities (objects, subjects, events, etc.) and more than 2.5 trillion. parameters.

Commissioned by DARPA, Raytheon has created a self-organizing knowledge base that allows you to automatically create dossiers on citizens and organizations by collecting information from open sources. From the end 2011. IBM and Recorded Future became the lead developers of such knowledge bases. Already today, by order of certain branches of the US military,

they have managed to create effective systems for early warning of crises.

Very high hopes are associated with the project for the automated detection of anomalous processes occurring on various scales. The sources of information for the program are, as usual, web 1 and web 2, as well as analysis of streaming video, financial transactions, etc.

The NSA and the American intelligence community have paid and continue to pay great attention to the project on information analysis and real-time forecasting. The best-known implemented automated system (with the participation of a human expert) within this project is Palantir, developed by Palantir Technologies, which is designed for data analysis and visualization. The system collects data streams from all available information channels about all recorded events related to people: banking transactions, credit card transactions, phone calls, e-mail, information from CCTV cameras, information about transactions in all federal and municipal databases, etc. P. In each data stream, data mining tools identify unusual events, the probability of which is small, and events from predefined "alarm lists". Then, information about unusual events from different databases is combined and linked. And as a result, if the calculated probability of the entire complex of associated unusual events is below a certain predetermined probability threshold, an alarm is issued indicating the specific person with whom the entire complex of events is associated.

The Palantir program is gradually finding application in the civilian sector. For example, Guy Chiarello, a spokesman for JPMorgan Chase, says that Palantir's programs help the bank identify fraudsters even before a crime happens. The contract with Palantir, according to him, is "the best deal in recent times."

At the same time, the main investors of the project Alex Karp (Alex Karp) and Peter Thiel (Peter Thiel) say that the greatest problem that they have managed to solve with the help of their programs is the ability to fight terrorism and violence while maintaining civil liberties. The information involved in their work is classified, and users have access only to its individual fragments.

For high-quality competitive intelligence by methods of analyzing texts from the Internet, it is necessary to formulate goals,

build databases for observations and research, and formulate queries. Note that one should not be limited to one information retrieval system even for the analysis of such information as Internet resources. We recommend using the best global and special information retrieval systems such as Google (www.google.com), Yahoo! (www.yahoo.com) or Microsoft Bing (www.bing.com). _ _ _ For special needs, it is also recommended to use legislative, address-nomenclature, price databases available both from the Internet and in local versions.

Let's show how queries related to competitive issues are formed using the example of search prescriptions for the InfoStream content monitoring system (www.infostream.ua, Fig. 74).

Usually, the search for information about a company or person always begins with the indication of various ways of writing the company name or full name. persons. Sometimes searching in operational and retrospective data for such "primitive" queries is quite enough, but the task becomes more complicated if it is necessary to investigate the state of a particular industry, a particular region, or even an entire country. In such cases, in accordance with the problem, queries are built, which are then iteratively refined.

As an example, we will give a number of concepts, and then we will put query fragments in correspondence with them and consider fragments of texts published by various sources, which can then be used to build various kinds of analytical references.



Figure 74 – InfoStream content monitoring system website

After finding documents containing references to the analyzed firms or brands, it is possible, by refining queries, to find out some important characteristics related to the activities of these companies. As an example, below are lookup queries related to the financial position of the companies mentioned in the web space:

Statutory ~to capital~/2/UAH
 Statutory ~ capital~/2/USD
 Statutory ~fond ~/2/UAH
 Statutory ~fond ~/2/USD
 Owned~/2/shares

The first two queries imply finding documents that include fragments containing the phrases "authorized capital" or "authorized fund", indicating the value in dollars or hryvnias ("~/2/"; in the language of requests, this means a distance of 2 or fewer words between expressions).

The search resulted in text documents containing the following fragments:

The Antimonopoly Committee approved the acquisition of Prominvestbank by Luregio Limited through the purchase from Fortify Financial Company LLC. In accordance with the information in the Unified State Register of Legal Entities and Individual Entrepreneurs, Luregio Limited on August 28 of this year registered Luregio Invest LLC with **authorized capital 350 million UAH**, the main activities of which are consulting and management, the provision of financial services and other ancillary commercial services. Lyudmila Nazarenko, who was deputy director for legal issues of TAS Group LLC and a member of the Supervisory Board of Universal Bank, owned by Tigipko, was appointed as the head of the enterprise.

Fin.org.ua 2021.02.05 18:15

The Ministry of Infrastructure has identified the first participant in a pilot project to allow private locomotives to operate certain routes on public railways. They became about LLC "Ukrainian locomotive building company". Ukrainian Locomotive Company LLC with **an authorized capital of UAH 3,000** was founded in Kiev in November 2016. Director Vyacheslav Yakubovsky is indicated as its beneficiary through Eurasia Investment Company LLC.

Business Censor 2021.02.05 18:02

On February 2, the Gambling and Lottery Regulatory Commission issued a license to SPACEICS LLC (TM Kosmolot) to carry out activities for organizing and conducting gambling casino games on the Internet. According to the YouControl portal, SPACEICS LLC was registered on May 28, 2020 in Kiev, the size of **the authorized capital is UAH 30 million**. Founder and leader – Sergey Potapov. The main activity of the company is the organization of gambling.

IPress.ua 2021.02.03 19:37

Centrenergo operates 23 units (18 pulverized-coal and five gas-oil) at Uglegorskaya, Zmievskaya and Trypilska TPPs with a total installed capacity of 7,660 MW. The state **owns 78.3% of** the company's shares.

Interfax-Ukraine 2021.02.04 21:01

The head of the NAC, Andriy Kobolev, said that from a legal point of view, the mechanism for dividing Ukrnafta is complicated, and assistance from state authorities will be needed to implement it. Naftogaz Ukrainy owns 50% + 1 share in Ukrnafta, and a group of companies associated with Igor Kolomoisky **owns about 42% of the shares**. The company has 25 drilling rigs, 1891 oil and 162 gas wells, and 537 filling stations.

Interfax-Ukraine 2021.02.04 21:01

Information about mergers and acquisitions in a particular business area, which allows you to monitor the expansion of competitors into new market niches, can be obtained as a result of working out such clarifying queries:

Buy ~ /2/ shares

buy~/2/package ~a stock

sales~/2/package ~a stock

(merger ~to companies) & (shares, assets)

Executing these lookup queries allows you to get documents containing, for example:

PJSC "Prominvestbank" was established in 1992. The Russian VEB became the owner of Prominvestbank in 2008, **having acquired a 99.7726% stake** in its shares. VEB estimated its investments in the development of the Ukrainian subsidiary bank at \$2.7 billion.

UaProm 2021.02.04 17:49

The current owner of the Italian team Gabriele Volpi is ready to consider selling. Pace recently **acquired an 84 per cent stake** in Premier League club Burnley. For this deal, he paid 170 million euros.

Dynamomania.com 2021.02.04 19:34

KYIV. February 2. UNN. President Volodymyr Zelensky gives no legal reason for blocking Chinese shareholders' access to the management of Motor Sich, so his actions are politically motivated. Viktor Suslov, ex-Minister of Economy, draws attention

to this in his publication, reports UNN. "From the point of view of legislation, it is not clear what "the president intends to prevent the purchase of a controlling stake" means," the expert comments on the personal intervention of the President of Ukraine in a legal deal to sell a controlling stake in **Motor** Sich. – After all, Zelensky does not say, that the Chinese did not act according to the law or violated some norms.

FinOboz 2021.02.02 22:34

In 2007 Wuhan iron and Steel, which co-founded the Baowu group nine years later, **acquired a 48.81% stake** in Kunming iron and Steel joint Stock – Kunsteel's largest steel company. Earlier it was reported about the intention of Baowu to take over another Chinese steel company Shandong Steel. With these acquisitions, the group's production capacity will exceed 150 million tons per year.

UaProm 2021.02.04 12:49

To identify publications about changes in financial condition and bankruptcy, you can use the following lookup queries:

issue~/2/stock
(increase ~ shutter) & (fund,capital)
boost~/1/dol~/2/share
announced~/2/bankruptcies

The development of such requests made it possible to find the following documents:

The PFTS exchange list included 119 issues of domestic government bonds, 12 issues of external government bonds, 65 issues of corporate bonds, 101 issues of corporate **shares**, 23 issues of securities of joint investment institutions, 15 issues of municipal bonds, one government derivative, one bond of a foreign government, two issues of bonds of a foreign issuer, three issues of shares of foreign issuers and three issues of securities of joint investment institutions of foreign issuers.

E-Finance 2021.02.05 09:30

The city council of Odessa **has increased the authorized capital** of KP "Odesgorelektrotrans". It grew by UAH 45 million – up

to UAH 108.5 million. The funds added to the authorized capital of the KP will be used for the next payment (must be made before March 25) under a loan agreement with the European Bank for Reconstruction and Development. Due to the loan, Odessa bought 47 new trolleybuses.

It's time to talk 2021.02.05 13:36

After the audit, Wirecard admitted that it wrote down 1.9 billion euros from non-existent bank accounts into the asset. The price of her shares on the stock exchange collapsed, and she herself **declared bankruptcy**. The Federal Office for Supervision of the Financial Sector called the situation a "catastrophe" and a "shame". The international rating agency Moody's withdrew the Wirecard rating.

Goodnews.ua 2021.01.30 06:42

The Furshet chain **has declared bankruptcy**. The Economic Court of the Dnipropetrovsk Region has begun the process of filing the bankruptcy of the Retail Center company, which is the owner of the Furshet chain. Suppliers of the Fursheta retail network received a letter signed by the director of Retail Center V. Kupchenko that the court has begun the bankruptcy proceedings of the company, in connection with which a moratorium is introduced on payments to creditors and a property manager has been appointed.

Freedom Matrix 2021.01.27 19:03

Content monitoring methods are the adaptation of classical content analysis methods to the conditions of dynamic information arrays, for example, information flows from the Internet.

A typical task of content monitoring is the construction of diagrams of the dynamics of the appearance of concepts over time.

Let's consider how the InfoStream system tracked the crisis in the food market in Ukraine in June 2011. To do this, a request was made "**crisis & buckwheat & Ukraine**", which was entered through the web interface of the system. In the special "Dynamics" mode, the corresponding diagram of the appearance of the concept was obtained (Fig. 75).



Figure 75 – The dynamics of the emergence of the concept

The diagram above shows that the mass appearance of reports of crisis phenomena occurred on January 12 2011 (while the prices for buckwheat rose sharply only in mid-March).

Of course, the prompt receipt of this type of data should have helped analysts in building short-term forecasts.

Similarly, you can monitor the financial market. For example, a simple query “dropped ~ to urs ~ hryvnia”, referring to a fragment of the information flow for the period from January to March 2015, produced a diagram showing the dynamics of the depreciation of the Ukrainian currency (Fig. 76). As you can see, the peak of publications falls on February 27, 2015, when, in particular, there was a “local” depreciation of the hryvnia to UAH 44/USD.

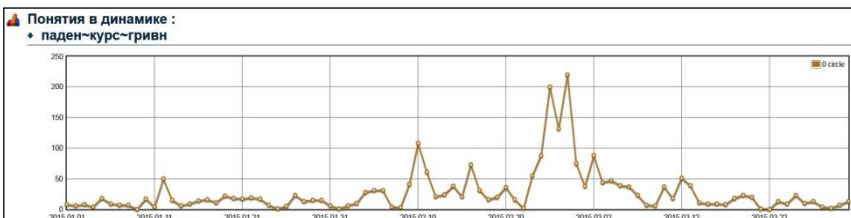


Figure 76 – "The fall of the hryvnia" in dynamics

3. Sources of information

In information and analytical work, the ability to access sources of data, information and knowledge is important. At the same time, the main problem is finding meaningful and reliable sources from all public sources. When such sources are found, the mechanisms for turning data into knowledge are activated, for which appropriate technologies are used. Data is usually understood as "raw", unprocessed information based on facts. This can be statistical data, facts from the biographies of key people, or, for example, information about the reporting of individual companies. Information is already processed and analyzed data in a certain way. The end product of any analytical work is knowledge – synthesized conclusions, recommendations for decision-making.

Information, as mentioned above, can be obtained from official, open sources, media, announcements, advertising, corporate, banking, government reports, databases, from experts through analysis or special processing of data, texts.

Below is a detailed list of types of information sources that are most often used in competitive intelligence [Nezhdanov, 2009].

1. Press releases of companies, official statements on behalf of companies about new technologies, new directions, transactions, prospects. Such press releases are created by companies for their own promotion, attracting the attention of potential customers, investors looking for profitable options for investing their funds. Often in such statements there is information about intentions, planned events. Press releases are available on the websites of companies, in PR services, on general and specialized specialized platforms for posting press releases.

2. Interviews with company employees, relevant materials in the media. In interviews, plans of companies are of interest. At the same time, the competitive intelligence service may initiate an interview with one of the employees of the object of interest.

3. Statements by company employees on forums, blogs, in private conversations. At the same time, company plans, personnel policy, the atmosphere in the team, etc. can be identified. Sources of information: 1) Internet resources (specialized forums, employee blogs), expert blogs, groups in social networks; 2) exhi-

bitions, conferences, advanced training courses, professional events.

4. Tenders, purchases. Procurement items, equipment, performers. Sources of information: 1) Internet resources (websites of companies, trading platforms, specialized forums); 2) partners of the company under study, those who participated in their tenders, customers and suppliers.

5. Patents, copyright certificates of the company and its employees. For the tasks of competitive intelligence, their content, focus, lists of co-authors are interesting. The information is posted on the respective websites. For Ukraine: <https://ukrpatent.org/>; Google Patents : <https://patents.google.com/>; Eurasian Patent Office: www.eapo.org. Patenting is possible in any country, the preferred options are the country of registration of the organization, the country of doing business, in addition to the USA, the European Union, Russia, Japan and China.

6. Developments of the company: ongoing, funded, developments in which the company is interested. The company's attempts to conduct research are subject to supervision: the purchase of specific equipment, the hiring of specialists, negotiations, visits to relevant organizations, etc.

7. The company's activity in the market of mergers and acquisitions (M&A). Information about which organizations are being taken over, planning to take over or negotiating a takeover. Information can be obtained from the Antimonopoly Committee (AMC) of Ukraine, similar departments in other countries, according to news reports on web resources dedicated to M&A.

8. Vacancies of the company (opening, closing), messages about the active search for employees, requirements for vacancies, conditions. Source of information: the company's website, job search sites and websites of agencies with which the company cooperates.

9. Refresher courses, staff training – an indication of the priorities in the development of the company. What is of interest is what is taught, which specialists are invited for training, what requirements are put forward when attracting trainers, what are the terms of training, how many personnel are trained.

10. Acknowledgments and awards of the company and its employees.

11. Participation in events (exhibitions, conferences, round tables, presentations). Finding out in which events companies participate, their focus, the circle of participants.

12. Participation in organizations (unions, associations, confederations, etc.) – information about which associations the company participates in, how actively it participates, what it receives from participation, what it expects, how it uses it.

Information is characterized by qualitative, quantitative and value indicators. Qualitative characteristics usually include: reliability, objectivity and unambiguity of information. Quantitative characteristics include its completeness (absence of unexplained gaps) and relevance (degree of compliance with the essence of the questions and tasks posed). Value characteristics are the cost and relevance of information.

The activity of competitive intelligence is based on the use of only legitimate sources of information, which are quite sufficient for making managerial decisions in the field of business, it is only necessary to carry out some information and analytical processing of the available open data. Among such sources of information are: statistical data, materials from websites, social networks, media, industry reports, etc.

Many competitive intelligence services are not always able to separate the illegitimate part of the information from the legal, and the customer, as a rule, is interested in the final results, the sources for him act only as confirmations, intermediate data. At the same time, reputable customers themselves are interested in the fact that information is obtained by legal means, so that the analytical report is legal.

In competitive intelligence in recent decades, a new information source has appeared and developed to an unprecedented scale – the Internet web space. Today, according to experts, the Internet is in first place in terms of the amount of information, ahead of the media, industry publications and news received from colleagues, special reviews, closed databases. At the same time, open sources and specialized databases available on the Internet contain most of the information necessary for conducting competitive intelligence, but the question of finding and effectively using it remains open. Recent studies of the web information space have shown that the trillion web pages available through traditional information retrieval systems are only the

"superficial visible part of the iceberg." About 40% of all information on the Internet is available for free. Navigation in this information space is provided by more than a million search engines and directories, but they cover only a small part of the information resources. There are much more hidden and invisible (deep, invisible) Internet resources – these are, first of all, dynamically generated pages, files of various formats, information from numerous databases. The "hidden" web can also include networks such as BitTorrent, DirectConnect, EMule, Napster, etc.

Today, the main sources of information for competitive intelligence are the Internet, the press, and open databases. Databases of state and statistical bodies, chambers of commerce and industry, privatization bodies, etc. are very popular among competitive intelligence specialists. Separate accessible databases of other authorities are also of great use. Recently, databases based on media archives, including online ones, have become increasingly popular. In Russia, for example, the largest media archive database of the Integrum service (integrum.ru) is very popular, containing several hundred million documents. With the help of another Russian database "Labyrinth" (labyrinth.ru), compiled on the basis of publications of leading business publications, one can obtain extensive information about specific persons, organizations and companies.

Traditionally, competitive intelligence relies on the following sources of information, such as published open access documents that contain product market overviews, information about new technologies, partnerships, mergers and acquisitions, job postings, exhibitions and conferences, etc. The information contained in the documents already available in companies conducting competitive intelligence, the results of marketing research, information obtained at conferences, when communicating with clients and colleagues are widely used. Much of this data ends up in the online press, press releases or published on corporate websites. Therefore, in recent years, databases based on mass media archives, including (and mainly) network ones, have gained great popularity.

3.1. Web sites

The web space, based on the physical infrastructure of the Internet and the HTTP data transfer protocol, unites hundreds of millions of web servers connected to the Internet (Fig. 77). In the early days of the web space, a small number of websites published information from individual authors to a relatively large number of visitors. Today the situation has changed dramatically, there has been a transition to the second generation web. Website visitors themselves are actively involved in the creation of content, which has led to a dramatic increase in the volume of information and dynamics of the web.

Today, the Web already has an information base freely accessible to users of a volume that was previously difficult to imagine. Moreover, the volume of this base exceeds by orders of magnitude everything that was available a decade ago. In August 2005, Yahoo! announced that it had indexed about 20 billion documents. Google's achievement in 2004 was less than 10 billion documents. Google has indexed over a trillion web documents today. According to the Netcraft Web Server Survey (news.netcraft.com, Figure 77), there are currently over 1,198 million website addresses (out of about 200 million active).

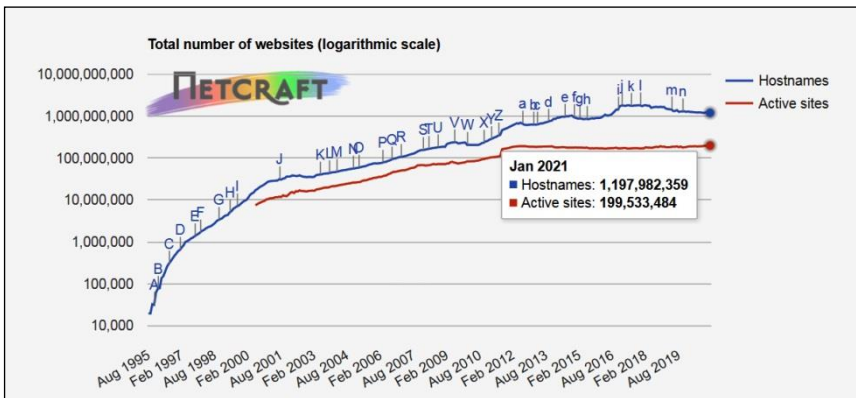


Figure 77 – Dynamics of growth in the number of web servers on a logarithmic scale (Netcraft, January 2021)

Open sources and specialized databases available on the web contain most of the information necessary for analytical research, but the questions of finding and effectively using it re-

main open. When using the web space as the most powerful source of information, as noted earlier, the most significant are the problems of volume, navigation, the presence of information noise and the dynamic nature of information on the Internet.

The possibilities of access to Internet resources, which attract with their openness, volume and content versatility, seem limitless at first glance. However, important developments in various fields indicate otherwise. It is in crisis situations that the Internet quite often fails. There are many problems – from congested network infrastructure – to virus attacks, vulnerabilities and denial of service of individual web servers. A number of problems are also generated by the volumes, variety of presentation and dynamics of the content segment of the information space.

Despite such qualities as openness and accessibility, the existing infrastructure of the web space cannot be considered reliable and reliable. Let's name a few more problems inherent in the web space:

- the problem of user access to heterogeneous web resources from a "single window" to obtain a generalized view of information flows on the required topics has not been solved;
- the possibility of timely “reminder” and “pushing” of profile information for the user, published on a large number of websites, is not provided;
- a sufficiently high probability of denial of service to critical web resources at the most necessary time.

It is known that today there are content integration technologies that allow to partially solve these problems, providing efficient search and navigation in the web space, monitoring and aggregation of open web resources. For professional search and aggregation of information from the web space, specialized software, information retrieval systems and services are used. Here are some examples of software products and services:

Avalanche (www.tora-centre.ru) is a family of web monitoring software. Avalanche technology is based on three main solutions: the concept of "smart folders" (Smart Folders), an autonomous intelligent search robot and an embedded database ("personal encyclopedia").

Newprosoft Web Content Extractor (www.newprosoft.com) is a web crawling and data extraction program that automates the data extraction process and allows you to save the extracted data in a user-selected format. Web Content Extractor is a powerful and easy to use web scraping tool. It allows you to extract certain data, images and files from any website. The web data extraction process is fully automatic. Web Content Extractor has a user-friendly interface, scanning rules and an extraction template ensure efficient and accurate data extraction.

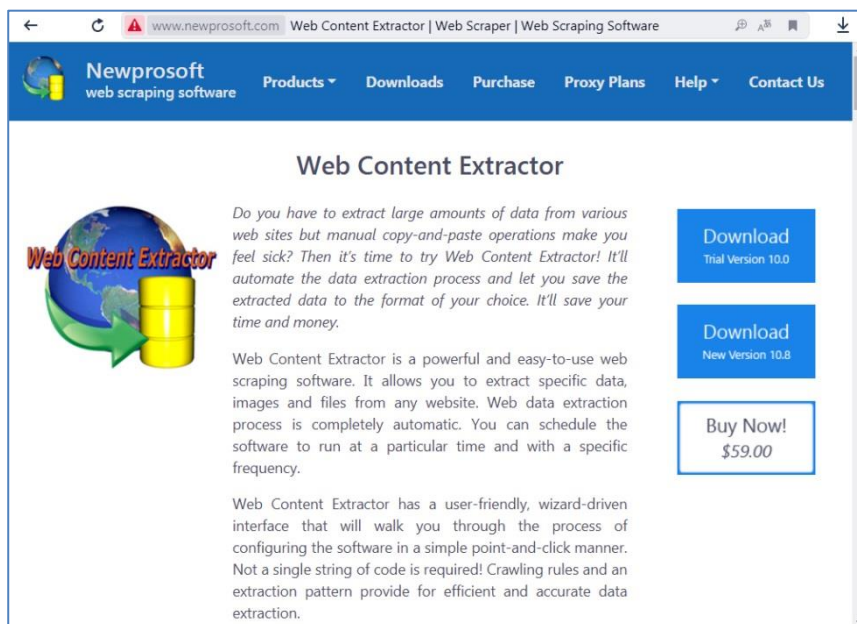


Fig. 78 – Fragment website Newprosoft Web Content Extractor

WebSite-Watcher (www.aignes.com) is a program that allows you to monitor websites, forums, local files, provides information filtering, as well as convenient visualization of monitoring results.

Service solutions include:

WatchThatPage (watchthatpage.com) is a free service that allows you to automatically collect new information from monitored web resources. The user chooses which pages to watch, and WatchThatPage determines which pages have changed and provides him with all the new content. New information is provided by email.

Newspaper Map (newspapermap.com) is a service that combines geolocation and an information retrieval system for media resources. When solving the problems of competitive intelligence, the user can select the region, language, the list of online versions of newspapers and magazines that are of interest to him, and directly go to the documents. The service supports the Russian language and has a user-friendly interface (Fig. 79).

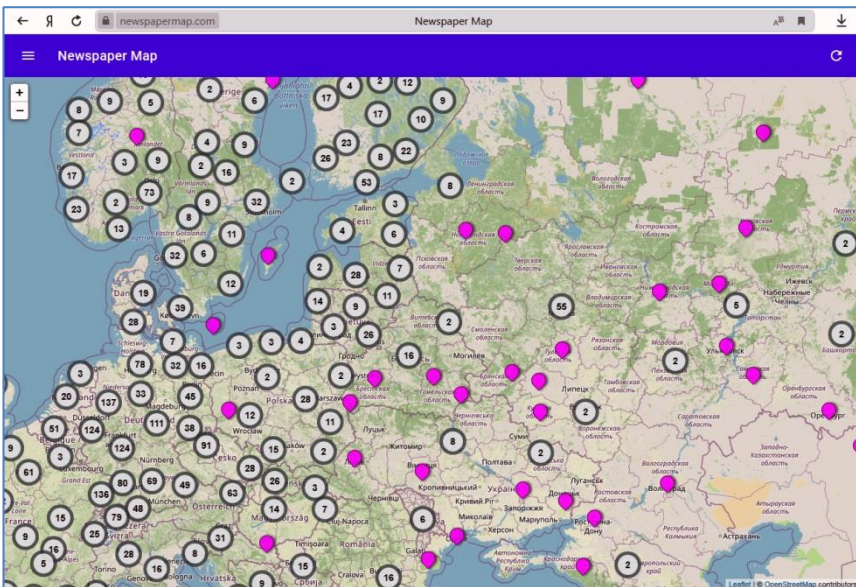


Figure 79 – Fragment of the geographical news aggregator Newspaper Map

WebSvodka (websvodka.ru) is a tool for automatic control of changes in Internet pages, which allows you to quickly find out:

about the appearance of new vacancies and announcements; on changes in price lists; about new orders, instructions, tenders, competitions; control mentions of the words you are interested in. The user sets the topic or page of interest to him, and WebSvodka regularly informs him of all developments on this topic. To increase the reliability and speed of work, the modules for loading content, analyzing pages, and storing results operate in parallel and are located on different servers.

InfoStream (www.infostream.ua) is a content monitoring service for web resources that provides access in search mode to information from 10,000 sources, information classification, extraction of concepts (persons, companies, toponyms), formation of story chains, assessment of message sentiment, analysis of the dynamics of publications for certain objects. The system's databases store over 500 million news documents over 25 years.

Webground (webground.su) is an aggregator of news information from the Russian-speaking segment of the web space. It can be used in competitive intelligence for tracking topics of interest, obtaining thematic stories, retrospective analysis of the development of topics over time.

3.2. Social networks, blogs

The term "social network" denotes the concentration of social objects, which can be considered as a network (or graph), the nodes of which are objects, and the links are social relations. The term was coined in 1954 by Manchester School sociologist J. Barnes in *Classes and Gatherings in the Norwegian Island Parish*. In the second half of the 20th century, the concept of "social network" became popular among Western researchers, while as nodes of social networks they began to consider not only representatives of society, but also other objects that have social connections. Today, the term "social network" denotes a concept that turned out to be wider than its social aspect, it includes, for example, many information networks, including the WWW. Consider not only statistical, but also dynamic networks, to understand the structure of which it is necessary to take into account the principles of their evolution.

Today, the term "social networks" (*Social Networks*) means, first of all, online services on the Internet, designed to form, display and streamline social relationships. Social media features:

- 1) providing users with a wide range of opportunities for information exchange;
- 2) creation of user profiles in which it is required to specify a certain amount of personal information;
- 3) friends in a social network are mostly not virtual, but real friends.

The social network web resource provides the following opportunities:

- 1) active communication;
- 2) creating a public or private profile (*Profile*) of the user containing personal data;
- 3) organization and maintenance by the user of a list of other users with whom he has some social relationship;
- 4) viewing links between users within the social network;
- 5) formation of groups of users by interests;
- 6) manage content within your profile;
- 7) content syndication;
- 8) connecting various applications.

Social media is a collection of online services and Internet applications that allow users to communicate with each other, including in real time. At the same time, users can exchange opinions, news, information, including multimedia among themselves.

Social media is based on the ideological and technological basis of web 2.0, which allows the creation and exchange of content created by the users themselves (User- Generated Content), in contrast to the previous concept of the web, which, as in the case of traditional media, implies the centralized creation of content supplied by reader users.

Obviously, social media is the most valuable source of information for competitive intelligence, providing absolutely legal conditions for versatile information about people, events, companies, brands, products. Recently, such phenomena as information operations, active information countermeasures within the framework of competition, network mobilization, which have

recently become widespread, are in many cases based on the manipulation of data precisely in social media.

There are seven types of social media, these are social networks; blogs; forums; review sites; photo and video hosting servers; virtual dating services and geosocial networks. It should be noted that the clear boundaries between these varieties are blurred.

A social network on the Internet (social networking service) is an online service designed to build, display and organize social relationships, providing a wide range of opportunities for exchanging information, the user's ability to provide information about himself (create his profile), build connections, find friends by interests, connect relatives, colleagues, classmates, etc.

A blog (blog, from web log) is understood as a website, the main content of which is entries (text, images or multimedia) periodically added by users. Blogs are characterized by short entries (especially in the case of so-called "microblogs") of temporary significance, blogs are usually public and involve third-party readers who can enter into a public debate with the author (in comments on a blog entry or on their blogs). The totality of all blogs on the Internet is called the blogosphere.

Web forums are web applications designed to organize communication between visitors to certain Internet resources (websites or portals). On the resources of the web forum, users set topics of interest to them, which are then discussed by other users by posting messages (posting) within these topics.

Review websites are created to improve the efficiency and quality of the provided (not necessarily online) services and products. Users, visiting review websites, leave their messages there, participate in surveys, form opinions about a particular service or product.

Photo hosting is a website that allows you to publish any images (most often digital photos) on the Internet. The main advantage of photo hosting is the convenience of displaying the posted photos. Accordingly, video hosting is a website that allows you to download and view video information in a web browser. Video hosting is gaining popularity due to the development of broadband Internet access.

A virtual dating service is an Internet service that provides services for virtual acquaintance of users with the goals of com-

munication, creating a family, serious relationships, etc. When using a virtual dating service, the user creates a profile in which he indicates his pseudonym (nickname) and other parameters requested by the service (gender, age, purpose of acquaintance, interests, photos). After registration, the user can communicate with other users, receive messages and reply to them.

Geosocial networks (Geo Social Network) are a type of social networks in which users leave data about their location, which allows them to unite and coordinate their actions based on information about what people are present in certain places, what events take place in these places.

3.2.1. Major social networks

The list of the largest social networks that may be of interest to competitive intelligence includes:

- Facebook;
- Twitter ;
- LinkedIn;
- Sina Weibo;
- YouTube;
- Telegram;
- medium;
- reddit;
- livejournal.

Facebook (www.facebook.com) is the largest social network founded in 2004 by M. Zuckerberg and his associates. Since September 2006, the social network has been available to Internet users. As of June 2017, *Facebook* had 2 billion users. Daily active audience in March amounted to 720 million people. Approximately 500 million people use the *Facebook mobile applications per month*. Every day on the social network, users leave 6 billion "likes" and comments and publish 300 million photos. The site recorded 200 billion "friendships". Facebook has more than 1 trillion monthly page views.

Twitter (twitter.com) is a service that allows users to send short text notes (up to 140 characters) using the web interface, SMS, instant messaging tools, or third-party client programs. Created by Jack. *Twitter* is owned by *Twitter Inc.*, headquartered

in San Francisco. As of January 1, 2011, the service has more than 200 million users. 100 million users are active at least once a month, of which 50 million use *Twitter* daily. 55% use Twitter on mobile gadgets, about 400 million unique visits per month directly to the site twitter.com. A feature of *Twitter* is the public availability of posted messages; it's called microblogging.

LinkedIn (www.linkedin.com) is a social network for finding and establishing business contacts. LinkedIn was founded by Reed Hoffman in December 2002 and launched in May 2003. On June 13, 2016, Microsoft announced the acquisition of *LinkedIn* for \$196 per share (total deal value \$26.2 billion), Microsoft's largest acquisition to date. At the end of July 2020, *LinkedIn* announced the layoff of 960 employees; the cuts were due to the effects of the global Covid-19 pandemic. This social network provides an opportunity for registered users to create and maintain a list of business contacts. Contacts can be invited both from the site and from outside, however, *LinkedIn* requires prior acquaintance with the contacts. The *LinkedIn* contact list can be used to expand connections, find companies, people and interest groups, post resumes and find jobs, recommend users, post jobs, create interest groups. The *LinkedIn* social network also allows you to publish information about business trips and conferences. As of 2020, the total number of *LinkedIn* users has reached 675 million, of which 310 million are active.

Sina weibo (Chinese新浪微博, <http://weibo.com>) is a microblogging service launched by Sina Corp on August 14, 2009, one of the most popular Internet services (social media platforms) in China and the world. In early 2018, it surpassed a market valuation of US\$30 billion. As of February 2013, the number of service users is over 500 million. In June 2020, Sina Weibo reached 523 million monthly active users.

Youtube (youtube.com) is a video hosting service that provides users with video storage, delivery and display services. Created in February 2005 by three former PayPal employees Chad Hurley, Steve Chen and Jawed Karim, the service was bought by Google in November 2006 for US\$1.65 billion. According to the Alexa Web Rankings, *YouTube* is the second most visited website after Google Search. Users can download, view, rate,

comment, add to favorites and share certain videos. As of 2019, about 300 hours of video are uploaded to *YouTube every minute*, and the number of daily video views has reached 5 billion.

Telegram (telegram.org) is a cross – platform messenger that allows you to exchange messages and media files in many formats. Users can send messages and exchange photos, stickers, voice and video messages, files of any type, as well as make audio and video calls. The number of monthly active users of the service, as of January 2021, is about 500 million people. As of March 2020 official clients for *Telegram* include:

- Mobile applications for Android and iOS /iPadOS;
- Desktop apps for Windows, Linux and macOS;
- Web app, Chrome app web apps, React web app.

Since January 28, 2021, *Telegram* has the ability to import chats and correspondence history from other messengers, including *WhatsApp*.

Medium (medium.com) is a platform for social journalism. The service was launched in August 2012 by *Twitter co-founders* Evan Williams and Biz Stone. Williams, previously the co-founder of Blogger and Twitter, originally developed *Medium* as a medium for posting emails and documents that exceed *Twitter's maximum 140 characters* (now 280 characters). There are 15 authors and editors among the 75 employees of the company. The platform releases editions of Matter, Cuepoint, Backchannel , Re:form, Vantage, and The Nib. As of May 2017, *Medium* had 60 million unique readers per month.

Reddit (reddit. com) is a social news site where registered users can post links to any information they like on the Internet. Like many other similar sites, *Reddit* is one of the most popular sites in the world, ranked 19th in terms of traffic according to Alexa Internet. *Reddit* was founded on June 23, 2005 by University of Virginia alumni Steve Huffman and Alexis Ohanian. *Reddit* has an estimated value of \$6 billion. In 2019, there were about 430 million monthly *Reddit users*, known as "redditors".

LiveJournal, LJ (www.livejournal.com) is an online journaling (blogging) platform created in 1999. American program-

mer Brad Fitzpatrick. *LiveJournal* provides users with the opportunity to publish their own and comment on other people's posts, maintain collective blogs ("communities"), add other users as friends and follow their posts in the "friends feed". Until the end of December 2016, *LiveJournal* servers were located in the United States and the system belonged to the American company LiveJournal, Inc., but since December 2016, *LiveJournal* has been hosted on the servers of the Russian company Rambler&Co. Among the options of "LiveJournal" should be highlighted: different types of entries and the possibility of commenting on them; specifying extended information about the user; friends and friends feed; user pictures; account security features. At the end of 2012, *LiveJournal* had over 40 million registered users, of which 368,805 were active.

3.2.2. Social media monitoring

Social media monitoring is the most important stage for successful business development, promotion on the Internet, competitive intelligence. With the help of social media, you can find out the most complete information about the audience of a product or service, its opinion about the work of the company.

Here is an example of several services for effective social media monitoring, focusing on the most accessible:

Socialmention (www.socialmention.com) is a platform for free search and analysis of information in social networks. The slogan is "Search and analysis in social networks in real time." The system searches for mentions in selected networks or in all networks at once. Provides sentiment analysis, related keywords, popular sources, and more. System coverage – more than 100 social media, including social networks, social bookmarks, blogs, forums and more.

Hootsuite (hootsuite.com, seesmic.com) is a social media monitoring service. The slogan of the service is "Social networks are your superpower". Supports monitoring of resources such as: Twitter, Facebook, LinkedIn, Chatter, Ping.fm. There are applications for both web and personal computer, iPhone, Android, Windows Mobile. HootSuite is a certified Twitter partner. Provides scheduled posting, the ability to track posts by keywords

and mentions. The HootSuite system also provides full integration with Facebook.

YouScan (www.youscan.io) – Russian-language social media monitoring system. The slogan is "Use the power of social media to make the right decisions." The YouScan system monitors mentions in blogs, forums, social networks (Facebook, VKontakte), Twitter, YouTube, and provides monitoring results in an analytical interface with the functions of simultaneous work of several employees. Provides reports on the number of posts with mentions of keywords, authors, sources, sentiment.

IQBuzz (www.iqbuzz.ru) – a service for monitoring social media – a large number of sources and sites, such as LiveInternet, LiveJournal, Twitter, Yandex.Blogs, RuTube and YouTube video hosting services, various news, entertainment, specialized, thematic and regional portals. The system provides round-the-clock monitoring, allows you to receive information practically in real time mode. IQBuzz System allows you to determine the tone of user messages, analyze the socio-demographic characteristics of their authors on based on information from social media profiles. It is possible to connect according to user requests new sources for monitoring.

Socialbakers (www.socialbakers.com) is a unified social media analytics marketing platform, a social media statistics collection service that calls itself " the heart of Facebook statistics ". The Socialbakers system is known for its brand rankings on Facebook, in different categories. In addition, the Socialbakers service provides the ability to monitor information on networks such as Twitter, Youtube, LinkedIn.

PeerIndex (www.peerindex.net) is a free social media analysis service, primarily Twitter. Determines the size of the "social capital" or influence of a company, professional, publication, etc. Offers the largest database of Twitter users, which allows you to discover online communities based on interests, demographics, location, profession.

PostRank (www.postrank.com) is a Google service that allows you to analyze data in real time on topics, trends, events related to a person or business.

Trackur (www.trackur.com) is a commercial online social media reputation monitoring and analysis tool. Allows you to track the reputation of brands on news websites, blogs, forums, social networks.

Babkee (www.babkee.ru) is a comprehensive automated social media monitoring system. Allows you to solve such tasks as protecting the company's reputation, creating a positive image, as well as performing marketing and PR, services in the field of Social Media Marketing.

Semantic force (www.semanticforce.net) – a service that provides monitoring of unstructured sources – comments in online media and online stores (Fig. 80). The system uses SemanticForce W3Monitor technology, which monitors changes on any resources, including sites without RSS, page fragments, comments on publications and forum discussions. SemanticForce automatically indexes the texts of articles referenced by messages (tweets) in microblogs. This allows you to find indirect references to the object and significantly expand the coverage. SemanticForce's own search algorithms are used to monitor popular social networks: Facebook, VKontakte, GooglePlus. Morphological features and the specifics of a particular network are taken into account, which can significantly increase the volume of tracked mentions.

To monitor high-frequency objects, Twitter Firehose technology is used, which allows you to receive data from Twitter without time delays and restrictions on the amount of uploaded information. SemanticForce technology allows you to track the behavior of the author of mentions and his attitude towards the monitored object, automatically search for his profiles on the Internet and collect history for the purpose of subsequent analysis and involvement. Geo-segmentation is also possible, which determines the geographical location of the author of the message. To analyze messages, object detection is used – Automatic selection and statistics on companies, products and persons

mentioned in the texts. Hierarchical clustering provides navigation through a large amount of data, highlighting clusters by individual words, which are often mentioned in the context of monitoring objects.

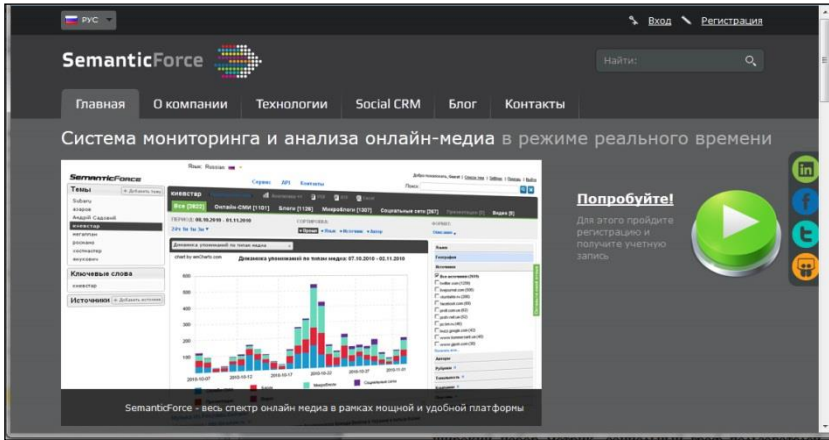


Figure 80 – SemanticForce service website fragment

Sentiment is determined not for the entire message in the system, but for a specific object in the mention, which allows you to form samples with different tonality – for example, in the case when in one message a certain brand is spoken positively, and its competitor is spoken negatively. The SemanticForce platform implements a special architecture for storing, searching and visualizing comments, which allows you to see comments under the original article or note to which they were originally left. The SemanticForce platform combines media and web analytics. The most popular web analytics service Google Analytics is integrated into the system. Analytical data from Google Analytics can be found in the source report.

Kribrum (www.kribrum.ru) is a system for monitoring and analyzing social networks that allows you to track and analyze mentions of a brand, products or services, key persons, events, geographical names. The system automatically determines the emotional coloring of statements and distributes publications by

tags and categories. The system belongs to Ashmanov & Partners and is positioned as a product for managing the company's reputation.

3.2.3. Social network analysis

As an example of the application of the possibilities of analyzing social networks, we present a fragment of the study of a network of connections between concepts (surnames of persons) extracted from corpora of unstructured texts – arrays of documents scanned from the Internet by the InfoStream content monitoring system [Grigoriev, 2007].

When constructing a network of concepts, algorithms for automatic extraction of concepts from unstructured texts were used. It should be noted that approaches to extracting various types of concepts from texts differ significantly both in the context of their presentation and in structural features. So, to identify whether a document belongs to a category of a thematic classifier, specially composed queries can be used, including logical and contextual operators, brackets, etc.

The identification of geographical names also involves the use of tables in which, in addition to the templates for writing these names, country codes, names of regions and settlements are used.

Another kind of concepts, such as "persons", is extracted from texts based on rules that take into account tables of valid names and surnames, initial patterns, possible options for joint spelling of initials / names and surnames.

It should be noted that the InfoStream system includes concept extraction tools and, among other things, provides users with results in the form of "information portraits" that include concepts such as keywords, place names, person names, company names, etc. Within the framework of this system, the properties of networks formed by concepts connected with each other by references in the same documents are analyzed.

The network formed by concepts extracted from text streams is not static, but depends on the volumes of documents from which the corresponding concepts are extracted. Therefore, to understand the structure of such a network, it is necessary to take into account its evolution.

The edges of the original network are assigned weight values equal to the number of documents in which there are mentions of persons corresponding to the nodes. To prevent "noise", edges with a weight less than 2 were not taken into account. With the development of a network with a fixed number of persons, with an increase in the number of considered documents, the average distance between nodes decreases accordingly, reaching its logical saturation.

An interesting fact is that the nodes of the network of persons under consideration with the maximum number of outgoing edges predominantly have the highest level of mediation and cannot be considered as the basis for building clusters with automatic grouping, but rather as elements connecting separate groups of nodes.

The information network of persons (the most mentioned) and their connections, obtained by analyzing the output stream on economic issues for a certain period of time, is presented in the form of a graph (Fig. 81).

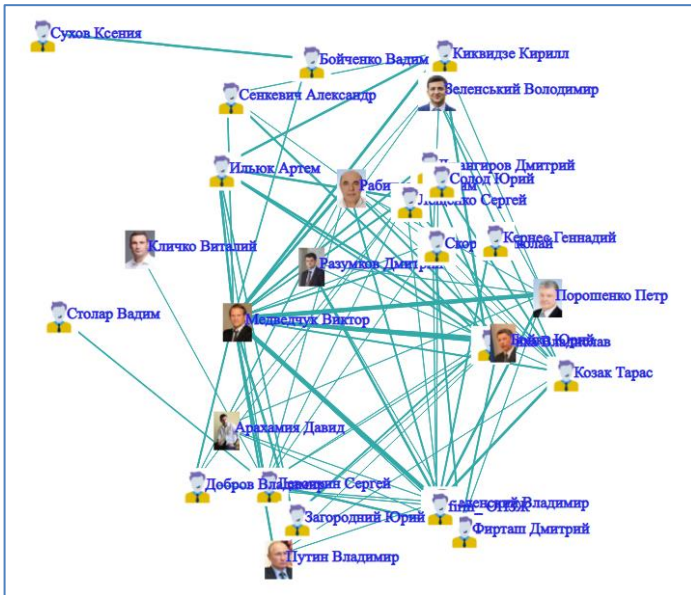


Figure 81 – Displaying the information network of persons

The empirical results obtained can be useful, for example, in modeling economic and social processes, identifying and visualizing implicit relationships between individual objects or subjects. The phenomenon of stabilization of the network of connections in practice allows, by analyzing a small array of documents, to identify stable connections, reduce the influence of noise factors. At the same time, the question of assessing the correlation of the obtained information relationships of persons, calculated by counting the frequency of documents in which persons are mentioned together, and real relationships remains open.

3.3. Deep web, special databases

Recent studies of the web space have shown that the more than a trillion web pages available through traditional information retrieval systems are just the "superficial visible part of the iceberg."

An important problem is the search for information in the "hidden" or "deep" web space, which, as noted above, contains an incomparably greater amount of data potentially interesting for competitive intelligence than in the open part of the Internet.

These are, first of all, dynamic web pages, information from numerous databases, which can be of great interest for analytical work. The "hidden" web also includes full-text information systems such as LexisNexis or Factiva.

The "hidden" resources of the Internet can also include peer-to-peer networks such as BitTorrent, EDonkey, EMule, Gnutella, Kazaa.

As noted earlier, there is much more information (including for competitive intelligence) on the Internet than it is covered by universal search engines.

It is assumed that, unlike the "cognizable" part of the Internet, the "hidden" part turned out to be hundreds of times larger.

A business analyst is often faced with a situation where he is aware of the existence of a document in the web space, but cannot find it using traditional search engines, which systems such as Google, Yahoo!, Bing, Baidu, Rambler can be considered today. or Meta. However, remembering or finding the address (URL) of this document in bookmarks, he can easily access it. That is, this document exists in the web space, but it cannot be

found in the usual way. The user encountered an invisible (*invisible*) resource for search engines.

3.3.1. The concept of "deep web"

The collection of sources on the web that are inaccessible to users of traditional search engines forms the so-called "deep web" – a concept introduced by Jill Ellsworth in 1994. Under the deep web (invisible web, deep web, hidden web) it is customary to understand that part of the web space that is not indexed by robots (web crawlers) of search engines. Using an analogy, information, being inaccessible to search, is "in depth" (English – deep). At the same time, you should not confuse the deep web with resources that are completely inaccessible from the Internet – this is the dark web (dark web), and we will not talk about it here. Some resources, access to which is open only to registered users, also belong to the deep web.

In 2000, the American company BrightPlanet (www.brightplanet.com) published a sensational report stating that there are hundreds of times more pages on the web than the most popular search engines at the time could index. The company developed the LexiBot program, which allows you to scan some dynamic web pages generated from databases, and when you run it, it received unexpected data. It turned out that there are 500 times more documents in the deep web than are available through search engines. Of course, these numbers are inaccurate. In addition, it became known that the average page of the deep web is 27% smaller than the average page from the visible part of the web.

Today the situation has changed, for example, leading search engines can index documents submitted in formats containing text. Of course, this is, first of all, pdf, rtf and doc. In 2006, Google patented a way to search the deep web: " Searching through content which is accessible through web – based forms " (Fig. 82). According to various authors, only 20-30% of the web space belongs to the visible web.

The screenshot shows the WIPO IP Services website interface. At the top, the WIPO logo and 'IP SERVICES' are displayed. Below this is a navigation bar with links for 'ABOUT WIPO', 'IP SERVICES', 'PROGRAM ACTIVITIES', 'RESOURCES', and 'NEWS & EVENTS'. A breadcrumb trail indicates the current page is 'Home > IP Services > PATENTSCOPE® > Patent Search'.

On the left side, there is a 'PATENTSCOPE®' menu with links for 'About Patents', 'PCT Resources', 'PCT Service Center', 'Database Search', 'PCT Applications', 'National Collections & PCT', 'External Databases', 'Patent Analysis', 'Glossary', 'Data Services', 'Publications', 'Projects & Programs', 'Patent Law', and 'Priority Documents'. Below this is a 'RELATED LINKS' section with links for 'WIPO GOLD', 'Patent Classification: IPC', 'Statistics', 'Life Sciences', and 'WIPO Standards'. At the bottom left is an 'E-NEWSLETTERS' section with a 'Subscription' link.

The main content area features a notice: 'This page is being phased out of production, but will remain available during the transition to our new system. Please try the new PATENTSCOPE® International and National Collections search page (English only)'. Below the notice is a heading: '(WO/2006/108069) SEARCHING THROUGH CONTENT WHICH IS ACCESSIBLE THROUGH WEB-BASED FORMS'. A navigation bar below the heading includes tabs for 'Biblio. Data', 'Description', 'Claims', 'National Phase', 'Notices', and 'Documents', with 'Biblio. Data' selected.

The selected tab displays the following bibliographic data:

Latest bibliographic data on file with the International Bureau		
Pub. No.:	WO/2006/108069	International Application No.: PCT/US2006/012734
Publication Date:	12.10.2006	International Filing Date: 04.04.2006
IPC:	G06F 17/30 (2006.01)	
Applicants:	GOOGLE, INC. [US/US]; 1600 Amphitheatre Parkway, Bldg. 47, Mountain View, CA 94043 (US) (All Except US). HALEVY, Alon Y. [US/US]; (US) (US Only). MADHAVAN, Jayant [IN/US]; (US) (US Only). KO, David H. [CN/US]; (US) (US Only).	
Inventors:	HALEVY, Alon Y.; (US). MADHAVAN, Jayant; (US). KO, David H.; (US).	
Agent:	PARK, A. Richard; 2820 Fifth Street, Davis, CA 95616 (US).	
Priority Data:	60/669,292 06.04.2005 US	
Title:	SEARCHING THROUGH CONTENT WHICH IS ACCESSIBLE THROUGH WEB-BASED FORMS	

Figure 82 – Fragment of the WIPO web resource describing the Google patent to deep web search

3.3.2. Causes

The deep web contains web resources that are not linked to other resources by hyperlinks, such as pages that are dynamically created by database queries, documents from databases that are available to users through web search forms (but not via hyperlinks). Such documents remain inaccessible to the robot, which is unable to correctly fill in the form fields with values in real time (form database queries).

Here is what the book [Price, 2001] says about the deep web : “Most of the pages on the invisible web can technically be indexed, but are not indexed because search engines have chosen not to index them... Most of the invisible sites have high-quality content. It’s just that these resources can’t be found with general purpose search engines...

... Some sites use database technology, which is really difficult for a search engine. Other sites, however, use a combination of files that contain text and media, and so some of them may be indexed and some may not.

... Some sites may be indexed by search engines, but this is not done because the search engines find it impractical – for

example, because of the cost or because the data is so short-lived that it simply does not make sense to index it – for example, weather forecast, exact arrival time specific aircraft landing at the airport, etc.. »

The main limitations associated with search engine robots can be explained by the following main reasons: for public search services, it is more important to ensure the accuracy of the search than completeness, sometimes it is more important to ensure that a response to a query is received in a reasonable time than accuracy. Hence, there are restrictions on the depth of scanning web resources, attempts to "filter" content by content, screening out pages containing excessive output hyperlinks, etc. At the same time, the child often splashes out with the water.

It is generally accepted that the value of the resources of the deep web is sometimes higher than the value of the resources of the visible part of the web space.

One more reason to replenish the deep web can be mentioned – the owners deliberately do not want their web resources to be found using search engines. Most often, such web resources represent something not quite legal, hacker forums, archives of unauthorized content, etc. It is clear that many of these resources are very interesting for business analysts to explore.

Many companies first connect to the public network, and only then spend a lot of money on protection. Site owners can try to prevent the indexing of certain pages of their resources by writing a prohibition command in the robots.txt file, but search engines can ignore it. Therefore, such resources either remove or remove hyperlinks, translating the resources into the deep web. For example, some business directories refuse to submit their ads to search engine robots, ie. protecting their information assets, companies transfer their resources to the deep web.

3.3.2. Types of Deep Web Resources

There are several types of deep web resources, for example, as noted above, these can be quickly outdated web pages. In addition, the deep web includes web resources that represent multimedia information. As is known, at present there are no satisfactory algorithms for searching for non-textual information. Dynamically generated on-demand pages also often find their way into the deep web. Often, without a request, such pages do not

exist; they are generated when querying databases. It turns out that information seems to be present in the web space, but it appears only at the time of processing the request, and there is no universal algorithm for filling them with search form robots. And, finally, if no links lead to a web resource, then search engine robots cannot find out about its existence in any way.

BrightPlanet founder Michael K. Bergman was able to identify 12 types of deep web resources that belong to the class of online databases. The list included both traditional databases (patents, medicine and finance) and public resources – job search ads, chats, libraries, reference books. Bergman also included specialized search engines that serve certain industries or markets, the databases of which are not included in the global catalogs of traditional search services.

The deep web also includes numerous systems of interactive interaction with users – help, advice, training, requiring the participation of people to form dynamic responses from servers. They also include closed (in whole or in part) information available to Web users only from certain addresses, groups of addresses, sometimes cities or countries. To the "hidden" part of the Web, many also include web pages registered on free servers, which are indexed, at best, only partially – search engines do not seek to bypass them in full to avoid advertising spam.

The deep web also includes the category of so-called "gray" sites that operate on the basis of dynamic content management systems (Dynamic Content Management Systems). Search engines usually limit the depth of indexing of such sites in order to avoid possible cyclic viewing of the same pages.

3.3.2. Deep Web Resources

So how do you find web resources hosted on the deep web? If the resources require filling in special forms, supplemented, for example, with captchas, then you need to go to the database that supposedly contains the necessary documents. Finding databases – sources of the hidden web can be done using conventional search engines by summarizing the query and entering clarifying words, such as "database", "data bank", "database", etc.

Let's take a well-known example: a user wants statistics on plane crashes in Argentina. A natural query on a traditional search engine returns a huge list of newspaper headlines. On request "aviation database", you can immediately go to the NTSB Aviation Accident Database (www.nts.gov/ntsb/query.asp).

For searching in the deep web, namely in the segment that makes up databases, some specialized resources already exist today. The leader among deep web navigators is BrightPlanet's CompletePlanet (www.completeplanet.com). This site is the largest directory with over 100,000 links. BrightPlanet also created a personal online database search utility, LexiBot, that can search thousands of deep web search engines. The same company's DeepQueryManager (DQM) metasearch package provides searches for more than 70,000 "hidden" web resources.

A study carried out in 2006 [He, 2007] showed that the Deep Web has over 300,000 sites linked to over 450,000 databases not covered by traditional search engines. The most interesting deep web resources for business analysts include: databases of legal entities and individuals; industry databases; reputation databases (black and white lists); criminological databases; databases of goods and services; product catalogs, etc. World famous business resources hosted on the deep web include: amazon.com, ebay.com, realtor.com, cars.com, imdb.com.

Here are some more examples of deep web databases and catalogs:

FindLaw (www.findlaw.com), one of the world's most popular legal websites, is a large directory of legal resources containing an annotated list of freely available databases of legal documents for which this resource is the "entry point". A fragment of the website of the FindLaw service is shown in Fig. 83.



Figure 83 – Fragment of the FindLaw website

Politicalinformation.com (www.politicalinformation.com) is a resource for journalists, politicians, students and politicians, a service that provides online search in 5000 selected political websites, providing news from dozens of authoritative sources.

The academic search engine Biefield Bielefeld BASE (<https://archive-it.org/>) is one of the world's largest search engines for academic web resources. Access to the full texts of about 60% of indexed documents is provided free of charge (open access). BASE is administered by the Bielefeld University Library.

CiteSeerX (<https://citeseerx.ist.psu.edu/index>) is an electronic scientific literature library and search engine. The service was created to disseminate scientific literature and improve the functionality, usability, accessibility, cost, completeness, efficiency and timeliness of access to scientific and scientific knowledge.

Data.gov (<https://www.data.gov/>) is the "house of data". Under the terms of the 2013 Federal Open Data Policy, newly created government data must be available in open, machine-readable formats while maintaining confidentiality and security.

Mednar

(<https://mednar.com/mednar/desktop/en/search.html>) is a free deep web search engine focused on medicine. Since Mednar is a public search, it cannot retrieve results from a personal subscription or additional medical resources.

World Wide Sicence (<https://worldwidescience.org/>) is a global scientific portal consisting of scientific databases and portals.

The peculiarity of most "hidden" resources lies in their narrow specialization. For searching, they use the same mechanisms as for the "surface" web, however, in most cases, search engine robots for the deep web include data access modules unique for each such resource.

A traditional search engine can most often give you the address of a database, but it won't tell you exactly what documents it contains. A typical example is information retrieval systems for Ukrainian (zakon.rada.gov.ua) or Russian legislation (www.kodeks.ru). Thousands of documents from databases become available only after logging in, and the robots of standard search engines are not able to index the content of databases.

Paradoxically, the archive of open web space resources can also be considered as one of the resources of the deep web. Such an archive, the Internet Archive, has been created since 1996 (www.archive.org). Today, the Alexa database exceeds 538 billion web pages (Fig. 84), 28 million books and texts, 14 million audio recordings, 6 million videos, 3.5 million images, 580,000 computer programs.

The Internet Archive's storage technology includes a number of state-of-the-art tools for managing a giant document repository. For example, using this technology, clustering of web resources is performed, i.e. formation of collections of documents related to topics. Of particular interest to users of the Internet Archive service is the Wayback Machine, which provides access to time slices of the web space. One of the most interesting practical applications of this technology is the recovery of documents

that were once published on the web but subsequently deleted. At the same time, the growth of the deep web threatens with serious problems of completeness in the system storage associated with an increasing number of sites using various content management technologies, dynamic publication of documents from databases, etc.

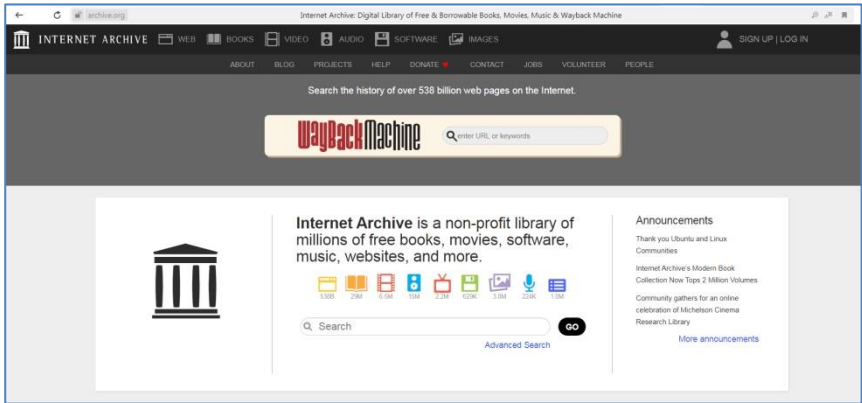


Figure 84 – The home page of the website www.archive.org

3.3.3. Deep web services

Traditional search engines are trying to narrow the space of the deep web, gradually capturing niches such as blogs, scientific sites, news agencies. So, as ancillary services for searching the deep web from Google, we can recommend: Google Book Search (books.google.com) – search for books, Google Scholar (scholar.google.com) – search for scientific publications, Google Code Search (code.google.com) – search for program code.

Goldfire Research System by companies Invention Machine Corp. (inventionmachine.com) allows you to process deep web content hosted on more than 2,000 US government, academic, research, and commercial sites. The Goldfire Research system knows about the mechanisms for accessing deep web databases and automatically generates queries to them.

Existing tools for analyzing and promoting web resources allow a new approach to assessing the ratio of visible and deep web volumes. For example, the website www.cy-pr.com provides information about the actual number of documents on the website under study, and about the number of documents indexed by various search engines, including Google. Having received a representative sample of sites, for example, according to the top100 rating (top100.rambler.ru), you can get an estimate of the ratio of the visible and deep parts of the web space.

As calculations show, the amount of information found in the deep part of the web space exceeds the amount of information from the visible part by about 3-5 times. It turns out, with rare exceptions, that the larger the resource, the more of it belongs to the deep web. In this sense, small web resources benefit from accessibility. Since a large proportion of news documents end up in the deep web, business intelligence tasks require special services to access such information. Such a service is provided by news content integration services – online media archives. Business analysts actively use the largest archives of information from open sources "Integrum" (integrum.ru) and InfoStream (www.infostream.ua). It is the use of open sources that allows competitive intelligence to operate within the legal framework, but, at the same time, be highly effective.

It can be stated that the faster the web space grows, the worse it is covered by traditional directories and search engines. Due to the growth in the number of websites and portals using databases, dynamic content management systems, the emergence of new versions of information presentation formats, the deep web is growing very rapidly. On the one hand, the Internet as a huge repository increases the amount of information available "in principle", but on the other hand, information chaos is growing, the entropy of the network information space is increasing. An ever smaller part of information resources becomes available to users in reality.

Leading search engines are still trying to find technical possibilities for indexing database content and accessing closed websites, however, their tasks objectively diverge from those of business analysts – the orientation of traditional search services to the mass service is justified in this case. Thus, the niche for search engines in the deep web is becoming wider.

3.3.4. Special databases

As a rule, in order to successfully conduct competitive intelligence, a data bank should be created and maintained, including the following main databases [Lande, 2005]:

1. Competitors (current and potential);
2. Information about the market (trends, nomenclature, price, address information);
3. Technologies (products, exhibitions, conferences, GOSTs, quality);
4. Resources (raw materials, human and information resources);
5. Legislation (international, central, regional and departmental regulations);
6. General trends (politics, economics, regional features, sociology, demography).

Today, the main sources of information for competitive intelligence are the Internet, the press, and open databases. But if access to ordinary Internet resources can be considered conditionally free, then, in most cases, access to databases requires not only registration, but also payment for such services. In addition, almost all of them can be attributed to the so-called "hidden" web space.

Databases of customs, tax and statistical authorities, justice authorities and courts, chambers of commerce and industry, privatization authorities and stock markets, information, rating, analytical and other agencies, etc. are very popular among competitive intelligence specialists. Separate accessible databases of other regulatory bodies and organizations are also of great benefit.

Traditionally, competitive intelligence relies on sources of information such as published open access documents that contain product market overviews, information about new technologies, partnerships, mergers and acquisitions, job postings, trade shows and conferences, and so on. Therefore, in recent years, databases based on media archives, including online ones, have become increasingly popular.

The "Big Three" of global services dedicated to providing users with access to business and analytical information include **LexisNexis**, **Factiva** and **Internet Securities**.

The world's largest full-text online information system **LexisNexis** (www.lexisnexis.com), which contains over 2 billion documents from 45 thousand sources with an archive of more than 30 years of business information and more than 200 years of legal information, belongs to the category of "hidden » web (Fig. 85). An additional 14 million documents are added to the archives every week. Unlike the unstructured masses of the "shallow" Web, users of LexisNexis can use powerful search tools to obtain reliable and classified information.

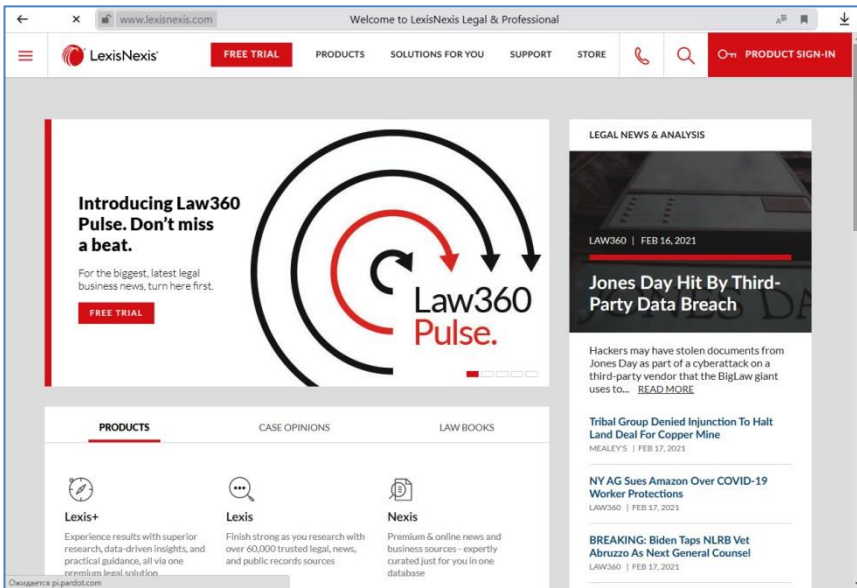


Figure 85 – Excerpt from the LexisNexis website

Factiva (global.factiva.com), a division **of** Dow Jones, provides access to business and analytical information. Factiva is based on over 35,000 primary sources from 159 countries. The Factiva database contains material on more than 36.5 million companies, as well as a complete collection of Investext information.

Internet **Securities** (www.internetsec.com), a brand of ISI Emerging Markets, covers 80 thematic information sections

formed from 16 thousand information sources – texts of articles, financial and analytical reports, corporate information, macroeconomic statistics, market data (Fig. 86). Main ISI Emerging Markets products : CEIC Data, Emerging Market Information Service (EMIS), Islamic Finance Information Service (IFIS), IntelliNews, ISI Compliance Edition, ISI DealWatch.

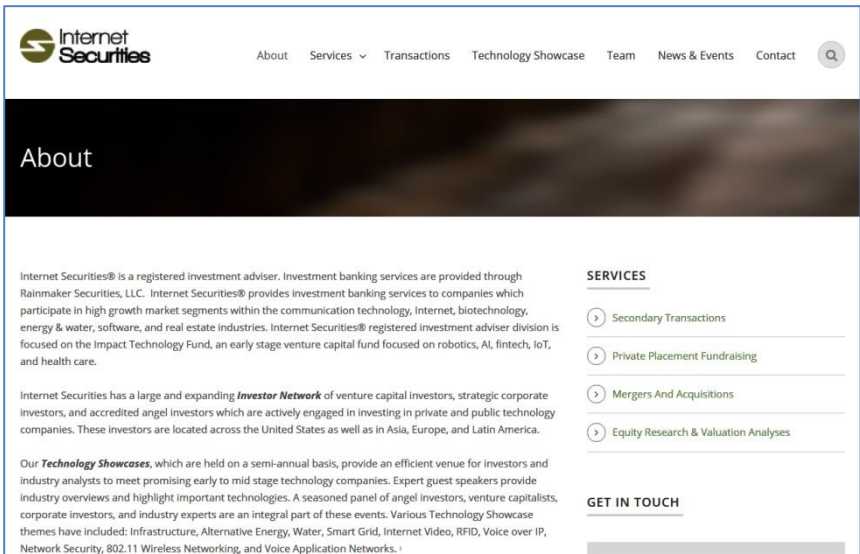


Fig. 86 – Fragment website _ _ Internet Security Services

Europages (www.europages.eu) – European Business Directory – B2B information retrieval system covering over 3 million suppliers, manufacturers and distributors in Europe and worldwide.

The task of a complete enumeration of all sources of information is practically impossible, since this market is very dynamic, new databases are constantly appearing, existing sources are merging, and weak ones are being absorbed by strong ones. At the same time, one of the rules of competitive intelligence is formulated as follows: “the more information is confirmed by a large number of independent sources, the more reliable it is.”

Along with databases, one of the most effective sources of information can be reports and references from outsourcing companies professionally engaged in competitive intelligence and collecting information about commercial structures and markets. Their products are, in fact, the result of competitive intelligence.

There are many such special companies in the world. One of these largest companies, which owns about 80% of the Western market, is the American company, **Dun & Bradstreet (D&B)**, whose database we mentioned above. Any company reference in this service is valued at an average of \$100 or more. A more serious market or competitor analysis can cost as much as \$10,000. Terms of execution – from several hours (information is present in the database) – up to several days for references and up to several months for analytical work.

Creditreform is no less famous, German **Schufa Holding AG** (479 million documents in the database, including 66 million about individuals), Austrian **Intercredit Information Holding**, Latvian **Coface IGK** (IGK System is known – a database of debtors, including information about current debts, lawsuits, as well as insolvency processes) and many others. Some of these companies combine competitive intelligence functions with other activities, such as credit bureau duties.

A common problem when applying for informational statements to Western agencies with representative offices in Ukraine is that, as a rule, the information provided in relation to Western non-residents is much more extensive and of better quality than that provided in relation to domestic firms. In this connection, in such cases it is advisable to contact local information companies, the results are cheaper and better.

There are a number of similar companies in Ukraine, among which are **Avesta-Ukraine**, **SIDCON Consulting Company**, **SKIF Interbank Security Service** and many others.

All domestic and foreign information companies have their own representative offices and accept orders on the Internet, and therefore they can be attributed to specific Internet sources.

It should also be noted that, despite the fact that in the case of ordering the services of an outsourcing company, it does most of the information work for the client, the final conclusions and decisions, recommendations for making management decisions still remain with him. Only the client can have all the necessary completeness of external and insider information.

4. Reputation analysis

4.1. The problem of reputation management

Reputation is a social assessment of a group of subjects about a person, a group of people or a company, formed on the basis of certain criteria.

The company's reputation is a set of evaluative ideas of the target audience about the company, formed on the basis of reputation factors that are important for this audience. According to the information letter of the Supreme Economic Court of Ukraine "On some issues of the practice of application of information legislation by economic courts" dated March 28, 2007, the business reputation of a legal entity is the prestige of its company (commercial) name, trademarks and other intangible assets belonging to it among the circle of consumers of its goods and services.

The success of a company is directly related to its reputation. Thus, a study conducted by Australian scientists P. Roberts and G. Dowling revealed that the higher the company's reputation, the longer the period during which it receives the maximum income from its activities, and, secondly, the less time is needed for a company to achieve industry average financial performance when introducing innovations. Reputational Capital is not only a marketing concept, it has no less relation to finance. The monetary equivalent of goodwill can be expressed in the form of goodwill. In accordance with International Financial Reporting Standards (IFRS), goodwill is the difference between the price paid for the enterprise by buyers and "fair value" (this value often differs significantly from the simple value of all the company's assets). For example, in accounting rules, reputation is defined as "the difference between an organization's purchase price and the balance sheet value of all of its assets and liabilities".

The financial return of a company is directly related to its reputation. Thus, a study conducted by Australian scientists G. Dowling and P. Roberts [Roberts, 2002] revealed two advantages of a favorable company image. Comparing the data of the ranking of the 500 best and most respected companies in the United States, compiled annually by the American Fortune magazine,

for 1984-1995 with the financial performance of companies over the same period, scientists have identified a relationship between the reputation of the company and its financial level. It turned out that the higher the reputation of the company, the longer, firstly, the period during which it receives the maximum income from its activities, and, secondly, the less time the company needs to achieve industry-average financial performance when introducing innovations..

To be able to ascertain the intangible value of a company, expert assessments of reputation are being developed. The cost of reputation can be determined by experts, for example, in this way. First, the income received by the company due to the brand is calculated (the difference between the real profit and the income that can be obtained by selling a non-branded product), and then the amount received is multiplied by a specially calculated coefficient (depending on the position of the company in the industry, the stability of financial indicators, etc.) The result is the price of the brand, which is an important part of the reputation.

There are also indirect assessments of the level of reputation of companies, for example, based on the results of a survey of company executives and analysts who evaluate companies in terms of such parameters as the quality of management and product, the ability to attract and retain qualified personnel, financial stability, efficient use of assets, investment attractiveness, application of new technologies, etc.

The concept of Online Reputation Management (ORM) is essentially a set of measures to detect negative content online and minimize it in social media and as a result of search results. This is a kind of PR campaign in cyberspace. A branch of ORM is SERM (Search engine reputation Management) – reputation management in search engines. In the West, such services are practiced very actively, and the growth of ORM per year is about 35-40%.

Today, according to Google statistics, 70% of users look for reviews about goods and services before buying them. Historically, the first company to practice two-way communication with customers on social networks was eBay. Based on the feedback, a rating of sellers was compiled, which buyers could rely on when making a purchase decision. In Russia, a vivid example of

displaying a company's reputation based on user reviews is the Yandex.Market system. More than half of Internet users, when choosing a particular product, company, customer, contractor, etc. based on information provided by other users.

Reputation management work is carried out both by specialized PR agencies operating in the vastness of the web space, and divisions of SEO agencies that launch PR campaigns aimed at finding and eliminating negative content. In addition, such services are provided by individuals – freelancers, specialists in the field of Internet marketing and SEO. Large companies have their own departments, whose work is aimed at managing the reputation of the company, brand, product, service.

The concept of "Online Reputation Management" (ORM) has already become an established term, and for these purposes in the West, a part of the budget of most large companies is annually allocated for this purpose. Along with the growing influence of social media on the views and preferences of people, the need for large companies to monitor their online image is also growing. Against this background, it does not seem strange that the ORM market is growing by 40% annually.

The main task of reputation management is to form a positive image about the company and its product. Since it's difficult to capture all user reviews and remove all the negatives, efforts are usually concentrated in three areas: search results, online media reviews, and social media mentions. We have to work both with content created by editors of various publications and ordinary users. To create a holistic positive image, the information from these three sources must be positive or neutral.

Reputation management in search engines – Search Engine Reputation Management (SERM) – a set of measures aimed at eliminating negative reviews about a company, product or service in the results of a search engine.

Reputation management service in search engines is necessary:

- companies wishing to exclude or minimize negative reviews about their activities (products);
- companies that want to generate positive reviews or increase their number and visibility for the target audience.

Negative information that harms the reputation on the network can be of various origins. Conventionally, there are three main groups of negative content origin [Kovalchuk, 2012]:

- unintentional negative – these can be both reviews of dissatisfied customers who have no intention of harming the company's reputation, but are simply not satisfied with the results of cooperation, as well as photos from corporate holidays carelessly posted on the Internet, statements of employees addressed to customers, etc. Usually this the negative does not pose a big threat, but in no case should it be ignored;
- intentional negativity with the aim of hurting reputation – in this case, the classic example is negative feedback from laid-off or resigned employees who are dissatisfied with the company's concept.
- a black PR campaign is the most dangerous type of negative content that deals a serious blow to reputation. Such PR campaigns are carried out by specialists who carefully study the competitor's business and know exactly where the Achilles' heel is hidden. Large raider seizures are being organized that can lead to a complete collapse not only of the reputation, but of the entire business as a whole. This service is ordered from PR-specialists by major serious competitors.

The most vulnerable topics in terms of attracting negative reviews are:

- banks, financial institutions;
- figures of politics and show business;
- tourism, travel (reviews about hotels, resorts, tour operators, air carriers);
- mobile technology and communications (operators, telephones, electronic tablets);
- Appliances;
- catering establishments (restaurants, cafes, bars).

How the monitoring space for reputation management is chosen by network resources where consumer reviews are posted:

- social networks, messengers;
- blogs and forums;

- thematic websites and portals;
- special review services.

Pages with positive content are promoted using standard search engine optimization tools (Search engine Optimization, SEO), such as link exchanges, buying, exchanging links to articles with thematic resources, posting announcements, news, etc. At the same time, positive content should be posted regularly, as negative content can reappear and damage reputation.

Search engine reputation management (SERM) is designed to fight negative content. The task of SERM is to oust web pages with unwanted information from search results, as a result of which the target audience will no longer see such resources, since users will not access them using search engines. To achieve this goal, materials with positive content are created, with the expectation that they will crowd out negative unwanted messages.

Online reputation management usually begins with monitoring search results and social media in order to identify information on a given object. There are several monitoring methods:

- manual monitoring of search engines by entering targeted search queries;
- use of alert systems integrated with search engines, such as Yandex.Blogs (blogs.yandex.ru) and Google Alerts (google.com/alerts). In these cases, relevant information is sent to the subscriber's email;
- use of special tools for monitoring the reputation of companies in social networks.

As a monitoring space for reputation management, network resources are chosen, where consumer reviews are posted:

- social media;
- blogs and forums;
- thematic websites and portals;
- special review services.

One of the criteria for the quality of the reputation monitoring service is the completeness of coverage – the share of information about the object, investigated during operation from the total amount of information in the network about the object. Traditional search engines are still the main tool for finding in-

formation, they cover a significant part of the Internet content, as well as some of the social media.

Today, there are hundreds of reputation monitoring systems all over the world, among which are Babkee, Brand-spotter, BuzzLook, Buzzware, IQBuzz, Kribrum, SemanticForce, Wobot, Youscan. In the research of Ken Barberi (Ken Burbary) and Adam Cohen (Adam Cohen) [Burbary, 2009-2013] provides a list of 230 reputation monitoring systems, many of which offer free trial periods to evaluate the quality of their work.

4.2. Reputation modeling in networks

Recently, in the framework of the theory of social network analysis, much attention has been paid to assessing the reputation of individual subjects (agents, social network nodes) and the level of trust between them [Rastorguev, 2006], [Gubanov, 2009].

Formally, a social network is a graph in which the set of vertices is a set of agents, subjects – individual or collective, and the set of edges is a set of relationships, a set of social connections between agents.

When modeling social networks, it becomes necessary to take into account the dynamics of social ties – the mutual influence of agents.

Influence in this case is considered as a process and result of a change by an individual (subject of influence) of the behavior of another subject – the object of influence, his attitudes and assessments in the course of interaction [Rastorguev, 2006]. Thus, influence is the ability to influence someone's ideas or actions [Novikov, 2002]. A distinction is made between directional and non-directional influence [Novikov, 2007]. Directed influence uses persuasion and suggestion as mechanisms of influence on another person. At the same time, the individual – the subject of influence – sets himself the task of achieving certain results from the object of influence. Non-directional (non-targeted, indirect) influence is an influence in which the individual does not set himself the task of achieving certain results from the object of influence.

The purposeful influence of social network participants (or subjects that are not part of the network, but use it as a tool of informational influence) is a special case of information management, which consists in shaping managed subjects of such

awareness that the decisions they make on its basis are most beneficial for the manager. subject.

The possibilities of influence of some participants in a social network on its other participants significantly depend on the reputation of the former. Reputation is “a general opinion that has been formed about the merits or demerits of someone, something, a public assessment” [Ermakov, 2005]. Reputation can be viewed as the "weight" of the community's opinion about an individual agent or a group of agents, determined by his views and activities (activity). At the same time, reputation can be both individual and collective.

Reputation increases if the agent's choice (answers to some key questions) matches what the community expects from him, and goes down otherwise.

Here is a formal definition of the reputation model [Gubanov, 2009].

Let be $\{a_1, a_2, \dots, a_n\}$ a set of agents – nodes of a social network that influence each other. We denote the influence matrix as $A = \|a_{ij}\|_{i=1, n}^{j=1, n}$ ($a_{ij} \geq 0$ denotes the degree of trust of *the i*-th agent to *the j*-th). At the same time, it is obvious that if *the i*-th agent affects *the j*-th, and *the j*-th influences *the k*-th, then this means the following: *the i*-th agent indirectly affects *the k*-th (transitivity), which allows you to build chains of indirect influences.

Assume that each agent at the initial moment of time has an opinion on some key issue. Let the opinion of the community of network agents reflect the vector of initial opinions b^0 of dimension n . The opinion of each agent changes under the influence of the opinions of other agents of the social network.

We will assume that the opinion of *the i*-th agent at the moment of time t is equal to

$$b_i^t = \sum_{j=1}^n a_{ji} b_j^{t-1}$$

In [Ermakov, 2005] it is shown that with multiple exchange of opinions, the opinions of agents converge to the resulting vector of opinions. $B = \lim_{t \rightarrow \infty} b^t$. Thus, the relation is true $B = Ab$.

Denote r_i is a parameter describing the reputation of the i -th agent in the social network (community), which can be defined as the normalized sum of his influences on all other agents of the social network (assumed to be $a_{ij} \geq 0$, $i, j = 1, \dots, n$), i.e.

$$r_i = \frac{\sum_{j=1, \dots, n, j \neq i} a_{ij}}{R}, \quad j = 1, \dots, n.$$

Here $R = \sum_k \sum_{j \neq k} a_{kj}$, $k, j = 1, \dots, n$, is the total mutual influence of all members of the social network on each other.

In accordance with the above expression, the agent i has the higher reputation, the higher his influence on all other members of the social network.

Modeling with the use of hypercomplex numerical systems (HCNS) makes it possible to use advanced tools from this area of mathematics.

Within the framework of the model based on the use of PNS, each subject (social network node) is characterized by its attitude to a number of important (key) issues (let their number be equal to N). Then, by analogy with [Lande, 2012], the subject A can be associated with a hypercomplex number with a dimension basis $2N$:

$$A = e_1 w_1^+ + e_2 w_1^- + \dots + e_{2N-1} w_N^+ + e_{2N} w_N^-.$$

At the same time, each question is assigned weight values w_i^+ and w_i^- , which correspond to the level of the subject's positive attitude to this question (w_i^+) or negative (w_i^-), which is a natural extension of the above approach. Both values can be in the interval $[0, 1]$, in some cases it can be assumed that $w_i^+ + w_i^- = 1$.

It is proposed to use the GChS of dimensions $2N$ with a basis $\{e_1, e_2, \dots, e_{2N}\}$ and a multiplication law, which can be presented in the form of a table:

	e_1	e_2	e_3	e_4	\dots	e_{2N-1}	e_{2N}
e_1	e_1	e_2	0	0	\dots	0	0
e_2	e_2	e_1	0	0	\dots	0	0
e_3	0	0	e_3	e_4	\dots	0	0
e_4	0	0	e_4	e_3	\dots	0	0
\dots	\dots	\dots	\dots	\dots	\dots	\dots	\dots
e_{2N-1}	0	0	0	0	0	e_{2N-1}	e_{2N}
e_{2N}	0	0	0	0	0	e_{2N}	e_{2N-1}

The model of the subject of the social network in this case is considered as a hypercomplex number of the form:

$$D = e_1 w_1^+ + e_2 w_1^- + e_3 w_2^+ + e_4 w_2^- + \dots + e_{2N-1} w_N^+ + e_{2N} w_N^-.$$

We can consider an assessment of the proximity of the opinions of two subjects $Est(A, B)$

$$A = e_1 a_1^+ + e_2 a_1^- + \dots + e_{2N-1} a_N^+ + e_{2N} a_N^- \text{ And}$$

$$B = e_1 b_1^+ + e_2 b_1^- + \dots + e_{2N-1} b_N^+ + e_{2N} b_N^-:$$

$$Est(A, B) = Norm\left(\frac{1}{N} \left(\sum_{i=1}^N (e_{2i-1} a_i^+ + e_{2i} a_i^-)(e_{2i-1} b_i^+ + e_{2i} b_i^-)\right)\right),$$

where $Norm(\cdot)$ is the function of the norm of a hypercomplex number: $Norm(e_{2i-1}) = e_1$, $Norm(e_{2i}) = -e_1$.

The attitude of most of the participants in the social network (society) to the selected issues is also represented as a hypercomplex number $Q = q_1 e_1 + q_2 e_2 + q_3 e_3 + \dots + q_{2N} e_{2N}$, just like a separate subject D . The greater the value of $Est(Q, D)$, the more loyal the subject, "relevant" to society.

One can also introduce another estimate of the proximity between hypercomplex numbers, by analogy with the norm of differences between ordinary vectors in a vector space:

$$Est_1(A, B) = Norm\left(\frac{1}{N} \left(\sum_{i=1}^N (e_{2i-1}a_i^+ - e_{2i-1}b_i^+)^2 (e_{2i}a_i^- - e_{2i}b_i^-)^2\right)\right).$$

In this case, the subject will be more loyal to society if the score $Est_1(Q, D)$ is lower.

At the same time, the second assessment in terms of content is less suitable for the tasks of identifying the mutual influence of subjects. For example, when comparing the relationship of the subject to society with the request, even with completely zero values of the weight coefficients related to the value of the entire social network (society), the sum in the expression given for the expression will not $Est_1(Q, D)$ be zero, i.e. depends entirely on the subject. Therefore, we confine ourselves to applying the first estimate.

Consider, for example, some simplified special cases in which the relations of society and the subject to one issue are analyzed.

1. Let the value corresponding to society be of the form:

$$Q = \frac{1}{2}e_1 + \frac{1}{2}e_2, \text{ i.e. the attitude to the chosen issue in society can}$$

be both positive and negative, with equal probability. Let the attitude of the subject to the same question be unambiguously positive, namely: $D = e_1$. In this case $Est(Q, D) = 0$, which corresponds to complete uncertainty.

2. Let the value corresponding to society, as in the previous case, have the form: $Q = \frac{1}{2}e_1 + \frac{1}{2}e_2$. Let the attitude of the subject

to the same question also have the form: $D = \frac{1}{2}e_1 + \frac{1}{2}e_2$. In this

case $Est(Q, D) = \frac{1}{4} + \frac{1}{4} - \frac{1}{2} = 0$, which, as in the previous case, corresponds to complete uncertainty.

3. Let the value corresponding to the society be of the form:

$Q = e_1$, and the attitude of the subject to the same question:

$$D = \frac{4}{5}e_1 + \frac{1}{5}e_2, \text{ then } Est(Q, D) = \frac{4}{5} - \frac{1}{5} = \frac{3}{5}.$$

It should be noted that not all zero elements of the above "ideal" table may actually be zero, but it is assumed that this table will be sparse. Rare non-zero elements in it can characterize the relationship of various issues.

The values of the coefficients for the basic elements of the images of the subjects of the social network can correspond to the probabilities of a positive (or negative) attitude of the subjects to the relevant issues. In this case, by renumbering the bases, the GChS multiplication table can be represented in the following form:

	e_1	e_2	e_3	e_4	\dots	e_{4N-3}	e_{4N-2}	e_{4N-1}	e_{4N}
e_1	B_1								
e_2									
e_3									
e_4									
\dots									
\dots									
e_{4N-3}						B_N			
e_{4N-2}									
e_{4N-1}									
e_{4N}									

where the block B_1 (and in the general case, any non-zero block) will have the form, supplemented by the coefficients w_i^j calculated during the training of the model:

	e_1	e_2	e_3	e_4
e_1	$w_1^1 e_1$	$w_1^2 e_2$	$w_1^3 e_3$	$w_1^4 e_4$
e_2	$w_2^1 e_1$	$w_2^2 e_2$	$w_2^3 e_3$	$w_2^4 e_4$
e_3	$w_3^1 e_1$	$w_3^2 e_2$	$w_3^3 e_3$	$w_3^4 e_4$

e_4	$w_2^2 e_1$	$w_2^1 e_1$	$w_3^2 e_4$	$w_3^1 e_3$
-------	-------------	-------------	-------------	-------------

At the same time w_1^1 - the weight of a positive attitude to the issue t_1 , w_1^2 - the weight of a negative attitude to the issue t_1 , w_2^1, w_2^2 - the weight of the relationships of the presence of contradictory simultaneous positive and negative attitudes to the issue t_1 .

Then for the document $A = a_1 e_1 + a_2 e_2 + a_3 e_3 + a_4 e_4$ and the document, $B = b_1 e_1 + b_2 e_2 + b_3 e_3 + b_4 e_4$ the proximity estimate will have the following form:

$$\begin{aligned} Sim(A \cdot B) = Norm & \left(\frac{1}{2} (e_1 (w_1^1 (a_1 b_1 + a_2 b_2) + w_1^2 (a_1 b_3 + a_2 b_4 + a_3 b_1 + a_4 b_2) + \right. \\ & + w_2^2 (a_1 b_4 + a_2 b_3 + a_3 b_2 + a_4 b_1)) + e_2 (w_1^2 (a_1 b_2 + a_2 b_1)) + \\ & \left. + e_3 (w_3^1 (a_3 b_3 + a_4 b_3)) + e_4 (w_3^2 (a_3 b_4 + a_4 b_4))) \right). \end{aligned}$$

In this case, the function *Norm* corresponds to the PNS, the table of which is used for the search. It should be borne in mind that the proposed table can be a component of a table of a larger dimension. The transition from a filled multiplication table to a weakly filled (sparse) one can be carried out by an isomorphic transition [Kalinovskiy, 2012], which will significantly reduce the number of operations when calculating the proximity function between subjects.

In general, the level of trust (proximity) between subjects $A = e_1 a_1^+ + e_2 a_1^- + \dots + e_{2N-1} a_N^+ + e_{2N} a_N^-$ and $B = e_1 b_1^+ + e_2 b_1^- + \dots + e_{2N-1} b_N^+ + e_{2N} b_N^-$, which can be interpreted as a degree of confidence, is also characterized by the function given above, taking into account the possible presence of non-zero off-diagonal elements of the table:

$$Sim(A, B) = Norm \left(\frac{1}{N} (e_1 a_1^+ + e_2 a_1^- + \dots + e_{2N-1} a_N^+ + e_{2N} a_N^-) \right) + (e_1 b_1^+ + e_2 b_1^- + \dots + e_{2N-1} b_N^+ + e_{2N} b_N^-)$$

The reputation of the subject $A_i = e_1 a_{i1}^+ + e_2 a_{i1}^- + \dots + e_{2N-1} a_{iN}^+ + e_{2N} a_{iN}^-$ within the entire social network (i.e. in relation to society) is defined as a normalized sum of levels of trust with all other subjects:

$$Trust(A_i) = \frac{\sum_{j \neq i} Sim(A_i, A_j)}{\sum_{k \neq j} Sim(A_k, A_j)}.$$

To assess the level of mutual influence of subjects of a social network, other methods can also be used, among which are: calculation of a measure of mutual information (mutual information), calculation of the modified Dice coefficient (modified Dice coefficient), occurrence of likelihood (log likelihood ratio), assessment (χ^2 Chi-square test). At the same time, without special modifications, none of these methods makes it possible to simultaneously take into account the level of positive and negative attitudes of one subject to key issues, take into account the mutual dependence of key issues, up to semantic synonymy.

The application of the model for determining reputation in social networks based on the use of PSN can provide: the possibility of training the system, taking into account some semantic synonymy, homonymy in key issues at the level of expanding the multiplication tables of the corresponding PSN; the possibility of applying the existing developments in the field of hypercomplex numerical systems, including isomorphic CNS, but more suitable for calculations.

4.3. Rating of Internet resources

The concept of information survivability is closely related to the problem of reputation management on the Internet. In turn, to manage the survivability of information objects, it is necessary to model their life cycle: formation and development, reactions to destructive influences, restoration, destruction.

Under the survivability of an information system is understood the ability of it (or its fragment) to adapt to new unforeseen conditions, resist undesirable influences while implementing the main function – targeted information. In addition, today such a

socially important problem as ensuring information security is associated with the survivability of information objects.

There are several mechanisms that ensure the persistence of information objects on the Internet.

Some of the most common survivability mechanisms are discussed below, which in reality are not used in their pure form, but, as a rule, in a combined one.

To study the problems associated with survivability, it is necessary to clearly define both this concept itself and to provide a formal model on the basis of which it is possible to calculate the level of survivability for such entities that are difficult to formalize as information objects.

5.3.1. Mechanisms for ensuring the survivability of information objects

The concept of survivability of the information component of the Internet implies the ability of information objects (news reports, articles, documents, videos, etc.) to perform their functions (informing) in a timely manner under the influence of destabilizing factors. Such factors may be the elimination of individual objects from the information space, the loss of their properties of relevance, accessibility [Dodonov, 2011], [Knight, 2003]. Let's consider some of them.

1. Copying data when placing it on the target resource. That is, the author places information that is copied by the hosting provider to a number of mirror servers. An example is the infamous WikiLeaks service (several hundred servers that store copy fragments).

2. Reprinting information (republications, "copy-paste") on other sites for the purpose of their content. As an example, the ratio of original information and the total amount of information scanned by the InfoStream system [Grigoriev, 2007] for the first four months is given 2012 by day (Fig. 87). At the same time, it should be noted that the most important and interesting information is reprinted hundreds of times, while irrelevant, uninteresting information is practically not duplicated.

3. Information once posted is permanently included in Internet archival services such as the Internet Archive (archive.org), which accumulates network information. The Library of Con-

gress (www.loc.gov) has bought the rights to store all public messages from the social network Twitter since 2006 and all tweets that will be published henceforth. The Library of Congress also runs the Digital Preservation National Digital Preservation and Distribution Project (www.digitalpreservation.gov - 1,400 data collections).

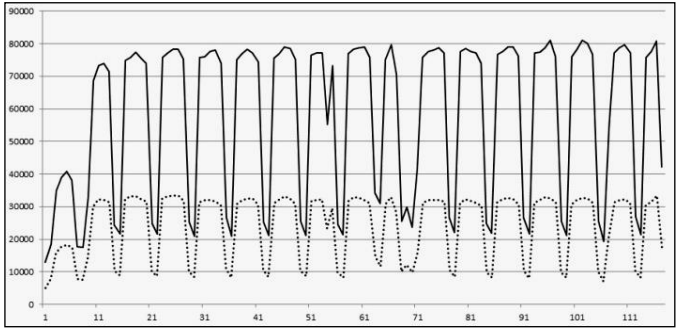


Figure 87 - The ratio of the original information (dashed line) and the total amount of information (solid line) scanned by the InfoStream system

4. Information often remains in search engine caches even if it is removed from a web page or social network page. Information is indexed by global information retrieval systems and remains in their cache memory, from where it is available to users. Only relatively recently, web resource administrators have the opportunity to independently remove their content from Google and Yandex caches. Often a lot, for example, about a person can be found in his blog, online reputation is a fashion brand today. As for the social network Twitter, [twitFlink](http://www.twitfink.com) (www.twitfink.com), for example, will quickly collect and issue patient tweets. The Google Replay service allows you to find and view thematic messages in microblogs for a specified period of time.

5. Finally, information from the website may be stored on the local computers of end users who have accessed it either directly or through information integrators.

5.3.2. Formal survivability models

information objects

It is known that the survivability of an information object can be assessed as the probability that the object will be intact during a certain period of time t under certain conditions [Li, 2012].

If an information object is stored on n servers (information carriers), then the probability of destruction of this object is estimated as:

$$F_{lost}(t) = \prod_{i=1}^n F_i(t).$$

In this product $F_i(t)$, is the probability of losing an information object on i the -th server in time t .

Accordingly, survivability is estimated as:

$$S_n(t) = 1 - F_{lost}(t) = 1 - \prod_{i=1}^n F_i(t).$$

Assuming that the probability of destruction of information objects is proportional to the time of their existence, and that the time of their destruction has a power-law distribution (in accordance with the Pareto law), it can be considered appropriate and justified to study a model with a power-law distribution of losses of information objects, which is fundamentally different from approaches which use the Poisson error flow (the theory of queuing systems) and the Weibull distribution of errors. In this case, survivability can be estimated as:

$$S_n(t) = 1 - \prod_{i=1}^n F_i(t) = 1 - \prod_{i=1}^n C t^{-\beta} = 1 - C^n t^{-n\beta},$$

where C , β are some constants.

The regularities of the statistical distribution of the lifetime of information objects allow us to draw conclusions related to their survivability, namely, to take into account the phenomena of self-similarity, the irregularity of information loss, the presence of a “heavy tail” in the distribution, which characterizes an extremely large number of actually obsolete information objects, etc.

When analyzing the life cycle of an information object, two more large classes of models can be used: Boolean and Markov.

In the Boolean model, we can assume that copies of the information object are stored on n servers, while *the* i -th server corresponds to the Boolean variable x_i , which can take the values $\{0,1\}$, i.e. $x_i = 1$ if the data object on server i is active, and 0 otherwise. The state of an information object is determined by the structure function of its availability (boolean function) $S(x_1, x_2, \dots, x_n)$, which takes the values 1 if the information object is available, and 0 otherwise.

If the availability of an information object is considered as a function of time, then the state of the information object on *the* i -th server can be considered as a random process $x_i(t)$ that takes the values 0 and 1 at arbitrary times. $t \geq 0$ The probability of its failure-free operation is determined for the system using the above formulas.

Among the shortcomings of Boolean models, one can name the assumption of only two states of an information object – activity (availability) and inactivity. In addition, in the general case, the nature of failures of individual copies of an information object depends on the state of other copies.

An information object can also be described by a Markov model. Let the system (the set of copies of the information object) have m possible states. Denote the set of states by $M = \{z_1, z_2, \dots, z_m\}$. For any fixed point in time, $t \geq 0$ the state of the system $z(t)$ is interpreted as a random variable. The set of all states M , the initial probability distribution vector $p(0)$, and the transition probability function are given. The probability of activity, “life” of the system at a given time t (system readiness) is determined. The applicability of Markov models also has its limits. The rates of transitions between individual states of the system can be non-stationary, the assumptions made in the calculation regarding the distribution of the failure rate can significantly reduce the accuracy of the results obtained; the number of states of the system can be so large that the calculation becomes practically impossible.

The assessment of the survivability of information objects can be carried out at all stages of their life cycle. There are sev-

eral approaches to the assessment of survivability, which are of the most general nature. Survivability can be evaluated relative to some standard external influence or relative to a variety of external influences. In this case, the problem of finding a set of characteristic vectors of states of an information object (in the simplest case, distribution over servers) is solved, in which the configuration is implemented that ensures the fulfillment of the purpose of functioning. The power of this set can serve as a measure of the survivability of the entire information object.

When analyzing the survivability of information objects, the problem of informing on their various aspects is considered, regardless of the presence or absence of unfavorable factors. In this regard, as a quantitative criterion for assessing survivability, it is advisable to use the ratio of the number of functions performed by an object in the presence of certain adverse effects or a multitude of such effects to the total number of functions of an information object, taking into account the criticality of the functions performed and not performed. The criticality of each specific function is determined individually for each specific information object based on its specifics. The quantitative indicator of the survivability of a particular information object under given conditions can be calculated by the formula:

$$S = \sum_{i \in \Delta} \alpha_i / \sum_{j \in \Theta} \alpha_j ,$$

where Θ is the set of all informing functions, Δ is the set of information object functions performed under given conditions ($\Delta \subseteq \Theta$), α_n and is the criticality of the n th function. Thus, the quantitative assessment of the survivability of an information object will change in the interval $[0, 1]$, the higher the survivability, the greater its quantitative assessment.

5.3.3. Digital footprints and shadows

Removing an information object from a web resource cannot guarantee its disappearance from the Internet. Not only "digital footprints" and "digital shadows" remain.

The expression "digital footprints" (Digital Footprint) refers to the information that is left by the user himself when working on the Web and by which you can not only identify him, but also "attach" him to certain actions, events, restore some fragments of his biography.

Often, users voluntarily indicate their full names, "linking" further information to their own personality, date of birth, marital status, education, profession, places of previous work, and much more, including contact numbers and email addresses. In addition to the "digital footprints" that users leave themselves, information about users is constantly replicated and without any participation.

Information about the user, created without his participation, is called "digital shadow" (Digital Shadow), which arise and accumulate whenever someone searches for a user through search engines, when an e-mail is sent to the lists in which he appears in many other cases. Indexing by search engine robots of pages with user information and their subsequent caching is also the creation of a "digital shadow" available to everyone. In addition to "public access digital shadows", "restricted access digital shadows" are created and accumulated – surveillance camera recordings, bank transactions, billing of online stores, ticket sales services, phone calls, etc.

According to the analytical company International Data Corporation (IDC), specializing in information technology market research, the volume of the "digital shadow", i.e. information about the Internet user, which is created without his participation, is already in 2007 exceeded the amount of information that the user creates himself.

More and more users face the problem of online reputation every day. This is also evidenced by the appearance of special sites (for example, www.suicidemachine.org), which allow you to simultaneously delete the registration and all the entries made on various forums and social networks. Such an operation is called "commit suicide on the Internet." However, this system is still imperfect. Recently, this concern has been taken over by special companies, the so-called "Internet cleaners", who establish contacts with the administration of leading search engines and social networks, individual websites, use programming interfaces for interacting with search engine caches.

As an illustration, we can cite the data of the administration of the social network (microblogging service) Twitter on the number of requests to remove content. According to analysts, in the first half of 2013, governments around the world sent 1,157 requests for information to Twitter. For the same period 2012 this figure was 849. At the same time, the number of requests to remove content increased by 10 times. Russia leads in terms of the number of requests to delete information. In addition, there has been a sharp increase in government requests. 78% of all requests for information (902) are from the USA. In second place and third place are Japan (87) and the UK (26).

The concept of survivability of an information object implies its ability to perform its functions (in this case, informing) in a timely manner under the influence of destabilizing factors. Such factors may be the elimination of individual information objects from the information space, the loss of their relevance, accessibility. It should be noted that drawing the attention of the audience to another topic, generating another information object can also reduce the relevance of the current information object.

At the same time, it should be borne in mind that the most important information, once on the Internet, remains there almost forever, and as practice shows, it is not necessary to count on its easy removal or change. The best method is to replace undesirable information with new plots, to conduct special events to correct errors in a meaningful way [Dodonov, 2010].

Considering the effect of the super-survivability of information on the Internet, it is worth considering several important points when dealing with negative content when managing reputation on the network:

- you can't just ignore negative content; As you know, information messages, especially those of a negative nature, are repeatedly duplicated in the network. Therefore, rebuttals, positive content are needed;
- Internet cleaners – services for removing negativity from the Internet can "mechanically" only partially solve the problem. Negative information vseravno somewhere will remain and once pop up. Therefore, negative content should be replaced with positive content;

- positive content should be truthful, objective. The Internet is a great lie detector;
- it is necessary to place "negative pushing" positive information on the network on various target resources, taking care of hyperlinks to this information.

The survivability of information objects and systems is difficult to notice under normal operating conditions. This property is manifested in relief only in cases of loss of information, violations in the structure of the information system, failure of its components, individual functions, purposeful destructive influences. Depending on the class of systems, their complexity, degree of organization, as well as on the chosen level of analysis, the property of survivability can be evaluated as stability, reliability, adaptability, fault tolerance.

The currently observed process in the field of intellectualization of automated systems, the transition from simple data processing to decision support processes requires new approaches. That is why a special place is occupied by tasks related to ensuring the survivability of both information systems and information objects in a network environment.

5. Legal Issues of Competitive Intelligence

5.1. Competitive intelligence in the legal field

Of course, competitive intelligence as a field of activity should be carried out within the legal framework of the state. The basis for this is the constitutional rights to search, receive, transmit and use information in all civilized states. At the same time, it should be noted that in a number of countries, legislation that restricts the collection and processing of information practically prohibits competitive intelligence.

At the same time, in Ukraine "everyone has the right to freely collect, store, use and disseminate information orally, in writing or in any other way – at his choice" (Constitution of Ukraine, section 2, art. 34).

Thus, in Ukraine, legal regulation in the information sphere, which, of course, includes competitive intelligence, is based on the following principles:

- 1) freedom to search, receive, transfer, produce and distribute information in any legal way;
- 2) establishment of restrictions on access to information only by the laws of the state;
- 3) openness of information about the activities of state bodies and local self-government bodies and free access to such information, except in cases established by the laws of the state;
- 4) according to the category of access, information is divided into open (publicly available) and with limited access. In turn, information with limited access by its legal nature is also divided into two types: information constituting a state secret; confidential information.

Despite the fact that competitive intelligence is today a recognized field of activity, there is no legalized concept of "competitive intelligence" in Ukraine today, although the collection, storage, processing and dissemination of information is regulated by a number of laws and regulations:

Law of Ukraine "On Information" dated 02.10.1992. No. 2657-XII (as amended on 13.01.2011), art. 5-7.

The Law of Ukraine "On other matters of mass information (press) in Ukraine" dated November 16, 1992, No. 2782-XII, Art. 6, 25.

Law of Ukraine "On the protection of activities" No. 4616-VI dated March 22, 2012 Art. 9, 13, 19.

Law of Ukraine "On protection of personal data" No. 2297-VI dated 01.06.2010

Civil Code of Ukraine (Art. 505), Criminal Code of Ukraine (Art. 231, 232), Code of Ukraine on Administrative Offenses (Art. 163, Art. 163);

Decree of the President of Ukraine "Nutrition for European and Euro-Atlantic Integration" dated April 20, 2019 No. 155/2019.

Decree of the President of Ukraine "On the National Coordination Center for Cyber Security" dated 07.06.2016 No. 242/2016.

We must not forget that the implementation of measures to ensure business security, even within the framework of competitive intelligence, can sometimes be perceived as conducting operational-search activities, which, according to the Law of Ukraine "On operational-search activities" dated February 18, 1992 No. 2135-XII, can be carried out by only the entities specified in separate articles of these Laws. At the same time, the list of subjects is exhaustive, and it is prohibited to conduct operational-search activities by other legal entities and individuals.

The Cybersecurity Strategy of Ukraine, approved by Decree of the President of Ukraine No. 96/2016 dated January 27, 2016, declares the main tasks for law enforcement agencies, including: "for the intelligence agencies of Ukraine – the implementation of intelligence activities to identify threats to the national security of Ukraine in cyberspace, other events and circumstances related spheres of cybersecurity", and also provides for "the creation of a system for the timely detection, counteraction and neutralization of cyber threats, including with the involvement of volunteer organizations", all this, of course, refers to the use of OSINT (or competitive intelligence) tools in this area.

At the same time, the current Criminal Code of Ukraine provides for criminal liability for illegal collection for the purpose of use or use of information constituting a trade secret, as well as for disclosure of trade secrets. Obviously, such information goes beyond competitive intelligence.

With a fairly broad interpretation of the legislation, any procedures for collecting, processing and storing information about competitors become, on the one hand, legitimate, practically unpunished, and, on the other hand, difficult. In Ukraine, access to a large layer of business information freely available in most countries, for example, about real estate (existing and mortgaged), land plots, bank accounts, etc. is actually closed. In these countries, much of the information can only be obtained through consultation with relevant experts.

Today, more than ever, the problem of criminalization of certain competitive intelligence services is acute. Many security services today use databases with information about persons. Such databases are used for quite good purposes, for example, to check data on employees, partners and competitors. Obviously, they will continue to use such databases in the future, but they will be forced to break the law and "go underground." Technically, the use and maintenance of such databases are provided by numerous systems such as Cronos (shells distributed quite legally). With the help of such tools, numerous databases that work with these shells become available to any interested Internet user.

As a result, the activities of companies engaged in competitive intelligence, there is increased attention from state regulatory authorities.

This is due to several groups of legal problems, which can be grouped by highlighting the problems associated with:

- 1) protection of trade secrets;
- 2) protection of personal data;
- 3) observance of copyrights;
- 4) the possibility of competition in the market of the most competitive intelligence.

It is also possible to distinguish three classes of main copyright problems related to competitive intelligence, these are problems related to such aspects:

- the legitimacy of using input information (sources of information), on the basis of which reports are generated – the results of competitive intelligence; Problems,
- copyright on the results of competitive intelligence;
- the rights to use (use) specialized software required for competitive intelligence.

In addition, one of the problems facing competitive intelligence services in Ukraine is the almost complete absence of anti-dumping legislation. Despite the fact that the entry of large international players into this market is difficult due to the lack of necessary connections, databases, archives, and even linguistic and legal training, they may exhibit dumping on competitive intelligence services.

The situation may change if a clear legal framework is created for activities related to the collection and analytical processing of information and, in particular, for competitive intelligence.

5.2. Competitive Intelligence and Trade Secret Protection

Important for the development of competitive intelligence was a number of articles of the Law of Ukraine " On Protection from Unfair Competition" No. 236/96-BP dated 06/07/1996, where (Article 15-1), "Illegal collection of commercial information", "Disclosure of commercial information » " Misuse of commercial information" (Chapter 4, Art. 16, 17, 19, respectively).

Decree of the Cabinet of Ministers of Ukraine dated August 9, 1993 No. 611 "On the list of information that does not constitute a commercial secret" defines a whole class of documents related to the activities of business structures that are actually open, in particular, constituent documents, reporting forms, information on the participation of founders and officials in other companies, etc.

Competitive intelligence efforts are often focused on obtaining the trade secrets of competitors. And although different formulations are given in various legislative acts, one can agree with the fact [I Vashchenko, 2006] that a trade secret is characterized by such a set of features: the information is secret, is unknown and is not easily accessible to persons who usually deal with the type of information, to which it refers; because it is secret, it has commercial value. Thus, a trade secret is information that is useful and is not generally known to the public. It has a real or commercial value from which profit can be made and for the protection of which the owner takes measures in all spheres of life and activity. Thus, it can be said that competitive intelligence

activities are sometimes aimed at extracting information that is not publicly available and is protected by law. These acts violate a huge number of articles of the Criminal Code of Ukraine, in particular, Article 231 "Illegal collection for the purpose of use or use of information constituting a commercial or banking secret."

Thus, commercial intelligence can legitimately use only those methods and means of collecting and processing information that do not contradict the law, i.e. the main functions of competitive intelligence are high-quality collection, systematization and, most importantly, analysis of information, and not surveillance, bribery and illegal hacking.

For the first time the right to keep commercial secrets was proclaimed by the USSR Law of June 4 1990 "On Enterprises in the USSR". In Art. 33 of this Law, the concept of a trade secret was disclosed as information that is not state secrets related to production, technological information, management, finance and other activities of enterprises, the disclosure (transfer, leakage) of which may harm their interests.

Currently, Ukrainian legislation on the protection of official and commercial secrets is a collection of articles contained in various legal acts generally devoted to the regulation of other public relations.

The Civil Code of Ukraine, in turn, defines a trade secret (p. 505, paragraph 1) as information "which is secret in the sense that it is generally or in a certain form and aggregate is unknown and not easily accessible to persons who usually deal with the type of information to which it relates, therefore has a commercial value and has been the subject of measures adequate to the existing circumstances relating to the preservation of its secrecy, undertaken by the person who legally controls this information.

In accordance with these definitions, as soon as information, as a result of any actions, gets, for example, on the pages of any website, it ceases to be considered a trade secret, as it becomes easily accessible.

Although many articles of the Criminal Code of Ukraine (Articles 231, 232, 232-1, 361, 363) establish criminal liability both for the disclosure of trade secrets and for the illegal collection and use of information related to it, however, the existing legal base does not clearly regulate what kind of information about the financial and economic activities of an enterprise is a com-

mercial secret (with the exception of banking secrecy, the definition of which is given in Article 60 of the Law of Ukraine “On Banks and Banking Activity”).

5.3. Competitive intelligence and protection of personal data

Government agencies, banks, large corporations are not always able to ensure the protection of their personal data bases, as a result of which a huge flow of confidential information enters the market. Ensuring the security of personal data is an objective need. Today, personal data, information about people is becoming the most expensive commodity. Such information in the hands of an attacker is a powerful weapon. That is, personal data must be protected.

Personal data is an important component of a broader concept – privacy. Therefore, the protection of personal data is an integral part of ensuring privacy. Privacy, along with freedom of speech and other rights, is one of the core values of humanity.

Today, the main European documents in the field of personal data protection are the Council of Europe Convention "On the Protection of Persons in Connection with Automatic Processing of Personal Data" and the Directive of the European Parliament "On the Protection of Individuals with Automatic Processing of Personal Data", ETS No. 108, which 1981 is obligatory for all member states of the European Union and which is a subject for imitation in the field of legislation, including our country. The EU countries are consistently bringing their legislation in line with the Directive. In the UK back in 1998, the Data Protection Act 1998 was adopted. His technical implementation – project Standard "Specification for the management of personal information in compliance with the Data Protection Act 1998" (BS 10012, 2009). In parallel with the British, their version of the personal data security standard was released in the United States. The US Government Draft Guide to Protecting the Confidentiality of Personally Identifiable Information (PII) (SP 800122) governs the implementation of The Privacy Act of 1974 and the Privacy Protection Act of 1980. Canada has released the "Privacy Code" – a set of documents for the implementation of legislation on the protection of information about individuals (The Privacy Act and PIPEDA).

In the Member States of the European Union, the definitions of personal data, as a rule, are as broad as possible, as a result of which citizens in practice often do not comply with the relevant legislation due to excessive “burden”. The relevant public authorities generally do not take any action except in special cases. Important questions remain about the emergence of conflicts between the requirements of privacy and the interests of freedom of speech. Modern European laws, as a rule, prohibit the collection, storage, use and distribution of critical personal data without the consent of the data subject.

The right to privacy is guaranteed by the Constitution of Ukraine. Article 32 of the Constitution of Ukraine states: "No one may be subjected to interference in his personal and family life, except in cases provided for by the Constitution of Ukraine." In addition, the Constitution of Ukraine provides for the protection of some more aspects of privacy. Thus, Article 30 protects the inviolability of the home (territorial privacy), Article 31 protects the secrecy of correspondence, telephone conversations, telegraph and other correspondence (communication privacy), Article 32 provides for a ban on the collection, storage, use and dissemination of confidential information about a person without his consent (information privacy), and Article 28 provides for the prohibition of subjecting a person without his free consent to medical, scientific or other research (protecting certain elements of physical privacy).

The Convention for the Protection of Persons with regard to Automatic Processing of Personal Data Strasbourg, 28 January 1981 (ratified on 06/07/2010) defines the provision regarding the transmission across national borders by any means of personal data subjected to automated processing or collected for the purpose of automated processing.

The following data is often used to isolate a specific person listed as a personal person by the US Office of Management and Budget:

- full name (meaning the name together with the surname)
- national identification number;
- IP address (in some cases);
- vehicle license plate;
- driver's license number;

- face, fingerprints, or handwriting;
- credit card numbers;
- digital identity (digital signature);
- Date of Birth;
- Place of Birth;
- genetic information.

According to the legislation of most European countries, personal data is divided according to the criterion of “sensitivity” into general data and “sensitive” (vulnerable) personal data.

General personal data:

- identification data (last name, first name, patronymic, address, telephone number, etc.);
- passport data;
- personal information (age, gender, marital status, etc.);
- family composition;
- education;
- profession;
- living conditions;
- Lifestyle;
- vital interests and hobbies;
- consumer habits;
- financial information.

"Sensitive" personal data:

- information about racial, ethnic origin and nationality;
- information relating to political, ideological and religious beliefs;
- information about membership in political parties, trade unions, religious or public organizations;
- health and sexual information;
- genetic and biometric data;
- location and ways of movement of the person;
- information on the application of measures to the person within the framework of the labor investigation;
- information about the commission of various types of violence against a person.

The constitutional norms define an exhaustive list of grounds for interference with privacy and the conditions for such interference. However, in the post-Soviet states there are many sectoral

norms of law that contradict the requirements of their Constitutions. It is these norms that do not comply with international standards, the practice of European legislation.

In accordance with Ukrainian legislation, personal data in Ukraine is full name, accompanied by any other identifying information, such as address, phone number, or educational status.

To find out what is the relation of an individual or a company to the protection of personal data, it is of great importance to determine the subjects of relations related to personal data (Article 4 of the Law of Ukraine No. 2297-VI): "The subjects of relations related to personal data are:

- subject of personal data;
- owner of the personal data base;
- manager of the personal data base;
- third party;
- authorized state body on personal data protection;
- other public authorities and local governments, whose powers include the protection of personal data.

Ukrainian legislation provides for the notification nature of the processing of personal data. The owner or manager (operator), prior to the processing of personal data, is obliged to notify the authorized body for the protection of the rights of personal data subjects of his intention to process personal data. Then the data about the owners or managers (operators) are entered into a special register of operators. The information contained in the register of operators becomes publicly available.

Personal data laws concern the majority of the population as participants in the process of "processing" data. And since every person is the subject of personal data, the Law is universal and applies to everyone.

This legislative act is directly related to the field of information technology and telecommunications, both contain controversial, contrary to established practice, seemingly unenforceable provisions. The requirements of the law apply to all legal entities and individuals, and the Internet sphere is no exception. The law on the protection of personal data may change the principles of op-

eration of Ukrainian Internet resources: e-mail services, dating, online stores and social networks, although market participants themselves hope that the sites will not be subject to the law. In order to comply with all provisions of the law on personal data, owners of Internet resources need to carefully consider the organization of their activities. Currently, there are many web services within which personal data is collected, stored, and used. Compliance with the requirements of the law is not an easy task for the owners of these resources, in particular, officials have the ability to oblige Internet companies to take written consent to the use of personal data from each user. It is no secret that many sites host information containing personal data of people (for example, job seekers, acquaintances), including those related to special categories, such as nationality or religion. The task of those who provide such services is to legitimately process such information and at the same time protect it in accordance with the requirements of the law.

In particular, personal data is widely used in social networks and email services. For example, it is very difficult for owners of web resources to comply with the legal requirement to obtain the consent of each user for the processing of his personal data. At the same time, the law imposes on the operator the obligation to prove the fact that he received such consent.

A modern Internet company collects and processes different categories of personal data – its employees, its counterparties under contracts, and some data of users of its services. People who post information about themselves on social networks or dating services deliberately make it open to all users of the resource, and by law it can be interpreted as “publicly available”, which means that it is not required to comply with a special confidentiality regime in relation to it, but in social networks there is also information that the user hides, making it available only to a separate group of users (“friends”). In this case, the Internet resource must provide for it special means of protection.

In the practice of competitive intelligence, one has to deal with numerous contradictions and incidents in the existing legislation, for example, the Ukrainian Law “On the Protection of Personal Data” (part 9 of article 6) states: “the use of personal data for historical, statistical or scientific purposes can only be car-

ried out for disembodied form." That is, entries in competitive intelligence reports should look something like this: "Person A negotiated with person B." References to other colleagues should not be made in scientific reports, even with their written consent. Causes certain difficulties and the need to notify the authority "on each change in the information necessary for the registration of the relevant database", which, among other things, includes information about all managers (users) of such a database.

In addition, many competitive intelligence services, quite legally creating a database of personal data to solve their designated task, are obliged to destroy the fruits of their work, having achieved the goal. But if the main goal, for example, when providing services to customers, is the fulfillment of these requests themselves, but the accompanying goal of any self-respecting organization is the development of a customer base. And this base often has its own commercial value. Numerous cases of legal resale of customer databases are known, for example, when the activity of the owner company is terminated. There is no strict article in the Ukrainian legislation, however, the conditions for the destruction of personal data are provided, among which (Article 15), "termination of legal relations between the subject of personal data and the owner or manager of the base...". And this means, for example, that the operator – the provider of the service must destroy the entire database accumulated during the performance of the service.

Therefore, the owners and managers of such databases reformulate their goals in a special way, for example, as "providing a service with the possibility of storing personal data during the warranty period...". Thus, the norms of the law are observed and the interests of the performer – the owner or manager (operator) of the personal data base are ensured.

Competitive intelligence units are engaged in the processing of personal data that are in open sources on the Internet, i.e. are public. For their processing, the consent of the subject of personal data is not required. However, at the same time, the obligation to prove that the personal data being processed is publicly available rests with the owner or administrator. And this means that it is necessary either to accumulate evidence that the data is taken from publicly available sources, or to obtain consent from the subject of personal data and then store this document. In

addition, you must have a document confirming the public availability of the source of personal data. At the same time, the question of proving that the owner of the information resource (website) has a written consent to processing remains unanswered.

The problem of criminalization of certain competitive intelligence services has become more acute than ever. Many security services today use databases with information about persons. Such databases are used for quite good purposes, for example, to check data on employees, partners and competitors. Obviously, they will continue to use such databases in the future, but they will be forced to break the law and "go underground." Technically, the use and maintenance of such databases are provided by numerous systems such as Cronos (shells distributed quite legally). With the help of such tools, numerous databases that work under these shells become available to any interested Internet user.

At the state level in the United States, the Department of Defense's primary legal mechanism for conducting open source intelligence is the Open Source Security Council (DOSC). It serves as a forum for coordinating and facilitating open source intelligence activities and programs for all services and combat teams. This council advises and reports to the Under Secretary of Defense for Intelligence on open source intelligence issues, new initiatives to improve the performance of the OSINT unit and the activities of the Department of Defense as a whole. The responsibilities of the Council include:

- coordinates the activities of the OSINT unit and approves its open source intelligence plan;
- determines the sequence of requirements for the process of conducting intelligence in open sources.

US Army Standard "ATP 2-22.9" establishes the general concepts, basic concepts and methods of collecting intelligence from open sources for the US Army. This paper highlights the characteristics of OSINT as an intelligence discipline, its links to other intelligence disciplines, and its application to joint operations.

The use of publicly available information is an important aspect of technical intelligence (TECHINT). Despite the fact that the intentions, capabilities and vulnerabilities of adversaries and

potential threats are subject to classification, the results of OSINT (in particular, the open Google Earth service) contribute to obtaining information about the most secretive states and organizations. Such examples testify to the responsibility of activities in this area.

Copyright is a form of protection for published and unpublished works under Title 17 of the US Code, which defines the authors of "authors' original works", including literary, dramatic, musical, and artistic works.

National copyright laws are restrictions on competitive intelligence. Infringement of rights, in particular, under chapter 17 of the US Code, copyright laws, still leaves the possibility of the legitimate use of competitive intelligence, which is determined by four factors:

- the purpose and nature of the use;
- properties of the author's works used;
- the number and parts of the author's work that are used;
- the impact of the use of copyrighted works on the potential market or value of these works.

5.4. Competitive intelligence and copyright protection

It is possible to single out three classes of the main copyright problems related to competitive intelligence, these are problems related to such aspects:

- the legitimacy of using input information (sources of information), on the basis of which reports are generated – the results of competitive intelligence;
- problems with copyrights on the results of competitive intelligence;
- the right to use (use) specialized software required for competitive intelligence.

In addition, one of the problems facing competitive intelligence services in Ukraine is the almost complete absence of anti-dumping legislation:

- the situation may change if a clear legal framework for competitive intelligence is created;
- Copyright is a form of protection for published and unpublished works under Title 17 of the US Code, which defines authors as "original works by authors," including literary, dramatic, musical, and artistic works. National copyright laws are restrictions on competitive intelligence. Despite this, there are still opportunities for the legitimate use of competitive intelligence, determined by four factors:
 - the purpose and nature of the use;
 - properties of author's works;
 - the number and parts of the author's work;
 - the impact of the use of copyrighted works on the potential market or value of these works.

6. Opposition to information operations

the term information operations has become popular, primarily because information technology plays an ever-increasing role in military operations. At the same time, information operations are defined as "actions aimed at influencing the information and information systems of the enemy, and protecting one's own information and information systems" [DoD, 2003]. Information operations are seen as combining the basic capabilities of electronic warfare, computer network operations, psychological operations, military operations and security operations in order to influence, destroy, distort information necessary for the enemy to make decisions, as well as protect one's own information.

Information operations cover a whole range of processes carried out in a variety of areas. At the same time, it should be noted that information operations are an essential and traditional component of combat operations. Although the formal definition in US Department of Defense documents is focused on the military aspects of information operations, it is quite applicable to almost any area of life.

Below we will consider such information operations that are implemented using information systems (IS). The survivability of these IS largely determines the survivability of information operations, which are implemented in the form of information impacts on people's consciousness.

Information is a reflection of the meaning invested in it, therefore today information has turned from an abstract term into an object, purpose and means of information operations, has become a critical concept in security issues. Former US Secretary of Defense William Cohen March 18 1999 stated that "the ability of the army to use information to dominate future battles will give the US a new key to victory for many years, if not for several generations" [Hill, 2000].

When modeling and conducting information operations, it is necessary to take into account the value of information for decision makers. The value of information includes its timeliness, accuracy and "analyticity". From a practical point of view, the value of information can also be defined as its relevance or applicability, suitability for use. The applicability of information is

understood as providing access for decision makers to ready-to-use information. The ISO 9241 standard (ISO stands for International Standards Organization) defines applicability in terms of efficiency and satisfaction of the needs of a specified set of users to solve a specified set of tasks in a specific environment. In practice, most of the useful information comes to decision makers from information and analytical systems that provide orientation in the situation and support in making decisions. According to the US War Department's Information Operations Field Manual (FM 100-6), "situational orientation means a combination of a clear understanding of the disposition of friendly and enemy forces with an assessment of the situation and intentions on the part of the command."

Information operations are carried out in a certain social environment, therefore, for their successful implementation, it is necessary to adapt to this environment, to overcome a certain barrier of not very strong attention to information impact. This barrier arises due to the so-called immune system of the environment, which may not miss information impacts if it is powerful enough and/or has already learned to defend itself from such impacts. Preparatory actions for conducting information operations may include the creation of an "immunodeficiency" of the social environment by influencing through the information space, for example, with the help of materials in the media. Very often, information influences use the mechanisms of "viral marketing", for example, in the form of rumors, when sensationally presented disinformation spreads at a great speed. It is the immune system that counteracts such information and operations. Very often, societies identify with the immune system the state, which is called upon to ensure the security of this society, i.e. in the presence of a strong state apparatus, the probability of success of antisocial information operations is significantly reduced. The reader knows perfectly well how such informational processes were counteracted in totalitarian states. In a democratic society, of course, totalitarian methods are not applicable. In this case, immunity is achieved through "learning", i.e. a democratic society must go through many informational attacks, influences, influences of stereotypes in order to develop the necessary immunity.

The level of readiness for information operations today is considered a key success factor for any social procedure, campaign.

A special purpose in carrying out information operations is the information and analytical systems of the subject of influence. By influencing such systems, it is possible to ensure that decision makers from the enemy camp make inadequate conclusions, and the required social process will change the trajectory in the direction necessary for the influencing side [Gorbulin, 2009] (Fig. 88).

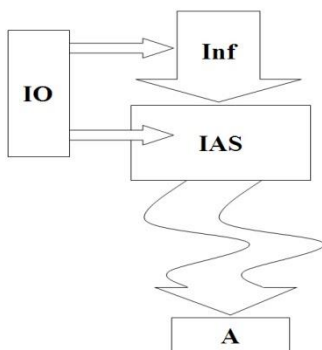


Figure 88 – Influence on the information-analytical system of the enemy: Inf – information space; IAS – information and analytical system; A – system subscriber – decision maker; IO -in information impact

In this case, direct information impacts may include the placement in the information space of documents compromising the opposite side, advertising (including hidden) of one's advantages, distorted data about the external environment, distorted information about intentions, etc.

Social procedures and processes tend to be difficult to evaluate and model because their outcomes are psychological and sociological rather than physical. It is this fact that also determines the problematic nature of predicting the results of modeling information operations. In addition, experimentation with information influences in the framework of information operations is more complex and dangerous than in the simulation of physical processes.

Actions to be effective in influencing adversary decision-making processes sometimes need to be taken for a long time before they take effect.

One of the main components of information operations is social influence, covering the whole variety of influence processes. Significant changes in people's beliefs or attitudes towards some problem or phenomenon are expected to lead to a change in behavior associated with this problem.

In 1948, Harold D. Lasswell [Lasswell, 1948] developed a communication transmission model consisting of five components:

- source – a person who influences or convinces other persons;
- message – with the help of which the source tries to convince the target;
- the target is the person whom the source is trying to influence;
- channel – message delivery method;
- impact – the reaction of the target to the message.

Although Lasswell was primarily interested in mass communication, his model of information transfer can be applied to interpersonal communication such as the Shannon-Weaver and Osgood -Schramm circular models, which include feedback loops in the communication process, stating that communication is circular, not linear [Schramm, 1974], [Osgood, 1954].

Modeling objective factors of social influence requires interdisciplinary approaches related to computer science, marketing, political science, and social psychology. The most famous models of public opinion formation and social influence are based on Latane's theory of dynamic social impact [Latane, 1981], [Latane, 1997], developed by many other authors, primarily in [Nowak, 1990], [Lewenstein, 1993], [Kacperski, 2000], [Sobkowicz, 2003].

Trying to justify the mechanism of social influence of messages, Latane [Latane, 1981] emphasized the importance of three features of the source-target relationship:

- power – social power, probability or level of influence on individuals;
- immediacy – physical or psychological distance between individuals;

- number of sources – the number of sources tending to the goal.

The current state of information operations modeling is characterized by a number of open problems, the main of which relate to understanding the concepts of information influence and impact.

6.1. Information influence, attacks and operations

The universal characteristics of objects are their state and the possibility of influencing other objects. The realization of the possibility of influence requires certain conditions, which are usually called its influence. At the same time, an object that can exercise its will is called a subject, and control is usually called an impact in relation to the object of influence, applied for a specific purpose.

When an individual is the target of influence from one or more sources, dynamic social influence theory states that the level of social influence on an individual can be represented by an equation that is the basis of the so-called person-centered model :

$$I_i = -S_i\beta - \sum_{j=1, j \neq i}^N \frac{S_j O_j O_i}{d_{i,j}^\alpha},$$

Where I_i – the magnitude (quantity) of social pressure exerted on the individual i , ($-\infty < I_i < \infty$); O_i And O_j represents the opinion of the individual i and j , respectively) on a topical issue – +1 or -1 – support or objection to this issue, respectively. S_i (S_j) represents the power of the individual i (j) or influence ($S_i > 0, S_j > 0$); β – resistance of the individual to changes ($\beta > 0$); $d_{i,j}^\alpha$ - distance between individuals i And j ($d_{i,j}^\alpha \geq 1$); α – indicator of distance reduction ($\alpha \geq 2$); N is the total number of agents (individuals that make up the community). The meaning β , of the tendency to maintain one's own opinion or resist change determines that individuals within the model may require more or less social pressure to change their opinion. Larger levels of val-

ue α correspond to the effect of increasing distance between source and target, which affects the amount of social pressure on the target.

On the basis of the terms introduced, the concept of the “information field of an object” is formulated [Kononov, 2003] and its characteristics are described. This makes it possible to define the information impact as an impact on the information field of the object. Exploring the information fields of objects and subjects of social systems, one can determine informational influences and controls. At the same time, information can be considered both as an object and as a means of influence. The use of information as a means of influence requires in the management process to prepare data, produce relevant information, and only then implement the created information in the form of impact (influence).

One of the main methods of conducting information operations is the information influence exerted for the purpose of information management. In this case, information management is understood as a control mechanism when the control action is implicit, indirect informational in nature and the control object is given a certain information picture, under the influence of which it forms its line of behavior. Thus, information management is a method of influence that encourages people to behave in an orderly manner, to perform the required actions.

In accordance with [Kononov, 2003], [Kulba, 2004] it is advisable to decompose the process of informational influence of one object on others into the following stages:

- generation by the source of influence of data, information elements and information sets;
- transfer of information by a source of influence;
- receiving information by the recipient;
- generation of a set of data, information elements and new sets of the object of influence;
- appropriate active actions of the object of influence.

Information impacts on elements of systems can be classified according to such features as sources of occurrence, duration of exposure, nature of occurrence, etc.

To select specific ways to implement information management, it is necessary to specify the tasks solved with the help of information impact, analyze the process of forming information

operations and develop criteria for their evaluation. Information management is considered as a process covering the following three interrelated areas:

- management of data exchange between the real world and the virtual world of the subject of influence;
- management of the virtual world of subjects of influence, decision-making mechanisms;
- managing the process of transforming decisions into actions by the subject of influence in the real world.

Information impact can be of two main types :

1) change in the required direction of the data that the information and analytical system of the object of influence uses when making decisions;

2) direct impact on the decision-making process of the target, for example, on decision-making procedures or individual decision makers.

The most important for information operations is the environment, the state of the objects of information impact, their mutual influence. In particular, if a certain electoral field is chosen as the objects of information operations, then it is important to take into account all electoral populations included in this field, which represent supporters (or opponents) of certain political forces. Despite the fact that some models will be considered in the future, in which the homogeneity of the environment is explicitly postulated, in the general case, in relation to information operations, the environment may consist of areas:

- dominant perception;
- hypersensitivity;
- indifference to the relevant informational influences.

6.2. Stages of information operations

Let us dwell separately on the stages of information operations. Obviously, there is no single "standard" plan for conducting both offensive and defensive information operations. One can only consider an approximate sequence of actions obtained by generalizing some already implemented information operations during their implementation.

In practice, an information operation as a process of information impact on the mass consciousness, as a rule, is implemented as follows: as a result of preliminary intelligence, a plan is developed for the next stage – operational control and appropriate operational intelligence measures are outlined, which are an approximate solution model, after which operational control of the enemy is implemented. At the stage of operational intelligence, the level of deviation of the original model from reality is determined, and if it is insignificant, then the original plan is implemented. Otherwise, a new plan of operational command and control of the enemy is being built. The cycle is then repeated until operational intelligence confirms the model used. In this case, the final decision is made with a certain operational risk.

Thus, the process of information impact covers the following main stages [Chkhartishvili, 2004] (Fig. 89):

- preliminary intelligence (reliminary intelligence, PI);
- identification of the current situation, the state of the enemy (Op);
- management of the enemy (M) (information impact on the enemy in order to transfer to him information corresponding to the intention of the manager);
- operational intelligence (OI) (checking the results of reflexive control);
- operational management (OM) – the actions of the manager to achieve the desired goal.

When planning or modeling social processes, in particular information operations, it must always be taken into account that the general behavior of social systems cannot be determined using exclusively refined mathematical models. This is mainly due to the fact that such processes are largely dependent on socio-psychological factors.

There are two main types of information operations – offensive and defensive. However, in practice, most information operations are mixed. In addition, most information operations procedures are both offensive and defensive. Each of the types of information operations, including the above main stages, implies some features and refinements.

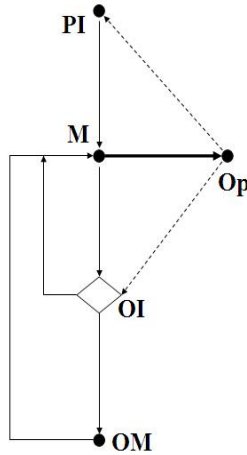


Figure 89 – Main stages of information operations

A feature of offensive information operations (information attacks) is that the objects of influence of such operations are determined and planning is based on sufficiently accurate information about these objects. An information attack most often requires finding or creating an informational occasion (for defensive informational operations, the reason may be the enemy's information attack itself), the promotion of this occasion, i.e. propaganda (as opposed to counter-propaganda measures in defensive information operations), as well as the need to take measures to prevent information counteraction.

Thus, the plan of a typical information operation includes such stages as evaluation, planning, execution, and the final phase, which coincide at the top level for both types of information operations. Let's give a more detailed list of components of information operations.

In offensive information operations, the following main phases can be distinguished:

1. Assessment of the need for an operation:
 - 1) definition of the goal, forecast of achievability, degree of influence;
 - 2) collection of information.
2. Planning.

3. Execution of information impact:
 - 1) finding or creating an information occasion;
 - 2) promotion and informational occasion (propaganda);
 - 3) operational intelligence;
 - 4) impact assessment;
 - 5) an obstacle to information counteraction;
 - 6) correction of information impact.
4. Final phase:
 - 1) performance analysis;
 - 2) the use of positive results of information impact;
 - 3) counteraction to negative results.

Typical defensive information information covers the following main stages:

1. Rating:
 - 1) analysis of possible vulnerabilities (goals);
 - 2) collection of information on possible transactions;
 - 3) identification of possible "customers" of information impacts:
 - determination of areas of common interest of the object and potential "customers";
 - ranking potential customers according to their interests.
2. Planning:
 - 1) strategic planning of a defensive operation (explicit or implicit):
 - definition of criteria for information impacts;
 - modeling of informational influences taking into account: object connections; impact dynamics; "special" (critical) points of influence;
 - predicting next steps;
 - calculation of consequences.
 - 2) tactical planning of counter-operations.
3. Execution – a reflection of the information impact:
 - 1) identification and “smoothing” of an information occasion;
 - 2) counter-propaganda;
 - 3) operational intelligence;
 - 4) assessment of the information environment;
 - 5) adjustment of information counteraction.
4. Final phase:

- 1) performance analysis;
- 2) the use of positive results of information impact;
- 3) counteraction to negative results.

The operational management of information operations using information and analytical systems can be illustrated using the diagram shown in Fig. 90.

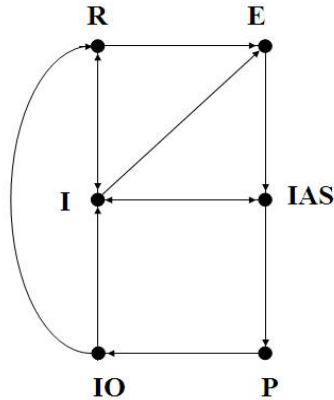


Figure 90 – Diagram of operational management using information and analytical systems

In accordance with the above diagram, information from the real world (R) enters the information space, in particular, to the media (I) or directly to experts (E), also through the media.

From experts or directly from the information space (for example, using content monitoring tools), information enters the information and analytical system (IAS). The information and analytical system transmits data to decision makers (P) that determine the measures of information impact on the information space and directly on real world objects (people, environment, computer systems, etc.).

6.3. Information Operations Modeling

Modeling can be seen as one of the ways to solve problems that arise in the real world, in particular, when planning and conducting information operations. Most often, simulation is used in cases where experiments with real objects are impossible

or too costly. Modeling covers mapping a real problem into an abstract world, learning, analyzing and optimizing the model, and mapping the optimal solution back to the real world.

When modeling, there are two alternative approaches – analytical and simulation. Ideal analytical models allow a rigorous analytical solution or at least a formulation, for example, in the form of systems of differential equations. However, analytical solutions are not always achievable. Therefore, especially in recent times, and especially in solving problems from the field of social dynamics, *simulation* modeling methods are increasingly being used. Simulation is a more powerful and almost indispensable tool for analyzing social procedures. The simulation model can be viewed as a set of rules that determine the future state of the system based on the current one. In this case, the modeling process consists in observing the evolution of the system in time according to these rules, and, accordingly, assessing the adequacy of the model, when possible.

The most promising direction in modeling information operations is the mathematical description of the self-organization of the environment for the perception and dissemination of information, taking into account the current conditions. Self-organizing environments, for which there is no central control mechanism, and development occurs due to many local interactions, are studied by the theory of complex systems. This theory covers such branches of knowledge as nonlinear physics, thermodynamics of nonequilibrium processes, and the theory of dynamical systems. Interactions between individual elements of complex systems determine the occurrence of complex behavior in the absence of centralized control. To study such behavior, the most modern methods are used, which are covered by the interdisciplinary basis of modern methodology – the concept of complexity. Currently, the theoretical and technological foundations of this concept include the theory of deterministic chaos, fractals and complex networks, synergetics, wave (wavelet) analysis, multi-agent modeling, the theory of self-organized criticality (studying the dynamic development to a critical state, characterized by strong space-time fluctuations, without external control [Bak, 1996]), percolation theory (Percolation – flow), etc.

Modeling social procedures (information operations, of course, belong to those) involves the conduct of computational

experiments, since most often there are significant limitations that make it difficult to conduct "field" natural experiments.

When modeling information operations, a computational experiment makes it possible to reduce the operations of clarifying constraints, selecting initial data, choosing the rules for the functioning of model components, etc. In this case, it becomes possible to take into account cases that are difficult to implement in practice, using real data only to identify the parameters of the mathematical model. At the same time, mathematical modeling has its limitations; the real world turns out to be difficult to model with a sufficient level of detail and accuracy, i.e. more or less reliable mathematical models are so complex and multi-parametric that they cannot be analyzed and evaluated by exact methods.

It is possible to work out mathematical models when planning information operations only in the process of modeling specific procedures, constantly comparing them with reality.

The expressed purpose of the information operations assessment methodology is to provide a timely and accurate analysis of possible discrepancies between the planned operation and the actual impact. When significant differences are found that affect the probabilities of success of the operation, the analytical system should report this to decision makers in order to correct current plans and decisions. At the same time, when planning information operations, it is impossible to act by trial and error, therefore, it is necessary to develop methods that allow generalizing retrospective data and, on their basis, to check the adequacy of models.

Successful information operations models are based on synergistic approaches. Indeed, society is a complex system, each component of which is characterized by many features, has many degrees of freedom. At the same time, an important property of this system is self-organization, which is the result of the interaction of such components as randomness, repetition, positive and negative feedback.

A feature of the mathematical modeling of information operations should be considered the comparative simplicity of interpreting the results. Such concepts as "the size of the electorate", "political weight", etc., are perceived on an intuitive level, even without getting to know the exact definitions, as far as they are pos-

sible here. And this makes it possible to make such an analysis of current situations a subject of wide discussion.

Due to the fact that some solutions are unstable with respect to their parameters, the values of such parameters must be determined with high accuracy. This requires a set of methods based not only on the processing of large volumes of statistical data, but also on versatile sociological research.

At present, the statement of the problem, which consists in using mathematical models to predict possible scenarios of the dynamics of social processes at a qualitative level, looks realistic. In this formulation, dynamics modeling occupies, as it were, an intermediate level between what is stated here and accurate forecasting. Nevertheless, it will be necessary to choose the values of the parameters that would, in some reasonable approximation, correspond to the situation under study, and in most cases the use of relative values turns out to be productive. So, of course, one cannot obtain reliable data on the future development of events, but, most likely, one can form a more or less adequate picture of what can happen and how. And this is not enough.

In order to achieve success in this case, individual information impacts must be considered as parts of a single information operation, just as shelling or air attacks can be considered as coordinated parts of a military operation.

At the same time, information operations have the following main features:

- information operations is an interdisciplinary set of methods and technologies in such areas as computer science, sociology, psychology, international relations, communications, military science;
- there are still no standards for conducting information operations;
- not only defense departments, but also many government and commercial organizations are interested in the development of information operations technologies;
- the task of forming a scientific approach to information operations is urgent and relevant.

When conducting information operations, it is essential to identify the content (knowledge) invested in information, taking into account a wide variety of aspects – social, political, reli-

gious, historical, economic, psychological, mental, cultural, inherent in various strata of society. Therefore, at present it makes sense to consider information operations more broadly, as operations based on knowledge (Knowledge Operations) [Burke, 2001].

A common network information attack in the web environment today is carried out as follows: as a rule, a website is created and operates for some time (let's call it the "original source"), while it publishes quite correct information. At hour X, a document appears on his page, usually compromising information on the object of attack, reliable or falsified. Then there is the so-called "laundering of information". The document is reprinted by online publications of two types – those interested in the attack and those who simply do not have enough information to fill their information field. In the case of claims, all reprinting publications refer to the "original source" and, in extreme cases, at the request / demand of the object of attack, remove information from their websites. The primary source, if necessary, also removes information or is completely eliminated (after which it turns out that it is registered on the Internet to a non-existent person). At the same time, the information has already spread, the task of the original source has been completed, the attack has started.

The modern information space is a unique opportunity to obtain any information on a chosen issue, subject to the availability of appropriate tools, the use of which allows you to analyze the relationship of possible events or events that are already taking place with the information activity of a certain range of information sources. On the other hand, in a retrospective analysis of any process or phenomenon, certain characteristics of its development are of interest, namely:

- quantitative dynamics inherent in a process or phenomenon, for example, the number of events per unit of time, or the number of messages related to it;
- determination of critical, threshold points, which correspond to the quantitative dynamics of the phenomenon;
- determination of manifestations at critical points, for example, identification of the main plots of publications in the media regarding the selected process or phenomenon;

- after identifying the main manifestations of the phenomenon at critical points, these manifestations are ranked, and the dynamics of the development of individual specific manifestations before and after certain critical points is studied;
- statistical, correlation and fractal analysis of the general dynamics and dynamics of individual manifestations is carried out, on the basis of which attempts are made to predict the development of the phenomenon and its individual manifestations.

To study the relationship between real events and publications about them on the Internet, the authors used the InfoStream system, which provides integration and monitoring of network information resources.

The number of web publications per day on any topic, and especially the changes (dynamics) of this value, sometimes allow even small experts in the subject area to draw more or less accurate conclusions.

You can get data on such dynamics, for example, by visiting the sites of news integrators daily (news.yandex.ru, webground.su, uaport.net). Of course, users of professional monitoring systems such as Integrum or InfoStream are in a better position. It is on the basis of the latter system that amazing statistics were obtained on the number of web publications on the subject of influenza epidemics in different periods.

As an example, consider the information campaign directed against Prominvestbank, which began at the end of September 2008.

With the help of the content monitoring system InfoStream (www.infostream.ua) [Grigoriev, 2007], which scans all the main information websites of Ukraine in real time, the dynamics of publications on the websites of messages that mention Prominvestbank was determined. for three months – September, October and November (Fig. 91). This dynamics testifies to a small number of publications in the first half of September, but then a number of publications began to compromise the chairman of the board V. Matvienko, which caused a relatively small resonance.

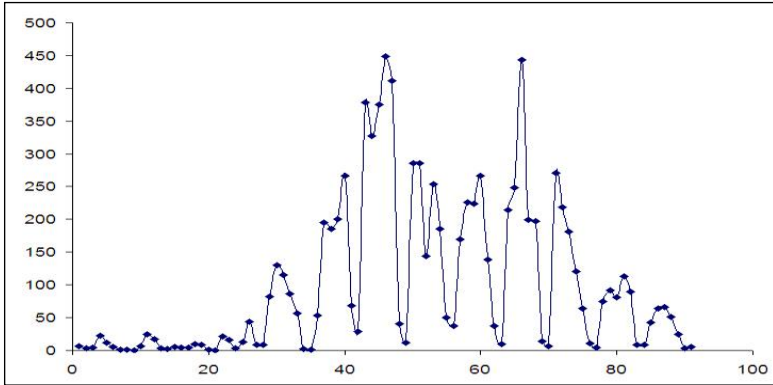


Figure 91 – Dynamics of publications on the topic "Prominvestbank" for three months 2008

As it turned out later, these publications were only "artificial preparation". On September 26, the first reports of a possible bankruptcy of the bank appeared (Figure 92), the number of which was quite consistent with an avalanche process, limited only by the number of websites capable of publishing such information. However, this process reached a stable-average level by December 2008.

It cannot be argued that only an information attack through the Internet led the bank to a sad state, but it was the first alarming messages that undermined the confidence of many depositors and forced them to massively withdraw their savings from the bank.

On September 30, it was reported that in order to save Prominvestbank, the National Bank of Ukraine (NBU) decided to allocate 5 billion hryvnias of refinancing to it, and on December 5, it was reported that Prominvestbank had a new owner (Fig. 93). After that, the volume of publications about Prominvestbank significantly decreased, which indicates not so much its recovery, but the systemic crisis of the banking system of Ukraine, which "dropped" many other credit and banking institutions.



В Донбассе вошла в активную фазу атака на Проминвестбанк. ПИБ заявляет, что это атака из-за рубежа

Сегодня в Донецкой области население организовано вышло к проходным **Проминвестбанка**.

Вести о том, что народные массы Донбасса штурмуют отделения ПИБа в Донецке, Авдеевке, Волновухе и пр. населенных пунктах Донецкой области, приходят в "Обком" с середины дня.

Никто из опрошенных нами экспертов не может пока сказать что-либо конкретное по данному поводу, кроме банальных констатаций: ПИБ - серьезный банк, он кредитует промышленный сектор Украины, Донецкое облотделение ПИБа - одно из крупнейших, борьба за него началась еще в середине 90-х годов... Ну а баннеры на киевских дорогах против нынешнего (неизменного) руководства ПИБа во главе с г-ном Матвиенко видели многие автомобилисты и пассажиры столичного транспорта.

"Обком" пока не готов сказать что-то определенное по поводу паники, которая охватила сегодня трудовой Донбасс - хотя сведения для определенных умозаключений, в принципе, имеются. Вместо этого мы предлагаем внимаю вкладчиков сообщение, поступившее от пресс-службы ПИБа:

"**Проминвестбанк** заявляет о стабильной работе, несмотря на дезинформацию в ряде СМИ о якобы приближающемся банкротстве банка.

Проминвестбанк, по оценкам зарубежных экспертов, стабильный банк и занимает в Украине второе место по надежности.

Массовая газетная атака на **Проминвестбанк** организована рейдерскими (бандитскими) группировками зарубежных агентов с участием высокопоставленных чиновников крупных государственных структур, которые по Конституции должны защищать отечественные предприятия и банки. Ложь, шантаж, направленные против банка, преследуют цель вынудить его к продаже иностранцам за комиссионное вознаграждение... Заявляем: банк не продается... **Проминвестбанк** останется украинским!" - говорится в сообщении.

Служба информационной поддержки **Проминвестбанка** также сообщает, что, несмотря на беспокойство вкладчиков, вызванное негативными публикациями о банке, все обязательства перед клиентами и вкладчиками выполняются, а структурные подразделения банка работают в нормальном режиме.

"Обком"

Figure 92 – One of the first alarm messages

Активная база данных: Система интеграции интернет-ресурсов

Главная Помощь Кабинет Источники Статистика Новости проекта

InfoStream Online

Период: Другой Убрать дубли Морфология

От: 200809 До: 200811

Найти Динамика Действительность

Очистить События Секеты

Риск запроса Прогноз

➤ (prominvestbank) & (2008.11.05)

Найдено документов - 443, страница 1 из 30

Статистика слов

ПРОМИНВЕСТБАНК - 26463, 2008.11.05 - 58209.

➤ Добавить канал

1. Матяевнюк поделился "Проминвестбанком" с братьями Клюевыми
 Печать 2008.11.05 22:08
 40% акций "Проминвестбанка" досталось братьям Клюевым. Эту информацию газете "Сегодня" подтвердил источник в руководстве Партии Регионов. По словам информатора, денег за это братья Клюевы не заплатили.
 Похожие документы - Оригинал
2. ВЧЕРА ОФИЦИАЛЬНЫЙ И РЫНОЧНЫЙ КУРСЫ ДОЛЛАРА ПОЧТИ СРАВНЯЛИСЬ, И В ОБМЕННИКАХ АМЕРИКАНСКУЮ ВАЛЮТУ ПРОДАВАЛИ ПО 5,85 ГРИВНИ
 Газета "Факты и комментарии" 2008.11.05 21:30
 А "Проминвестбанк" обрел новых владельцев Роберт ВАСИЛЬ "ФАКТЫ" Национальный банк в среду продолжил свою деятельность по укреплению курса гривны на наличном и межбанковском рынке и параллельно по ослаблению официального курса американской валюты. Курс доллара, установленный Нацбанком, вчера вырос приблизительно на 3,5 копейки и достиг значения 5,8261 гривны за доллар.
 Похожие документы - Оригинал
3. "Проминвестбанк" сменил собственника
 Газета "День" 2008.11.05 21:22
 Правительство утверждает, что не будет расходовать деньги налогоплательщиков на рекапитализацию банка Наталья БИЛЮСОВА. "День" Вчера об этом официально сообщил на своем сайте Национальный банк Украины (НБУ).
 Похожие документы - Оригинал
4. Кабинет министров Украины утвердил правила для капитализации банков
 УРА-Информ 2008.11.05 21:13
 Кабинет министров Украины утвердил порядок участия государства в капитализации банков. Соответствующее постановление от 4 ноября 2008 г. N 960 размещено на сайте правительства.
 Похожие документы - Оригинал
5. Дмитрий Фирташ покупает почти 90% банка "Надра"
 Времена.info 2008.11.05 20:17
 Вчера, 4 октября, украинский предприниматель Дмитрий Фирташ, совладелец швейцарского газового трейдера RosUkraine, подписал предварительное соглашение о покупке 86,7% акций банка "Надра".
 Похожие документы - Оригинал
6. У Проминвестбанка поменялся владелец (5.11.2008 18:00)
 ИДУ (рус.) 2008.11.05 19:45
 лев У Вас есть видео, которое Вы хотите показать всему миру? Вам сюда В Проминвестбанка изменился владелец. Факт продажи акций банка подтвердили в НБУ. СТВ
 Похожие документы - Оригинал
7. Население будет покупать доллары по официальному курсу
 АМИ Новости-Украина 2008.11.05 19:45
 Национальный банк Украины своим постановлением N353 от 5 ноября обязал коммерческие банки продавать населению наличные доллары по курсу не выше официального, сообщает "Украинский новини".
 Похожие документы - Оригинал

Информационный портрет	
Уточнить запрос	
Языки (1)	
Страны источников (1)	
Источники (1)	
Размер (1)	
Цифровая насыщенность (1)	
География (1)	
Компани (1)	
Слова (12)	
Классификатор-навигатор	
ОБКМ	
ВОЛНОВАХА	
АТАКА	
ФАЗА	
РУБЕЖ	
ВКЛАДЧИК	
ДОНБАСС	
ГОТ	
ВОЛНОВАХА	
АТАКА	
ФАЗА	
РУБЕЖ	
ВКЛАДЧИК	
ДОНБАСС	
УМОЗАКЛЮЧЕНИЕ	
ВОЛНОВАХА	
АТАКА	
ФАЗА	
РУБЕЖ	
ВКЛАДЧИК	
ДОНБАСС	
ПРОХОДНАЯ	
ВОЛНОВАХА	
АТАКА	
ФАЗА	
РУБЕЖ	

Figure 93 – Messages that completed extreme dynamics intensity of publications on the topic "Prominvestbank"

Literally a week after the events described above, another public landmark information attack took place in Ukraine, this time on the insurance market. It was a real information operation against the National Joint Stock Insurance Company (NASK) "Oranta". In this case, the primary source of compromising material was not a website, but an informational message sent by e-mail to thousands of Internet users. As a result of the use of special technical methods, it diverged from the designation of the address of the press service of the object of attack. So, on December 10, 2008, around 11:30 am, an informational message was sent in the form of spam, stating that the Oranta insurance company was declaring bankruptcy. According to preliminary

data, the information was scattered to 1000 addresses, naturally, the data got to competitors and the media. The message said that from December 31, 2008, the company ceases to fulfill its obligations to customers.

In connection with the incident, NJSIC "Oranta" appealed to law enforcement agencies with a request to investigate this incident and punish those responsible. What happened with NJSIC "Oranta" was very similar to the situation with "Prominvest-bank", numerous experts agreed with this. After all, both the banking business and the insurance business are based on the trust of customers, which is most easily undermined by information attacks. According to Oleg Spilka, Chairman of the Supervisory Board of NJSIC "Oranta", "This event was deliberately prepared in order to discredit the insurance company and undermine its reputation." Without going into details of the possible targets of the attack (change of owners, struggle for a blocking stake, destruction of the company, etc.), with the help of a retrospective analysis, we will follow the dynamics of publications on the Internet that mentioned NJSIC "Oranta".

On fig. 94 shows the daily dynamics of the number of relevant publications. On this diagram, among other things, a decline in the intensity of publications on this topic in early December is 2008 clearly visible, which can be perceived as some kind of "calm before the storm".

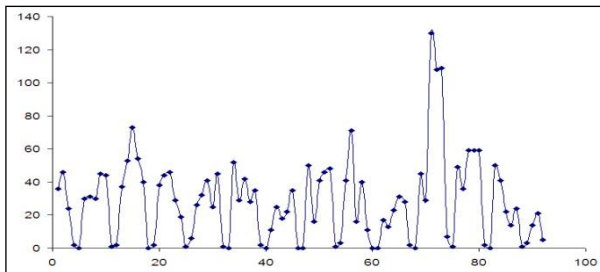


Figure 94 – Intensity of publications on the Internet on the topic "Oranta"

For the analysis of time series in the framework of the study, the authors used ΔL the -method. On fig. 95 shows the scalogram of the dynamics of the process under consideration using

the method (ΔL -method) for the second half of 2008. Despite separate peaks on the 16th and 55th days of the quarter, the extremum falling on December 10–12 is of the greatest interest.

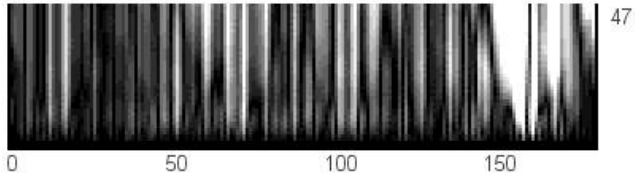


Figure 95 -- ΔL diagram of a number of publications on the topic "Oranta"

More detailed statistics of publications on the topic "Oranta" for December 2008 was obtained through the user interface of the InfoStream content monitoring system (Fig. 96).



Figure 96 – Detailed intensity chart publications on the topic "Oranta"

Let us follow the course of the information operation by examining messages published at different time intervals.

On Fig. 97 is a list of publications on the topic "Oranta" during the first hours of the attack. According to Oleg Spilka, within

6.4. Identification of information operations

For the operational analysis of the information environment in order to identify information operations, specialized systems for monitoring the information space (content monitoring) are used. Such systems provide, firstly, efficiency that traditional search engines cannot provide (the time for indexing network content, even by the best of them, ranges from several days to several weeks). Secondly, completeness (both in terms of sources and presentation of source materials), which is not always provided by ordinary news aggregators. And, thirdly, the necessary analytical tools that allow the user to create analytical reports based on publications on a given topic in the required period of time.

In terms of prevention of information operations, one should carefully monitor the dynamics of publications about the target company, if possible, taking into account the tone of these publications, use available analytical tools, for example, wavelet analysis. At the same time, one should be guided by possible models of information attacks, for example, if this model covers the phases: "background publications" – "lull" – "artillery preparation" – "lull" – "attack" (Fig. 52), then already for the first three components can predict future events with a high probability.

The above plan is obviously the ideal one, focused solely on web resource content monitoring data.

Of course, users of professional content monitoring systems are in a better position. Many modern information-analytical systems contain means of displaying statistics of occurrences in databases of concepts that correspond to user requests. In particular, the authors used the statistics subsystem as part of the InfoStream web space content monitoring system, which implements this functionality.

When studying trends in information operations, it is precisely the series by the number of thematic publications for a certain period of time (most often, per day) corresponding to these information operations that are considered as time series. Therefore, in order to identify trends, information flows are studied that correspond to the topics of information operations – thematic information flows.

given in [Gorbulin, 2009] corresponding to the stages of the information operation are shown in Fig. 2. 100. At the same time, analysts should focus on such models, for example, if monitoring allows you to determine the phases: "background" – "calm" – "artillery" – "calm" – "attack", then the first three components can be used to predict future events with a high probability.

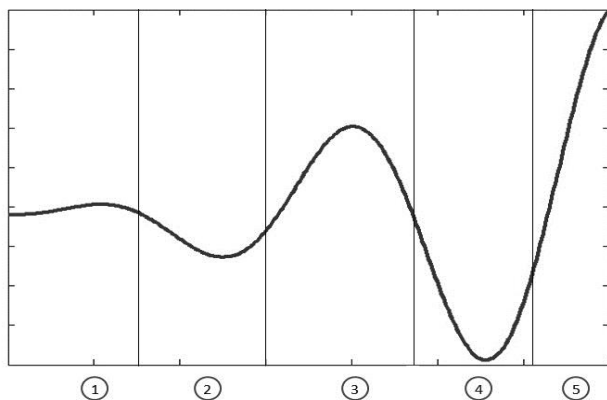


Figure 100 – Dynamics of the number of thematic messages during the information operation: 1 – background; 2 – calm; 3 – "artillery preparation"; 4 – calm; 5 – attack/growth trigger

It should be noted that such dynamics of the number of thematic messages during information operations is well described by the well-known equation for the propagation of electromagnetic waves:

$$y = A + Bx \sin(x),$$

where x is time, A and B are constants determined empirically.

As is known, at present, innovation activity is also indirectly measured by the number of publications related to innovation; there are several models of innovation processes, among which the innovation diffusion model can be distinguished [Bhargava, 1993]. At the same time, the introduction of innovations can also be considered as information operations. Therefore, we turn to the results of relevant studies. On fig. 101 shows the diagram of

the number of publications substantiated in [Khoroshevsky, 2012], which corresponds to the trend of innovation activity.

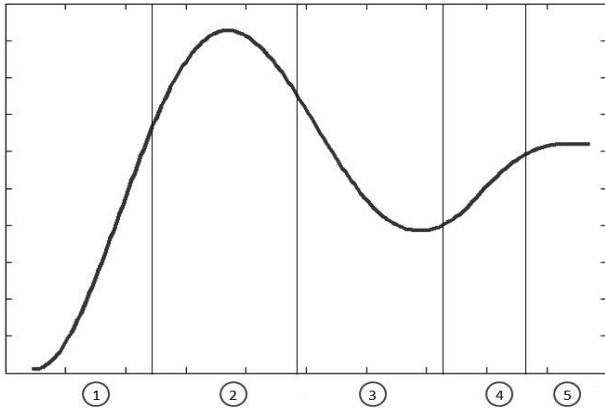


Figure 101 – Chart of the number of publications corresponding trend of innovation activity: 1 – attack/growth trigger; 2 – the peak of inflated expectations; 3 – loss of illusions; 4 – public awareness; 5 – productivity / background

Combining graphs corresponding to the beginning of the information operation (Fig. 100) and the trend of innovation (Fig. 101), you can get a complete graph corresponding to the display of information operations in the information space (Fig. 102).

The proposed models are fully consistent with real data, which are extracted by content monitoring systems [Dodonov, 2009], [Lande, 2007]. Therefore, the given dependencies can be used as templates for identifying information operations, both by analyzing the retrospective fund of online publications, and for real-time monitoring of the appearance of some of their signs. As you know, to identify information operations, one should carefully monitor the dynamics of publications on the target topic and, if possible, use available analytical tools, digital data processing and pattern recognition tools, for example, wavelet analysis or Kunchenko polynomials [Chertov, 2009].

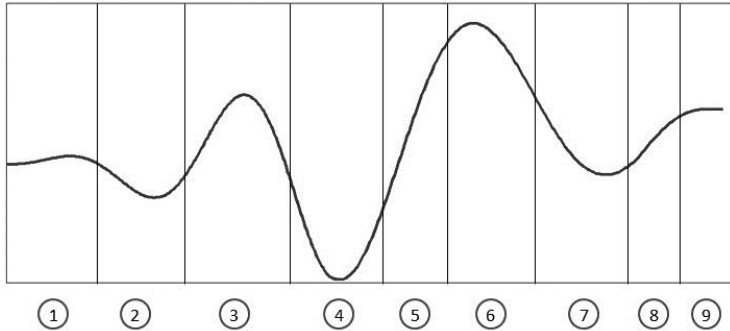


Figure 102 – A generalized diagram corresponding to all stages of the life cycle of information operations: 1 – background; 2 – calm; 3 – "artillery preparation"; 4 – calm; 5 – attack/growth trigger; 6 – the peak of inflated expectations; 7 – loss of illusions; 8 – public awareness; 9 – productivity / background

As an example, in fig. Figure 55 shows the dynamics of publications in RUNet – thematic information flows on requests "Banks, Cyprus", "Offshore", "Virgin Islands" for March-April 2013, during the period of known crisis events, obtained using the InfoStream system. As can be seen from Fig. 103, the peak of publications related to the banking crisis in Cyprus falls on March 17-18, 2013, while most of the publications on the Virgin Islands fell on April 4-5, when there, with a much smaller scale, began to appear events similar to those in Cyprus. At the same time, it should be noted that there is a weak correlation between the dynamics of information flows related to Cyprus and the Virgin Islands. In this case, the cross-correlation coefficient of the corresponding numerical series was only 0.3. At the same time, there is a high level of cross-correlation of the series corresponding to the topics "Offshore" and "Banks of Cyprus" (0.73), as well as "Offshore" and "Virgin Islands" (0.77).

Apparently, the manifestations of information operations in the field of offshore banks in this case are best seen when analyzing a more general topic – "Offshores". The graph of the corresponding numerical series clearly shows two areas of local extremes corresponding to the crisis situations in Cyprus and the Virgin Islands, as well as phases corresponding to "lulls" and "artillery preparations".

It can be assumed that if the dynamics of a private information flow at some point begins to differ significantly from the dynamics of a flow corresponding to a more general topic (as in the case under consideration, “Banks of Cyprus” and “Offshore”), then signs of the beginning of an information operation may appear, relating to a narrow topic.

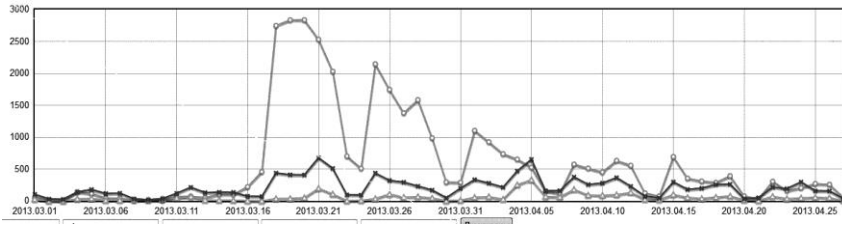


Figure 103 – Diagram of the dynamics of thematic information flows on requests: o – "Banks of Cyprus" ; Δ – "Virgin Islands"; x – "Offshore"

When performing the wavelet analysis [Astafieva, 1996], [Buckheit, 1995] (Fig. 56), it was decided to use the “Mexican hat” wavelet, as it is similar in shape to the diagram shown in Fig. 5b. 104.

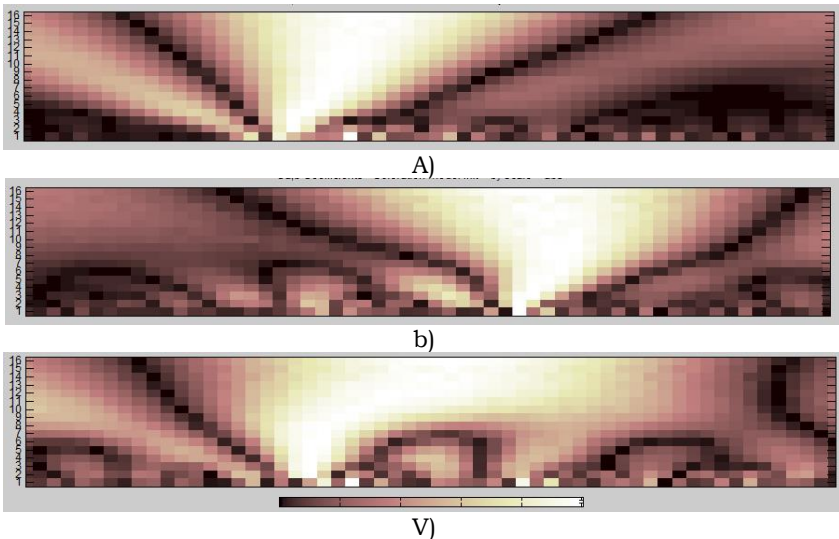


Figure 104 – Wavelet spectrograms corresponding to the dynamics of thematic information flows on requests: a – "Banks of Cyprus" ; b – "Virgin Islands"; in – "Offshore"

The processes under consideration are clearly visible both on the wavelet spectrograms and on the corresponding skeletons (plots of extremum lines).

The presented models and methods are suitable for describing general trends in the dynamics of information processes, however, the problem of forecasting remains open. Apparently, more realistic models can be obtained taking into account an additional set of factors, most of which are not reproduced in time. At the same time, the structure of the rules underlying the functioning of most of the available models allows making appropriate adjustments, for example, artificially modeling random deviations.

Note that the reproduction of results in time is a serious problem in modeling information processes and forms the basis of scientific methodology. Currently, only a retrospective analysis of already implemented information operations remains a relatively reliable way to verify them.

Naturally, in practice, focusing only on a single type of source can lead to a lack of information necessary for decision-making, inaccuracies, and sometimes misinformation. Only the use of complex systems based on the use of numerous sources and databases, along with the above capabilities of the content monitoring system, can guarantee effective information support when countering information operations.

The selected patterns of behavior of the intensity series of thematic publications can be considered as patterns (samples) of functional dependence. These templates can be taken as a single basic element of some linear space, i.e. as a generating element e for modeling using Kunchenko polynomials [Chertov, 2009].

Whereas a linear combination of linearly independent transformations $f_1(e), f_2(e), \dots, f_n(e)$ of the corresponding generating element can be constructed by an P_n approximation polynomial n of the t th order to a part of the output signal $f_s(e)$:

$$P_n = \sum_{\substack{k=0, \\ k \neq s}}^n c_k f_k(e),$$

where the coefficients c_k are determined from the condition of ensuring the minimum distance between the polynomial under construction and the signal. Element c_0 is defined by the expression:

$$c_0 = \frac{\langle f_s(e), f_0(e) \rangle - \sum_{k=1, k \neq s}^n c_k \langle f_k(e), f_0(e) \rangle}{\langle f_0(e), f_0(e) \rangle},$$

and other coefficients c_k - as a solution to the system of linear equations:

$$\sum_{k=1, k \neq s}^n c_k F_{i,k} = F_{i,s}, \quad i=1, \dots, n, \quad i \neq s,$$

where the centered correlates $F_{i,k}$ are also calculated using the appropriate transformations:

$$F_{i,k} = \langle f_i(e), f_k(e) \rangle - \frac{\langle f_i(e), f_0(e) \rangle \cdot \langle f_k(e), f_0(e) \rangle}{\langle f_0(e), f_0(e) \rangle}.$$

A numerical characteristic that can be used in the quality criteria for matching a signal with a selected template, i.e. as a measure of the approximation of the Kunchenko polynomial P_n to the signal $f_s(e)$, can be considered the efficiency coefficient d_n :

$$d_n = \frac{\sum_{k=1, k \neq s}^n c_k \langle f_k(e), f_s(e) \rangle}{\langle f_s(e), f_s(e) \rangle}.$$

The considered method of recognizing certain patterns by constructing a space with a generating element and searching for the coefficients of the corresponding Kunchenko polynomial can

be used in any problem area in which certain characteristic patterns can be identified a priori in the time series.

Thus, having built typical models of the behavior of the intensity series of thematic publications during information operations and comparing the patterns obtained on their basis, it is possible to use the method based on Kunchenko polynomials to determine (and prevent) a possible information attack.

The dynamics of thematic information flows is determined by a complex of both internal and external non-linear mechanisms that should be reflected in the modeling (perhaps in an implicit form). Often, a simplified understanding of the thematic information flow as some time-dependent quantity, the behavior of which is described analytically by nonlinear equations, turns out to be satisfactory. Today, when modeling information flows, mainly analytical nonlinear models are used, methods of nonlinear dynamics, the theory of cellular automata, percolation, and self-organized criticality are used [Lande, 2009], [Dodonov, 2011].

To analyze the dynamics of real thematic information flows (TIF), and, accordingly, to evaluate their models, it is necessary to somehow obtain the corresponding statistics presented in the form of time series.

The dynamics of real thematic information flows (TYP), for example, is displayed by a multi-agent model, in which individual TYP documents are associated with agents whose life cycle is the life cycle of documents in the information space. Accordingly, the entire space of the multi-agent model is associated with the thematic information flow.

the evolution of the population of agents occurs during discrete moments of time. In this case, individual agents can:

- 1) "self-generated" (be born for reasons that arise outside the considered multi-agent space);
- 2) "spawn" new agents;
- 3) "die" – disappear from the space of agents (corresponds to the loss of relevance of documents);
- 4) receive links from other agents.

Each agent has a "potential", depending on his age (current lifetime – t), authority (links put on him – ns) and fertility (number of agents directly generated by him – k). The agent's potential Pot is determined by the formula:

$$Pot = \frac{1 + ns + k}{t}$$

On fig. Figure 105 shows an example of the possible dynamics of a multiagent system: the processes of birth of new agents from existing ones are indicated by solid arrows, the processes of placing links to agents are represented by dotted arrows, living agents are shown by black circles, and “dead ” agents by the time $t = 5$ are indicated by empty circles.

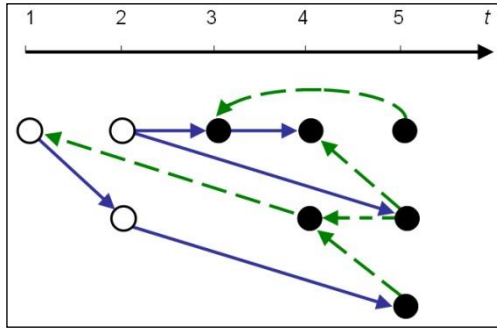


Figure 105 – Fragment of the multi-agent space

So, the control parameters of the model are as follows:

- the probability of "spontaneous generation" P_1 ;
- the probability of "birth" from the existing:
 $P_2 \cdot Pot$;

is the probability of the agent's "death": P_3 / Pot ;

- probability of a link to an agent: $P_4 \cdot Pot$.

The variation of these four parameters P_1 , P_2 , P_3 made P_4 it possible to model typical profiles of TYP behavior.

On Fig. 106 presents the results of numerical simulation of the number of agents (y-axis on the graph) in the considered multi-agent system depending on the number of cycles of the model (abscissa axis).

The considered model of agent space evolution for different values of control parameters is consistent with the dynamics of

real thematic information flows determined using the InfoStream system.

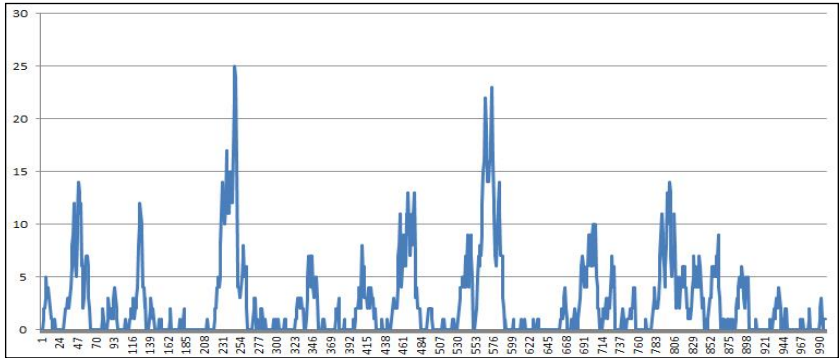


Figure 106 – Dynamics of changes in the number of agents in the model

In practice, focusing only on a single type of sources and mathematical models can lead to a lack of information necessary for decision-making, inaccuracies, and sometimes misinformation. Only the use of complex systems based on the use of numerous sources, databases, mathematical models, along with the above capabilities of content monitoring systems, can guarantee effective information support when countering information operations.

6.5. Ways to counter information operations

The considered practical examples made it possible to develop some general methodology for conducting a defensive information operation using a web resource content monitoring system. Suppose the object of an aggressive information operation is the company "ABV". The following 12 countermeasures are proposed:

- 1) collection of information with publications in "foreign" (not related to "ABV", non-affiliated) media about the company;
- 2) building a graph – the dynamics of the appearance of messages about the company "ABV" in the online media;

3) and analysis of dynamics with a retrospective of 6–12 months using time series analysis methods. After that, the content of publications is analyzed at threshold points, the moments, duration, frequency of impact are determined, the binding of the moments of impact to other events from the area of interest of the object;

4) identification of sources publishing the largest number of negative (publications with a negative tone) about the company "ABV";

5) determination of the "primary sources" of publications in the media – those sources that were the first to publish negative information ;

6) identification of probable "customers" – owners or persons influencing the publishing policy of individual media;

7) identification of areas of common interest of the ABV company and potential "customers" (by identifying common information characteristics – intersections of "information portraits" of the InfoStream system, built for the object and the "customer"), ranking potential "customers" according to their interests;

8) determination of criteria for information impacts based on the most rated interests;

9) modeling of information impacts, for which the connections of the "customer" are found – the persons and organizations most associated with it, the dynamics of the impact on the part of the customer is analyzed and a forecast of this dynamics is built, the content of publications is analyzed at the threshold points of the dynamics curve – critical impact points are determined. ;

10) further steps of impact are predicted by analyzing the similar dynamics of publications for other companies in the retrospective database of the InfoStream system;

11) taking into account the realities and publications from the retrospective database, the likely consequences are assessed;

12) informational (and not only) counteraction is organized. Examples of publications in the context of resistance are in the retrospective database.

6.6. Examples of information operations

Antimonopoly activities, the creation of a competitive environment in the state, involve the fight against manifestations of monopoly in the markets for goods and services, including the reflection of relevant information operations conducted by monopolists, the conduct of offensive information operations.

To carry out antimonopoly activities on the part of the state, to create a competitive environment, it is necessary to use all available and legal information and software tools. However, today there is a real shortage of operational market information, determined both by weak communications between individual authorities and by the incompleteness and inaccuracy of the relevant official databases. On the other hand, there is a huge information resource – the web space.

Obviously, despite the advantages of speed and wide coverage of information, this resource cannot be a demonstrative source, however, it cannot be rejected in some important applications. The agility inherent in the web, in particular, is critical to the implementation of the OODA governance concept, also known as the Boyd cycle. Translated, the abbreviation OODA means "Observation – Orientation – Decision – Action" [Ivlev, 2008]. The OODA concept is widely used throughout the world in the management of information confrontation, the prevention of information operations. Obviously, this concept can and should be applied in antimonopoly activities through the implementation of rapid response centers for monopoly manifestations.

It is well known that antimonopoly activity is a set of measures aimed at limiting the activities of monopolies throughout the state, as well as creating appropriate legislation, while competitive intelligence is aimed at increasing the competitiveness of only individual business entities. According to this, the particular tasks of competitive intelligence can be generalized to the level of antitrust activities at the state level as follows:

- 1) collection of information and timely information support of the relevant state bodies;
- 2) identification of risk factors, threats to the competitive environment of the state;
- 3) identification of factors influencing the obtaining of monopoly advantages by individual companies;
- 4) development of forecasts and recommendations that affect the development of the competitive environment;

5) strengthening of favorable and localization of unfavorable factors for the development of a competitive environment.

With methods competitive intelligence, which becomes modern direction of research behavior competitors on the market, are created alternative market models to characterize his participants and optimization tactics and strategies development of subjects management on certain markets. Achievement such purposes requires the use effective tricks work with information and its elements. Information in this sense becomes object in the process of research market and creation his models.

All the above tasks are implemented within the framework of a closed scheme of interaction between the market environment and the virtual information space.

As you know, market reality is reflected in the virtual information space, it is with it that expert analysts work, who prepare information, forecasts for decision makers, which, in turn, provide a targeted impact on the market environment.

Apparently, all of these functional components of competitive intelligence can also be used for general tasks facing the antimonopoly authorities of the state.

We will illustrate the possibilities of using competitive intelligence tools, in particular, content monitoring tools, in antitrust activities using the example of the buckwheat price collapse at the beginning 2010 in Ukraine. Antimonopoly Committee of Ukraine only in October 2011 (after a year and a half!) discovered and punished participants in the collusion in the buckwheat market (Fig. 107), while hundreds of users of the InfoStream content monitoring system could see the figures of the case already in February 2010 in the "information portrait" of this system (Fig. 108).

Definitely a support system antimonopoly activities like _ And systems competitive intelligence, using Internet as one of information resources, must tune in for the specifics concrete markets. She must include relevant classification, flexible search engines, operational data delivery, and quality evaluation information.

One of the most important tasks of information analysis in this case is to determine its reliability, i.e. solving the problem of analyzing and filtering noise and false information. After analyzing the reliability of information, assessments of its accuracy and

importance should follow. The main criterion for the reliability of data in practice is the confirmation of information by other reliable sources.

Документ по запросу

Укррудпром 2011.10.25 09:56
http://www.ukrudprom.ua/digest/AMKU_nashe_viditeley.html

AMКУ нашел вредителей
 [Экономические известия, No 186, 25 октября 2011]

Гречневый ценовой коллапс стал новостью номер один среди покупателей. Теперь же нашлись и виновники событий полугодовой давности.

Антимонопольный комитет Украины (АМКУ) оштрафовал восемь предприятий в целом на 590 тыс. грн. за антиконкурентные согласованные действия на рынке гречневой крупы. Как сообщают в Антимонопольном комитете, в частности, к ответственности привлечены такие субъекты хозяйствования, как "Родной продукт" (ТМ "Хуторок"), "Сельхозсервис" (ТМ "Фабрика крул"), ДП "Крупозавод Озерянка" (ТМ "Озеряночка"), "Сквирский комбинат хлебопродуктов" (ТМ "Сквирянка"), ЗАО "Нива" (ТМ "Добродия"), КП "Белоцерковхлебопродукт", фермерское хозяйство "Дар земли", ДП "Новоукраинский комбинат хлебопродуктов" ГАК "Хлеб Украины". Как установил АМКУ, с января по февраль 2010 г. указанные субъекты хозяйствования безосновательно одновременно повысили оптово-отпускные цены на гречневую крупу.

Такие действия предприятий, утверждая в комитете, привели к общему экономическому необоснованному повышению цен на гречку и привели к ущербу интересов потребителей. "В результате ответчики прекратили нарушения", - сообщается в письме комитета.

Напомним, что прошлой зимой в Украине резко выросли цены на гречневую крупу - до 20-25 грн. Кабинет решил закупить гречку в Китае, после чего цены на крупу снизились до 16-18 грн. Однако китайская гречка оказалась не дешевой. Предпродажная себестоимость продукта достигала почти \$2100/т.

Сохранить
 Распечатать
 Отправить

О документе

Агротром Рубрики (2)
 Экономика Украины

русский Языки (1)

средний Размер (1)

Цифровая насыщенность (1)
 малая

Украина География (2)
 Китай

Персоны (5)
 Ярославский
 Спассен
 Присяжнюк
 Колесник
 Арасланов

Figure 107 – The culprits of the crisis are found (Oct. 2011)

От: [201001] До: [201002]

🔍 (цен-2/гречк) & Украина

Найдено документов - 76, страница 1 из 6

Статистика слов
 ЦЕН - 1686954, ГРЕЧК - 4086, COUNTRY.UA - 1227403

Добавить канал

1. В Запорожье гречка подорожала из-за поста?
 Экспресс Запорожье 2010.02.26 14:11
 Подорожание продуктов питания, особенно гречки, цена которой возросла почти на 20% объясняется повышением закупочной цены, которую запорожские покупатели не устанавливают и не контролируют
 Похожие документы - Оригинал

2. Александр Син: Да пусть они задавятся своей гречкой!
 Политсовет 2010.02.26 12:56
 Мария Жартовская Заместитель губернатора Запорожской области Александр Син возмущен ростом цен на продукты питания в Запорожской области. В частности, за неделю стоимость продуктов выросла во всех районах области. Особенно Александра Сина возмутил рост цен на гречку. Стоимость гречки за последнюю неделю выросла на 20% - она уже в три раза дороже, чем крупа.
 Похожие документы - Оригинал

3. Александр Син: Да пусть они задавятся своей гречкой!
 Политсовет 2010.02.26 10:57
 Мария Жартовская Заместитель губернатора Запорожской области Александр Син возмущен ростом цен на продукты питания в Запорожской области. В частности, за неделю стоимость продуктов выросла во всех районах области. Особенно Александра Сина возмущен рост цен на гречку. Стоимость гречки за последнюю неделю выросла на 20% - она уже в три раза дороже, чем крупа.
 Похожие документы - Оригинал

4. Обмелели молочные реки
 Рабочая газета 2010.02.26 09:13
 Несмотря на заверения премьер-министра Тимошенко о том, что после выборов цены на продукты питания стремглав полетят вниз, этого не происходит. До 7, а то и 10 гривен за kilo подорожала в последние дни картошка.
 Похожие документы - Оригинал

5. За несколько дней "гречка" в Донецкой области подорожала почти вдвое
 КИД 2010.02.25 16:53

Информационный портрет
 Уточнить запрос

Рубрики (7)

Языки (2)

Страны источников (2)

Источники (50)

Размер (2)

Цифровая насыщенность (1)

География (17)

Персоны (16)

Компании (21)

AND NOT

Нива ***

Сквирский комбинат хлебопродуктов ***

ИПС ***

Родной продукт ***

Сильхозсервис **

Сельхозсервис *

Озерянка

Сильхозсервис

Ридный продукт

Figure 108- Reflection of the "buckwheat crisis" 2010

Conditions market research by using electronic funds, in particular during antimonopoly activities must correspond modern conditions competitive intelligence :

First, methods and software tools for researching information obtained from open sources should be applied in compliance with legal requirements and ethical standards.

Secondly, the success or failure in solving the practical problem of modeling the state of the market depends on the simplification of the integrated information that needs to be processed.

Thirdly, achieving success in market research is associated with the problem of overcoming the difficulty of accessing information resources from open sources, including the Internet.

The methodology for identifying anti-competitive actions of market participants based on the results of the analysis of the market state model should correspond to the capabilities of available computer tools and competitive intelligence methods.

For example, data, information, and knowledge resulting from antitrust research should be presented in a manner that is consistent in structure and form with intelligence information.

Modern tools used in competitive intelligence in a network environment provide:

- the availability of the necessary part of the information;
- huge scope of information;
- efficiency, taking into account the dynamics of information flows.

At that same time, these funds cannot replace All tools needed for antimonopoly activities. For acceptance decisions in this areas use required complex systems that allow mine and generalize information about objects research from different sources.

Thus, we can conclude that competitive intelligence complements the technology of searching for data and information in the Internet space and the targeted extraction of useful concepts about the state and development of the product market with methods for collecting, storing, processing and analyzing data, creates a space of integrated information for analysis. and formation of competition policy.

The goals and means of antimonopoly activities determine the practical requirements for the creation of new mechanisms and technologies and require the combination of competitive in-

telligence tools of various nature in accordance with various research algorithms.

Conclusion

The relevance of competitive intelligence has recently increased significantly. This is due to such processes as the globalization of the economy, and, consequently, competition, the virtualization of the economy, and the development of information technology.

Legislative acts of many countries of the world also contribute to the widespread introduction of computer competitive intelligence systems. For example, in the United States back in 1996, the Freedom of Information Act was adopted, which obligated federal agencies to provide citizens with free access to all their information. The restrictions apply only to materials related to national defense, personal and financial documents, and documents of law enforcement agencies. Denial of access to information can be challenged in court. Information must be submitted within ten days, and disputes must be resolved within 20 days.

For more than 20 years all over the world, it has been believed that competitive intelligence is the most important function of modern management and the main condition for dynamic and sustainable business development. At the same time, according to Roman Romachev, CEO of R-Techno, "While 10 years ago, competitive intelligence officers first of all checked whether business partners had criminal connections, now they, like in the West, are mostly extracting commercial information." This is confirmed by the data of a study conducted by the International Conference Center (ICC) OnConference: most companies use competitive intelligence to study the state of the market (74 % of respondents) and competitors (64 %). The search, collection and analysis of information helps to form a complete picture of the competitive environment, to establish cause-and-effect relationships.

Currently, competitive intelligence on the Internet provides accessibility, a huge coverage of information and high efficiency. But it cannot replace other types and tools of business intelligence. To make serious decisions, it is necessary to use complex systems that allow compiling and summarizing information about the object of research obtained from different sources using different technologies.

Numerous publications, trainings, conferences speak about the relevance of competitive intelligence based on Internet resources. Today, the tasks of competitive intelligence stimulate the development of knowledge management systems, deep analysis of data and texts, on the other hand, the most developed of these systems explicitly contain analytical blocks specifically focused on the tasks of competitive intelligence. Therefore, users have a wide choice of automation tools for analytical activities. Moreover, the levels of functionality of such systems can be very diverse – from simple information retrieval programs, necessary at the stage of formation of competitive intelligence systems, to expensive and resource-intensive knowledge management systems and in-depth analysis of data and texts.

To effectively analyze modern information processes based on monitoring information flows from global computer networks, modern methods based on nonlinear analysis should be used, many of which have been successfully used in the natural sciences. Modern approaches make it possible to apply methods that have been tested primarily in the natural sciences for the analysis and modeling of even social and information systems. Analysis of information flows is the foundation of such areas as modeling, design and forecasting.

At the same time, the above models and methods are suitable for describing general trends in the dynamics of information processes, however, the problem of forecasting remains open. Apparently, more realistic models can be obtained taking into account an additional set of factors, most of which are not reproduced in time. At the same time, the structure of the rules underlying the functioning of most of the available models allows one to make appropriate adjustments, for example, artificially simulate random deviations. Note that the reproduction of results in time is a serious problem in modeling information processes and forms the basis of scientific methodology. Currently, only a retrospective analysis of already implemented information operations remains a relatively reliable way to verify them.

At present, it is already obvious that a real breakthrough in the field of intensification of information and analytical work, as in science, is possible only as a result of aggregation of various areas.

Brief glossary

Automatic Referencing [Automatic text summarization] – automatic generation of a summary of the source text material, either by extracting fragments of information content and then combining them, or by generating text based on the identification of knowledge from the original.

Copyright is a set of legal norms governing relations arising in connection with the creation, use (publication, performance, display, etc.) of works of science, literature or art – the results of people's creative activity. Computer programs and databases are also protected by copyright.

Analysis [SNA] is a methodology for analyzing social networks. The subject of analysis in SNA, unlike most traditional sociological studies, is not the attributes of individuals, but the structure of their relationships within a particular community (working group). As part of the analysis of social networks, social relations are considered from the point of view of the theory of networks, consisting of nodes – personalities, network participants and connections – the relationship between them.

Antimonopoly activity [Antitrust Assets] – a set of measures aimed at limiting the activities of monopolies, as well as the creation of appropriate legislation.

Database is a set of data organized according to certain rules, providing for general principles of description, storage and manipulation, independent of application programs. It is an information model of the subject area.

Analytical database – a database that contains information obtained from other databases in the form of summary information, is of the greatest interest to a user or group of users.

Full-Text Database – a database that stores records of full-text documents or their parts.

Factual database [Factographic Database] – a database containing factual data – information related directly to the subject area.

Benchmarking is a competitor analysis tool. The process of identifying, understanding and adapting existing examples of the company's performance in order to obtain information that

would help take steps to improve the company's performance. It equally includes two processes: evaluation and comparison. One of the directions of strategically oriented marketing research.

A business process is a system of consistent, purposeful and regulated activities in which, through a control action and with the support of certain resources, process inputs are converted into outputs that are of value to consumers.

Business Intelligence – 1) collection and processing of data from different sources to develop managerial decisions in order to increase the competitiveness of a commercial organization ; 2) a structural subdivision of the enterprise that performs these functions.

Blog [Web Log] is an online diary of one or more authors, consisting of entries in reverse chronological order. Using the blog service, you can create your own online diary, read and comment on the diaries of other users, take part in communities on certain topics, and create your own communities.

Blogosphere – a set (collection) of all blogs on the Internet ; common name for the collection blogs.

Big Data – a series of approaches, tools and methods for processing data of large volumes and a significant variety to obtain human-readable results, effective in the conditions of their continuous growth, distribution over numerous nodes of a computer network . As defining characteristics for big data, the “three Vs” are noted: volume, in the sense of the size of the physical volume, speed in the sense of both the growth rate and the need for high-speed processing and obtaining results, diversity, in the sense of the possibility of simultaneous processing of various types of structured and semi-structured data.

Web analytics – measurement, collection, **analysis** , presentation and interpretation of information about *website visitors* in order to improve and optimize them. The main task of web analytics is _ *monitoring* of the operation of websites, on the basis of which a web audience is determined and the behavior of web visitors is studied to make decisions on the development and expansion of the functionality of a web resource.

Web space – a set of sites on the Internet (Internet hypertext space, www-space).

Website – a set of web pages that make up a single whole (dedicated to the same subject or belonging to the same author),

as a rule, hosted on the same server, having the same domain name and interconnected cross references. The HTTP protocol was developed for direct client access to websites on servers.

Web forum is a class of web applications for organizing communication between visitors, a website designed for online discussions.

Virtual dating service is an Internet service that provides Internet users with services for virtual communication with other users, an analogue of real dating services .

Visualization – to a set of methods for presenting the results of data analysis in the most convenient form for perception and interpretation. It can be used to monitor the process of building and operating various analytical models, testing hypotheses, and other purposes related to analysis.

The input degree of the node is the number of graph edges that enter the node.

The node's output degree is the number of graph edges that exit the node.

Geosocial network is a type of social network that uses geocoding. Users leave data about their location, which allows them to unite and coordinate their actions based on what people are present in certain places, or what events occur in these places.

Data Mining is a technology for analyzing data in databases or data warehouses, based on statistical methods and serving to identify previously unknown patterns, as well as to support the adoption of strategically important decisions.

Text mining is a technology for extracting information from text data based on the detection of patterns in them. As a rule, it includes the stages of structuring the source text (usually by syntactic analysis, adding some linguistic structures and deleting others, followed by inserting the results into the database), searching for patterns in the data, evaluating and interpreting the results.

Deep Web [Invisible Web, Deep Web, Hidden Web] is a part of the web space that is not indexed by search engine robots. Information, being inaccessible to search, is located "in depth" (eng. – Deep). Consists of web pages dynamically generated in response to online database queries.

Communication Graph – in social networks – a graph designed to identify connections between their participants. Graphs can be used to visualize these relationships. The graph of connections is built thanks to the exchange of content between people.

Digest – an information product (publication, article, selection) containing brief annotations and main provisions of articles or in which the content of the most interesting publications for a certain period is concisely transmitted.

Descriptor [(from Latin D escriptio – description)] – a lexical unit (word, phrase, code) of an information retrieval language that serves to express the main semantic content of documents (text). It is used for coordinate indexing of documents and information queries for the purpose of subsequent search.

Graph diameter – the maximum of the distances between pairs of its vertices. The distance between vertices is defined as the smallest number of edges that must be passed in order to get from one vertex to another.

System survivability – the ability of the system to perform the established minimum scope of its functions under external influences not provided for by the conditions of normal operation, to select the optimal mode of operation at the expense of its own internal resources, restructuring, changing the functions of individual subsystems and their behavior.

Knowledge extraction is the process of obtaining knowledge from data in the form of dependencies, rules, models. Stages: consolidation, purification, transformation, modeling and interpretation of the results.

Extraction (extraction) of information – a kind of information retrieval, in which some structured information is extracted from electronic documents, i.e. categorized, semantically meaningful data on a problem or issue.

Extraction of facts, concepts (Feature Extraction) is a technology that provides information in a structured form. Includes three main methods: Entity Extraction – extracting words or phrases that are important for describing the content of the text; Feature Association Extraction – identifying relationships between extracted concepts; Event and Fact Extraction – entity extraction, fact and event recognition.

Company image – a sustainable image that a company creates about itself through advertising, forming a favorable image among the target audience. This is a stable representation of consumers, clients, partners and the public about the prestige of the company, the quality of its products and services, the reputation of managers.

Simulation Modeling is a research method in which the system under study is replaced by a model that describes the real system with sufficient accuracy. This model is used for experiments in order to obtain information about the real system. Simulation is a special case of mathematical modeling. There is a class of objects for which, for various reasons, analytical models or methods for solving relative to the obtained model have not been developed. In this case, the mathematical model is replaced by a simulator or simulation model – a logical-mathematical description of the object.

Internet is a global information network, parts of which are logically connected by a single address space based on the TCP/IP protocol stack, their subsequent extensions, or other IP-compatible protocols. Provides, uses or makes available, publicly or privately, a high-level communication service. Consists of many interconnected computer networks.

Internet intelligence is a segment of competitive intelligence, covering the procedures for collecting and processing information carried out to support managerial decision-making, increase the competitiveness of commercial organizations exclusively from open sources from computer networks, most of which are built on top of the Internet.

Internet cleaners – specialists or services that can remove data from information resources on the Internet (usually negative customer information).

Information security is the state of information, information resources and information systems, in which information (data) is protected with the required probability from leakage, theft, loss, unauthorized destruction, distortion, modification (forgery), copying, blocking and etc. It has three main components: confidentiality, integrity and availability.

Information and analytical activities is a branch of human activity designed to meet the information needs of society with the help of analytical and information technologies by pro-

cessing input information and obtaining qualitatively new knowledge.

IAS [Information Analytical System] – class with information systems designed for analytical data processing, and not for automating the daily activities of the organization. Combines, analyzes and stores information extracted both from the organization's databases and from external sources. The data warehouses included in the IAS provide the transformation of large volumes of detailed data into generalized information suitable for decision making.

Information retrieval system, IPS is a system designed to provide search and display of documents presented in databases. The core of the IPS is a search engine – a software module that searches on request. IPS integrated with web technologies are the basis for building information retrieval web servers.

Informational influence – excitation (inhibition) in a controlled system of such processes that stimulate a choice desirable for the controlling party. This method of influencing the subject does not imply, for example, direct disabling of some of the elements of his system, but is the transfer of such information to him that will prompt him to choose a certain solution in which these elements will lose their effectiveness.

Information operations – informational impact on the mass consciousness (both hostile and friendly), impact on the information available to the competitor and necessary for him to make decisions, as well as on the information and analytical systems (IAS) of the competitor, including actions aimed at the physical defeat of the IAS, disable the means of computer and telecommunications infrastructure.

Informational resources – individual documents and individual arrays of documents, documents and arrays of documents in information systems, recorded on the appropriate media, as well as language tools used to describe a specific subject area and to access data and knowledge.

Information objects, IO – objects containing (carrying) information . They can be described directly or in the form of an algorithm for their generation.

Information portrait – a document that characterizes in a compact form the main content of the text – the objects described in it, persons, situations, etc.

Captcha is a fuzzy graphic representation of letters and numbers that you need to enter from the keyboard in a specific field.

Classification – a system for distributing objects into classes in accordance with a certain attribute (classification basis). Objects need to be classified to identify the general properties of the information object, which is determined by the information parameters (requisites). When classifying, the following requirements must be observed: completeness of coverage; uniqueness of details; the ability to include new objects.

Cluster Analysis – a multidimensional statistical procedure that collects data containing information about a sample of objects and then arranges the objects into relatively homogeneous groups (clusters).

Cliques are subgroups or clusters in which nodes are more strongly connected to each other than to members of other cliques.

Trade secret – information of a confidential nature from any sphere of activity of a state or private enterprise, the disclosure of which may cause material or moral damage to its owners or users (legal entities).

Competitive Intelligence – planned actions for the systematic collection and analysis of information, carried out in order to support managerial decision-making, increase the competitiveness of commercial organizations.

Competitive environment is the result and conditions of interaction of a large number of market entities. It is formed not only and not so much by the actual subjects of the market, the interaction of which causes rivalry, but, first of all, by the relations between them.

Consulting – the activity of specialized marketing companies that advise manufacturers, sellers, buyers on issues in the field of economics, management, marketing, pricing, product promotion, etc.

Consulting company (firm) – a company (firm) that provides consulting services for market research and forecasting, development of marketing programs, and finding ways out of crisis situations; and etc.

Consolidated information – obtained from several sources and integrated heterogeneous information resources (knowledge),

which together have the signs of completeness, integrity, consistency and constitute an adequate information model of the problem area in order to analyze its processing and use in processes decision support.

Content – content of information resources (eg websites) – texts, graphics, multimedia. Content parameters are its volume, relevance and relevance.

Content analysis is an analysis of the content of documents, which is aimed at measuring a number of qualitative and quantitative characteristics of the text and analyzing the dependencies between them.

Content monitoring is a systematic, time-continuous scanning and content analysis of information resources.

Confidential information – *information* that is a commercial or personal secret and is protected by its owner.

Confidentiality – the property of protecting information from unauthorized access and attempts to disclose it by users who do not have the appropriate authority.

Coefficient of clustering – value corresponding to the level of connectivity of nodes in the network. P shows how many nearest neighbors of a given node are closest neighbors for each other, and is equal to the ratio of the real number of edges that connect the nearest neighbors of the given node, to the maximum possible.

Intermediation coefficient [Betweenness] – parameter showing how much $\frac{1}{2}$ of the shortest paths passes through the node e . Points to the role of this node in establishing links in the network.

Centrality coefficient [Centrality] – a parameter that shows the "importance" or "influence" of a particular node (cluster of nodes) within the graph (network). Standard methods for measuring "centrality" cover the calculation mediation centrality, proximity centrality, eigenvector centrality, degree centrality, etc.

The decision maker is a subject (manager) endowed with certain powers and responsible for the consequences of the adopted and implemented management decision. Decision maker – one or more people (team) who are responsible for the decision.

Small world is one of the types of graphs in which most of the nodes are not pairwise neighboring, but can be connected to each other due to a small number of transitions along the edges

of the graph. A graph (network) is considered a small world if the distance between any two randomly selected nodes in most cases does not exceed the binary logarithm of the total number of nodes.

Mathematical modeling is the process of building and studying mathematical models – mathematical representations of reality.

Media activity – the activity of an individual in searching, receiving, consuming, transmitting, producing, disseminating information.

Metasearch engine – search engine, not having its own index, capable of transmitting user requests to several search servers simultaneously, selecting and selecting the most relevant results, and combining them And present to the user in the form of a document with links.

Multi-agent system, MAC – a system formed by several interacting intelligent agents. MACs can be used to solve problems that are difficult or impossible to solve with a monolithic system.

Unfair competition is a violation of generally accepted rules and norms of competition. Forceful and illegal methods of competition (depriving competitors of raw materials, sales markets, knocking down prices, industrial espionage, etc.).

Unstructured text information is a non-standardized and non-formalized text consisting of natural language sentences. With the content of the text – full-text presentation of ideas, meanings and plots (free text).

Knowledge discovery is a technique for extracting knowledge (information) from information sources .

Knowledge discovery in text is the process of discovering new, potentially useful and understandable patterns in unstructured textual data.

Knowledge discovery in databases [KDD] is the process of discovering useful knowledge in databases. This knowledge can be represented in the form of patterns, rules, forecasts, relationships between data elements, etc. The main KDD tool is Data Mining technologies.

Society of Analysts and Competitive Intelligence Professionals is a regional public organization of specialists in the field of competitive intelligence, created on the basis of the Depart-

ment of Social Informatics of the Kharkiv National University of Radio Electronics in 2002.

Ontology – formalization of a certain area of knowledge using a conceptual scheme, consisting of a data structure, their relationships and rules adopted in this area. Scope of application – business process modeling, Semantic Web, artificial intelligence.

Open sources – information sources that legally distribute information that can be accessed legally. Legality and legitimacy are considered only in the context of the jurisdiction of the territory in which economic or other operations are conducted.

Personal Information – any information relating to a natural person identified or determined on the basis of such information, including his surname, name, patronymic, year, month, date and place of birth, address, family, social, property status, education, profession, income, other information. Personal data is classified as confidential information; information or a set of information about an individual who is identified or can be specifically identified.

Peer-to-peer network [P2P] is a computer network based on the equality of participants. In such a network, there are no dedicated servers, and each node (peer) is both a client and a server. Unlike the client-server architecture, it allows you to keep the network working with any number and any combination of available nodes.

Search Engine Optimization [SEO] is a set of measures to improve the position of a website in the results of web search engines for certain user requests.

Building a semantic network is one of the main tasks solved in Text Mining – searching for key concepts of the text and establishing relationships between them, forming a structure for representing knowledge in the form of a directed graph, in which the vertices are concepts, and the arcs are relationships. Such a network can be contextually navigated.

Industrial counterintelligence – activities to prevent industrial espionage.

Industrial espionage is a form of unfair competition in which the illegal receipt, use, disclosure of information constituting a commercial, official or other legally protected secret is car-

ried out in order to obtain advantages in carrying out business activities, as well as obtaining material benefits.

Intelligence service is the practice and theory of gathering information about an adversary or competitor for security and military, political, or economic advantage. Intelligence can use both legal methods of collecting information and illegal operations that fall under the concept of "espionage".

Intelligence activity – an activity that includes collecting information, assessing its reliability and combining individual facts into a big picture.

Intelligence information – meaningful information based on the collected, evaluated and interpreted facts, obtained as a result of the selection, comparison, logical linking and generalization of intelligence data and information in accordance with the task of the consumer.

Intelligence cycle – within the framework of competitive intelligence – processes describing: target designation, collection, processing And analysis of intelligence information , bringing targeted information and conclusions to the customer.

Relevance – a measure of the conformity of the result obtained with the desired one. In information retrieval, a measure of the correspondence of search results to the task set in the search query.

Reputation is a social assessment of a group of subjects about a person, a group of people or a company, formed on the basis of some criteria.

Company reputation is a set of evaluative representations of the target audience about the company, formed on the basis of reputation factors that are important for this audience.

Retrospective Information – information contained in datasets accumulated over a significant period of time, or obtained as a result of searching in these arrays.

Retrospective analysis is an analysis that examines trends over a period of time in the past.

Risk is a situational characteristic of an activity, consisting in the uncertainty of its outcome and possible adverse consequences in case of failure.

Data collection – the process of identifying and obtaining data from various sources, grouping the received data and presenting them in the form necessary for entering into a computer.

Semantic Network [Semantic Network] is a way of representing knowledge in the form of a directed graph, in which the vertices correspond to the semantic units of the language (concepts, objects, actions, situations, etc.), and the edges correspond to properties or relationships between them.

Network analytics – a set of tools and methods for collecting from the network environment (in particular, from the Internet), transforming, storing, analyzing, modeling, delivering and tracing data, information and knowledge when working on tasks related to decision making.

Network Mobilization – the process of combining the efforts of social network participants to solve some problems, for example, organizing mass demonstrations, repelling aggression, helping the victims, etc. The possibilities of network mobilization depend on the structure of the network, its topology, parameters, the dynamics of information circulating in it, the possibility and probability of information perception by network nodes, the possibility of information transformation in network nodes, the possibility of restoring links in the network after a destructive impact on them.

Competitive intelligence system [CIS] – infrastructure to carry out competitive intelligence ; an integrated information and analytical system (IAS) for decision support in terms of analyzing changes in business conditions and political activities, on the basis of which a strategy and tactics of preventive measures are developed aimed at achieving competitive advantages and preventing the impact of negative factors in the economic and political environment.

Weakly social connections – a property of social networks, which consists in the presence of connections (edges with small weights) between nodes that are remote in some sense (for example, relationships with distant acquaintances and colleagues). If these connections are ignored, the network will fall apart into separate fragments. Weak ties are the phenomenon that links the social network into a single whole.

Mergers and acquisitions [M&A] – a class of economic processes of consolidation of business and capital, occurring at the macro- and microeconomic levels, as a result of which larger companies appear on the market instead of several less significant ones.

Community of Practitioners of Competitive Intelligence [CPCI] is a community of competitive intelligence practitioners in Russia that has existed since 2004. Initially, the SPKR was de facto created at the Business Intelligence Internet Forum, then it expanded, accepting new specialists from Russia, Ukraine, and Belarus. One of the priorities in the activities of the SPKR is the active propaganda and promotion of competitive intelligence in Russia and the CIS countries.

Social network is a social structure consisting of nodes (which are social objects) and links between them. Network objects can be enterprises, people, Internet resources, etc. There are many social networks that have their own specifics, their unique features, however, modern methods of data analysis are applicable to any of them, regardless of the specifics.

Social media is a set of online services and Internet applications that allow users to communicate with each other, including in real time. At the same time, users can exchange opinions, news, information, including multimedia among themselves. Social media is based on the ideological and technological basis of web 2.0, which allows the creation and exchange of content created by users themselves (User- Generated Content).

strategic business intelligence [Strategic BI] is intelligence that assists management in developing their overall plans and in testing the effectiveness of the vision process. Covers environment scanning, industry structure analysis, competitive analysis, scenario analysis (action plans), issue management, technology forecasting, development of competitive personality types, etc.

Scenario [Script] _ – process execution plan ; defines a sequence of commands that tells the program how and in what order to execute a particular procedure.

Scenario planning– planning of scenarios (scenarios).

Tactical business intelligence [Tactical BI] – intelligence that helps the company in its daily work, used by employees directly in their areas at the level of daily control. It covers the analysis of the needs of the buyer, the price of a competitor, the analysis of products and services of a competitor, the production of a competitor, etc.

Theory of complex networks is an interdisciplinary field of knowledge that has arisen on the basis of empirical research real

networks, especially computer and social. Within this theory complex network is a graph (net) with non-trivial topological features that do not occur in simple networks such as gratings or random graphs but often found in reality. _ The theory of complex networks studies the characteristics of complex networks, taking into account not only the topology of networks, but also statistical phenomena, the distribution of weights of individual vertices and edges, the effects of leakage and conduction in networks, etc.

Knowledge Management – the processes by which the essential elements of intellectual capital necessary for the success of an organization are created, stored, distributed and applied. There are five main technologies that support knowledge management: Business Intelligence, Collaboration, Knowledge Transfer, Knowledge Discovery, and Expertise Location.

Reputation management – methods of monitoring the reputation of a person or company, identifying facts that harm it, and using consumer feedback channels to react or early identify possible negative consequences for reputation.

Online reputation management [ORM] is one of the modern ways to manipulate Internet content (popularization of informative sites, writing press releases, articles and reviews) in order to create a positive or negative image of a company or person on the Internet. To implement ORM, specialists are required : copywriters, editors, SEO specialists, designers and programmers. The goal of specialists ORM – get an attractive image of a company or person, increase the profitability and efficiency of the enterprise.

Risk Management is the process of making and implementing managerial decisions aimed at reducing the likelihood of an unfavorable result and minimizing possible losses caused by its implementation. The purpose of risk management in the economic sphere is to increase the competitiveness of economic entities by protecting against the realization of risks.

Decision Management is a directive act of purposeful influence on the management object, based on the analysis of data characterizing a specific management situation, determining the goal of actions, and containing a program to achieve the goal.

Factographic Database – a database containing factual data – information related directly to the subject area.

Factual Information [Factographic Information] – description of facts grouped according to certain system-forming features.

Photo hosting is a website that allows you to publish images (such as digital photos) on the Internet. Photo hosting can be used to host, store and display images to other network users. The main advantage that photo hosting provides users is the convenience of displaying photos. The author can easily share a hyperlink leading to a photo with anyone with Internet access.

Digital shadow – information about the user, created without his participation, which arises and accumulates when someone searches for the user through search engines, e-mails are sent to the lists in which he appears and in many other cases. In addition to "open access digital shadows", "restricted access digital shadows" are created and accumulated – surveillance camera recordings, bank transactions, billing of online stores, ticketing services, phone calls, etc.

Digital footprint – information that is left by the user himself when working on the Web and by which you can not only identify him, but also "attach " him to certain actions, events, restore some fragments of his biography.

Expert evaluation is a quantitative or qualitative assessment of processes or phenomena that are not directly measurable based on the judgments of specialists.

Extractor – a program that collects data from source systems (highlighting complex elements and special constructions in the text that differ in a special type of spelling – names of legal entities, goods, addresses, numbers, etc.).

Node eccentricity – the largest of the geodesic distances (minimum distance between nodes) from a given network node to others.

Literature

[Astafieva, 1996] Astafieva N.M. Wavelet analysis: fundamentals of theory and examples of application // Uspekhi fizicheskikh nauk, 1996. – 166. – No 11. – P. 1145-1170.

[Bird, 2007] Byrd K. OSINT model // Computerra, 2007. – No. 22.

[Gorbulin, 2009] Gorbulin V.P., Dodonov O.G., Lande D.V. Informational operations and security of support: threats, pro-tides, modeling: monograph. – K.: Intertekhnologiya, 2009. – 164 p.

[Grigoriev, 2007] Grigoriev A.N., Lande D.V., Borodnikov S.A., Mazurkevich R.V., Patsyora V.N. InfoStream. Monitoring news from the Internet: technology, system, service: scientific and methodological manual. – Kyiv: Start-98, 2007. – 40 p.

[Gubanov, 2009] Gubanov D.A., Novikov D.A., Chkhartishvili A.G. Models of reputation and information management in social networks // Mathematical theory of games and its applications, 2009. – No. 2. – P. 14-37.

[Jilad, 2010] Gilad B. Competitive Intelligence. How to recognize external risks and manage the situation – St. Petersburg: Peter, 2010. – 320 p.

[Dodonov, 2009] Dodonov O.G., Lande D.V., Putyatin V.G. Information flows in global computer networks. – K: Nauk. Dumka, 2009. – 295 p.

[Dodonov, 2010] Dodonov A.G., Lande D.V. Vitality of information plots // Proceedings of the XI International Scientific and Practical Conference "Information Security". – Part 2. – Taganrog: Publishing House of TTI UFI, 2010. – P. 179-183.

[Dodonov, 2011] Dodonov A.G., Lande D.V. Vitality of information systems. – K.: Nauk. Dumka, 2011. – 256 p.

[Dodonov, 2013] Dodonov A.G., Lande D.V., Kozhenevsky S.R., Putyatin V.G. Computer information-analytical systems and data storages. Dictionary. – K.: Phoenix; IPRI NAS of Ukraine, 2013. – 554 p.

[Doronin, 2011] Doronin A. Business intelligence. – M.: Os-89, 2003. – 704 p.

[Dudikhin, 2004] Dudikhin V.V., Dudikhina O.V. Competitive intelligence on the Internet. – M.: AST, NT Press, 2004. – 240 p.

[Ermakov, 2005] Ermakov N.S., Ivashchenko A.A., Novikov D.A. Models of reputation and performance standards. M.: IPU RAN, 2005. – 67 p.

[Ivashchenko, 2006] Fundamentals of the methodology for investigating illegal collection and disclosing commercial secrets // Legal Journal, 2006. – No. 8. – P. 48-66.

[Ivlev v, 2008] Ivlev A.A. Fundamentals of Boyd's Theory. Directions of development, application and implementation (monograph). – M., 2008. – 64 p.

[Kalinovskiy, 2012] Kalinovskiy Ya.A., Boyarinova Yu.E. High-dimensional isomorphic hypercomplex numerical systems and their use to increase the efficiency of computations. -K.: Infodruk, 2012. – 183 p.

[Kiselev, 2005] Kiselev S. Business Intelligence Information System Model // Open Systems, 2005. – No. 5-6. – S. 60-66.

[Kovalchuk, 2012] Kovalchuk A. Practice and secrets of making money on the Internet. Reputation management // Issue 30, 2012 (on – line : <http://www.trustlink.ru/subscribe/show/35>) –

--

[Kondratiev, 2010] Kondratiev A. Intelligence using open sources of information in the USA // Foreign military review, 2010. – No. 9. – S. 28-32.

[Kononov, 2003] Kononov D.A., Kulba V.V., Shubin A.N. Basic concepts of modeling information management in social systems // Proceedings of the international scientific-practical conference "Theory of Active Systems". – M.: Institute of Management Problems. V.A. Trapeznikova RAN, 2003. -T 2. – P. 125-129.

[Kochergov, 2009] Kochergov D. One step that may be the last // Business Economics, 2009. – No. 13 (9279).

[Kuznetsov, 2006] Kuznetsov S.V. How to conduct business intelligence in the "invisible" Internet? // "CNews", 07.09.06.

[Kulba, 1999]: Kulba V.V., Malyugin V.D., Shubin A.N., Vus M.A. Introduction to information management. Educational and methodical edition. – St. Petersburg: Publishing House of St. Petersburg University. 1999. – 116 p.

[Kulba, 2004] Kulba V.V., Kononov D.A., Kosyachenko S.A., Shubin A.N. Methods of formation of development scenarios from socio-economic systems. – M.: SINTEG, 2004. – 296 p.

[Lande, 2005] Lande D.V. Search for knowledge on the Internet. Professional work. – M.: Dialectics, 2005. – 272 p.

[Lande, 2007] Lande D.V., Snarsky A.A., Braichevsky S.M., Darmokhval A.T. Modeling the dynamics of news text flows // Internet Mathematics 2007: Collection of works of the contest participants. – Yekaterinburg: Ural Publishing House. un-ta, 2007. – S. 98-107.

[Lande, 2009] Lande D.V., Snarsky A.A., Bezsudnov I.V. Internet: Navigation in complex networks: models and algorithms. – M.: Librokom (Editorial URSS), 2009. – 264 p.

[Lande, 2010] Lande D.V. Deep web – information environment for business analyst // Information technologies for management, 2010. – No. 9. – C. 28-32.

[Lande, 2013] Lande D.V. Method of visualization of instability zones in measurement series // Information technologies and safety. Status assessment: Proceedings of the international scientific conference ITB-2013. – K.: IPRI NAS of Ukraine, 2013. – S. 105-113.

[Lande, Braichevsky, 2010] Lande D.V., Braichevsky S.M., Darmokhval A.T., Zhigalo V.V. Architecture of the system for covering information links of monitoring objects // Computational Linguistics and Intelligent Technologies: Based on the materials of the annual International Conference "Dialogue". – Issue. 9 (16). – M.: Publishing house of the Russian State Humanitarian University, 2010. – C. 272-278.

[Lande, Prishchepa, 2007] Lande D., Prishchepa V. School of web intelligence. Tools and sources // Telecom, 2007. – No. 7-8. – S. 46-49.

[Lande, Furashev, 2012] Lande D.V., Furashev V.M. Fundamentals of information and social and legal modeling: monograph. – K.: PanTot, 2012. – 144 p.

[Nezhdanov, 2010] Nezhdanov I. Intelligence technologies for business. – M.: Os-89, 2009. – 400 p.

[Novikov, 2002] Novikov D.A., Chkhartishvili A.G. Theory of management of organizational systems – M.: Sinteg, 2002. – 227 p.

[Novikov, 2007] Novikov D.A. Theory of management of organizational systems. 2nd ed. – M.: Fismalit, 2007. – 584 With.

[Pechenkin, 2004] Pechenkin I.A. Information technologies in the service of intelligence // *Confident*, 2004. – No. 4. – C. 28-41.

[Prescott, 2003] Prescott John E., Miller Stephen H. *Competitive Intelligence: Lessons from the Trenches*. -M.: Alpina Publisher, 2003. – 336 p.

[Rastorguev, 2006] Rastorguev S.P. Information war. Problems and models. *Existential mathematics*. – M.: Helios ARV, 2006. – 240 p.

[Robinson, 2016] Robinson Jan, Weber Jim, Eifrem Emil. *Graph databases: new opportunities for working with related data*. – M.: DMK Press, 2016. – 256 p.

[Khan, 2000] Khan U., Mani I. *Automatic Referencing Systems* // *Open Systems*, 2000. – No. 12.

[Khoroshevsky, 2013] Khoroshevsky V.F. *Semantic Technologies: Expectations and Trends* // *Open Semantic Technologies for Intelligent Systems Design – Open Semantic Technologies for Intelligent Systems (OSTIS-2012): Proceedings of the II Intern. scientific-technical conf. (Minsk, February 16-18 2012)*. – Minsk: BGUIR, 2012. – S. 143-158.

[Chernykh, 2013] E. Chernykh. We can't hide from the NSA, we can't hide // *Komsomolskaya Pravda*, June 24, 2013.

[Chertov, 2009] Chertov O.R. *Polynomials of Kunchenko for recognition of images* // *Bulletin of NTUU "KPI" Informatics, management and calculation technology*, 2009. – No. 50. – P. 105-110.

[Chkhartishvili, 2004] Chkhartishvili A.G. *Game-theoretic models of information management*. M.: CJSC " PMSOFT ", 2004. – 227 p.

[Bak, 1996] Bak P. *How nature works: The science of self-organized criticality*. – New York: Springer-Verlag Inc., 1996. – 212 p.

[Bhargava, 1993] Bhargava SC, Kumar A., Mukherjee A. *A stochastic cellular automata model of innovation diffusion* // *Technological forecasting and social change*, 1993. – 44. – No. 1. – P. 87-97.

[Bjorneborn, 2004] Bjorneborn L., Ingwersen P. *Toward a basic framework for webometrics*. *Journal of the American Socie-*

ty for Information Science and Technology, 2004. -55(14): 1216-1227.

[Boyd, 2012] Boyd, D.; Crawford, K. (2012). "Critical Questions for Big Data". *Information, Communication & Society*. 15(5): 662-679. doi : 10.1080/1369118X.2012.67887

[Buckheit, 1995] Buckheit J., Donoho D. Wavelab and reproducible research // Stanford University Technical Report 474: Wavelets and Statistics Lecture Notes, 1995. – 27 p.

[Burbary, 2009] Burbary K., Cohen A. A Wiki of Social Media Monitoring Solutions // (on-line: <http://wiki.kenburbary.com/>)

[Burke, 2001] Burke MM Knowledge Operations: above and beyond Information Operations. 6th International Command and Control Research and Technology, June 19 – 21, 2001. – 16 p.

[Clauset, 2008] Clauset, A., Moore, C., Newman, MEJ Hierarchical structure and the prediction of missing links in networks // *Nature*, 2008. – 453, 98-101.

[Dean, 2004] Jeffrey Dean, Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", December 2004, <http://labs.google.com/papers/mapreduce.html>.

[DoD, 2003] Information operations roadmap – DoD US, 30 October 2003. – 78 p.

[Erdős, 1960] Erdős P., Rényi A. On the evolution of random graphs, *Publ. Math. Inst. hungar. Acad. sci.* 5, 1960. – P. 17-61.

[Ghemawat, 2003] Sanjay Ghemawat, Howard Gobioff, Shun-Tak Leung, "The Google File System", October 2003, <http://labs.google.com/papers/gfs.html>.

[He, 2007] He B., Patel M., Zhang Z., Chang KC-C. Accessing the Deep Web: A Survey // *Communications of the ACM (CACM)*, 50(5):94-101, 2007.

[Hill, 2000] Hill JMD, Surdu JR, Ragsdale DJ, Schafer, JH Anticipatory planning in information operations // *Systems, Man, and Cybernetics*, 2000 IEEE International Conference, 2000. – 4. – P. 2350-2355.

[Kacperski, 2000] Kacperski K., Holyst JA *Physica A*. Phase transitions as a persistent feature of groups with leaders in models of opinion formation // *Statistical Mechanics and its Applications*, 2000. – 287, Issues 3-4. – P 631-643.

[Knight, 2003] Knight JC, Strunk EA, Sullivan KJ Towards a Rigorous Definition of Information System Survivability // Pro-

ceedings of the DARPA Information Survivability Conference and Exposition (DISCEX'03), 2003.

[Lande, 2012] Lande DV, Kalinovskiy Ya.A., Boyarinova Yu. E. The model of information retrieval based on the theory of hypercomplex numerical systems // Preprint Arxiv v 1205.3031. (on-line: <http://arxiv.org/abs/1205.3031>).

[Lande, 2019] Dmytro Lande, Ellina Shnurko-Tabakova. OSINT as a part of cyber defense system // Theoretical and Applied Cybersecurity, 2019. – N. 1. – pp. 103-108.

[Lasswell, 1948] HD The structure and function of communication in society // The Communication of Ideas. / Ed.: L. Bryson. – New York: Harper and Brothers, 1948.

[Latane, 1981] Latane B. The psychology of social impact // American Psychologist, 1981. – 33. – P. 343-356.

[Latane, 1997] Latane B., Nowak A. Causes of polarization and clustering in social groups // Progress in communication sciences, 1997. – 13. – P. 43-75.

[Lewenstein, 1993] Lewenstein M., Nowak A., Latane B. Statistical mechanics of social impact // Physical Review, 1993. – A, 45. – P. 763-776.

[Li, 2012] Li Y., Miller EL, Long DDE Understanding Data Survivability in Archival Storage Systems // Proceedings of the 5th Annual International Systems and Storage Conference (SYSTOR 2012), June 4–6, 2012, Haifa, Israel.

[Milgram, 1967] Milgram S. The small world problem, Psychology Today, 1967.-2.-P. 60-67.

[Newman, 2003] Newman MEJ The structure and function of complex networks // SIAM Review, 2003. – 45. – P. 167-256.

[Nowak, 1990] Nowak A., Szamrej J., Latane B. From private attitude to public opinion: A dynamic theory of social impact // Psychological Review, 1990. – 97. – P. 367-376.

[Osgood, 1954] Osgood Ch. E. Psycholinguistics. A Survey of Theory and Research Problems // Supplement to the International Journal of American Linguistics. Vol. 20. No. 4. Oct. 1954, mem. 10. Baltimore: Waverly Press, 1954.

[Price, 2001] Price G., Sherman C., Sullivan D. The Invisible Web: Uncovering Information Sources Search Engines Can't See. – Information Today, Inc., 2001. – 439 p.

[Roberts, 2002] Roberts PW, Dowling GR Corporate reputation and sustained superior financial performance // Strategic Management Journal, 2002. – 23. – No. 12. – P. 1077–1093.

[Shvachko, 2010] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler. The Hadoop Distributed File System. Proceedings of MSST2010, May 2010.

[Scaling, 2008] "Scaling Hadoop to 4000 nodes at Yahoo!", http://developer.yahoo.net/blogs/hadoop/2008/09/scaling_hadoop_to_4000_nodes_a.html.

[Sobkowicz, 2003] Sobkowicz P. Effect of leader's strategy on opinion formation in networked societies // Preprint Arxiv (online: <http://arxiv.org/pdf/cond-mat/0311566>)

[Schramm, 1974] Schramm W., D.F.Roberts (eds.) The Process and Effects of Mass Communication. Univ. of Illinois Press, 1974.

[Watts, 1998] Watts DJ, Strogatz SH Collective dynamics of "small-world " networks. // Nature, 1998. – 393. – P. 440-442.

[Yahoo, 2008] "Yahoo! Launches World's Largest Hadoop Production Application, February 19, 2008, <http://developer.yahoo.net/blogs/hadoop/2008/02/yahoo-worlds-largest-production-hadoop.html>.

Competitive intelligence websites

1. International Society of Competitive Intelligence Professionals SCIP (www.scip.org)
2. Competitive Intelligence Academy Fuld-Gilad-Herring, Cambridge (www.academyci.com)
3. Community of Practice of Competitive Intelligence, SPKR (razvedka-open.ru)
4. Russian Society of Competitive Intelligence Professionals, ROPKR (www.rscip.ru)
5. Competitive Intelligence Institute, Germany (www.institute-for-competitive-intelligence.com/start.html)
6. Business – school Skema, France (<http://www.skema-bs.fr/faculte-recherche/centre-intelligence-economique-et-influence>)
7. Canada- headquartered association of competitive intelligence professionals Competia (www.competia.com)
8. Kharkiv Regional Public Organization "Society of Analysts and Competitive Intelligence Professionals" (www.scip.org.ua)
9. "Knowledge Camp & Competitive Intelligence Camp" – Ukrainian BarCamp on competitive intelligence and knowledge management (barcamp2010.scip.org.ua)
10. R-Techno Private Intelligence Company (www.r-techno.com)
11. Competitive Intelligence Agency "Informant" (www.informnn.ru)
12. Intelligence technologies for IT2B business (www.it2b.ru)
13. Alt Marketing: Competitive Intelligence Library (alt-marketing.ru/articles/index-competitiveintelligence.shtml)
14. Corporate Information Service (z-filez.info)
15. Competitive Intelligence, E.L. Yushchuk (ci-razvedka.ru)
16. Competitive intelligence on the Internet. Author's course of A. Masalovich (<http://www.tora-centre.ru/razvedka.htm>)

Addresses of mentioned web resources

www.aignes.com – WebSite-Watcher – a program for monitoring websites, forums, local files.

www.anbr.ru – information and analytical system "Semantic Archive".

www.archive.org – Internet – archive (Internet Archive).

archive-it.org is Bielefeld's academic search engine Biefield BASE is one of the world's largest search engines for academic web resources.

attackindex.com is a system for monitoring information operations.

www.babkee.ru – Babkee – a system for monitoring mentions in social media.

www.base.ukrpatent.org/searchINV – interactive database "Inventions (utility models) in Ukraine".

blog.capterra.com/top-8-free-and-open-source-business-intelligence-software/ – Overview funds business intelligence. _

books.google.com – Google Book Search – search books.

brigh-tplanet.com – American company BrightPlanet, one of the first to publish a report on the "deep web".

www.ciradar.com/Competitive-Analysis.aspx – CIRadar – Competitive Intelligence Retrieval Systems on the **Deep** Web.

[citeseerx.ist.psu.edu / index](http://citeseerx.ist.psu.edu/index) – CiteSeerX is an electronic scientific literature library and search engine.

www.data.gov is a US government website that provides access to open government data.

code.google.com – Google Code Search – search for program code.

www.cy-pr.com – site analysis service.

www.digitalpreservation.gov is an American national project for the preservation and distribution of digital content Digital Preservation.

www.dtsearch.com – dtSearch is a search program that allows you to process static and dynamic data in all MS Office formats.

www.eapo.org – Eurasian Patent Database.

www.elastic.co is an Elastic company and developer of three related projects – the Elasticsearch search engine, the Logstash data collection and analysis engine, and the Kibana analytics and visualization platform.

www.europages.eu – Europages – European Business Directorate.

facebook.com – Facebook is the largest social network.

www.fmsasg.com is Sentinel Vizualizer, a visualization program for connections and relationships.

global.factiva.com is a division of Dow Jones Inc. dedicated to providing access to business and analytical information through its information and analytical services.

www.findlaw.com – FindLaw is a directory containing a list of freely available databases of legal documents.

gephi.org is a network and graph visualization and analysis program.

google.com is Google's global information retrieval system.

google.com/alerts – Google alerts.

hootsuite.com – Hootsuite is a multifunctional social media service.

hrazvedka.ru is a blog about intelligence technologies in business.

www.ibm.com/products/i2-analysts-notebook is a visual data structure design system for storing data about various individuals and organizations.

www.infongen.com – InfoNgen is an information aggregator customizable to unique topics.

www.i-teco.ru – X-Files dossier management system.

infostream.ua – InfoStream – a system for monitoring web resources.

www.integrum.ru is the largest archive database of Integrum mass media.

www.internetsec.com is an Internet Securities service that provides business information from over 16,000 sources.

inventionmachine.com – from the Goldfire Research system – a deep web content processing system.

www.iqbuzz.ru – IQBuzz is a service for monitoring social media with the ability to connect new sources at the request of users.

www.kodeks.ru is an information retrieval system for Russian legislation.

www.kribrum.ru – Kribrum is a technology that allows you to track and analyze mentions of brands, products, services, etc.

www.labyrinth.ru is the Russian database "Labyrinth", compiled on the basis of publications of leading business publications.

www.lexisnexis.com is the world's largest full-text online information system LexisNexis.

www.linkedin.com – LinkedIn is a social network for finding and establishing business contacts.

www.livejournal.com – LiveJournal, LJ, LiveJournal, LJ – a platform for keeping online diaries (blogs).

www.loc.gov – US Library of Congress.

medium.com is a platform for social journalism.

www.megaputer.ru – PolyAnalyst – a family of products for data mining.

www.mlg.ru – "Medialogy" – a service that provides online access to the media database with the ability to independently monitor and express media analysis.

mednar.com/mednar/desktop/en/search.html is a free deep web search engine focused on medicine.

modusbi.ru – Modus BI is a platform for business intelligence that allows you to collect and visualize data from various sources, generate reports and create forecasts.

neticle.com/textanalysisapi/ – Neticle Text Analysis is a technology for extracting information from unstructured texts.

newspapermap.com – Newspaper Map – a service that combines geolocation and an information retrieval system for media resources.

www.nts.gov – aviation base data NTSB Aviation Accident Database.

www.newprosoft.com – Newprosoft Web Content Extractor – a program for scanning and extracting data from websites.

www.oracle.com/business-analytics/cloud.html – and an integrated set of analytical tools from Oracle.

patents.google.com is a Google search engine that indexes patents and patent applications.

www.peerindex.net – PeerIndex is a social media analysis service that determines the size of a company's influence.

photoinvestigator.co is a service for extracting metadata and other information from photographs.

www.politicalinformation.com is a search service for 5,000 selected political websites.

www.postrank.com – PostRank is a global social media analysis system.

www.rco.ru – RCO – a system for identifying factual information from unstructured texts.

www.rocketsoftware.com/products/rocket-folionxt is a program that allows you to identify entities, their mutual relationships and events in unstructured texts.

www.sap.com/sapbus_inessobjects – Sap Businessobjects Text Analysis is a program that allows you to extract information about dozens of types of objects and events.

scholar.google.com – Google Scholar – search for scientific publications.

www.scip.org.ua – The Society of Analysts and Competitive Intelligence Professionals.

screen-scrap.com is a program that allows you to automatically extract all information from web pages, download the vast majority of file formats, and automatically enter data into various forms.

www.semanticforce.net – SemanticForce is a monitoring service for unstructured information sources.

www.socialmention.com – Socialmention is a platform for searching and analyzing information in social networks.

www.softpedia.com – Website-Finder, a program that searches for websites that are poorly indexed by Google.

sphinxsearch.com is a full text search engine for big data.

telegram.org is a cross-platform messenger that allows you to exchange messages and media files in many formats.

www.socialbakers.com – a unified marketing platform based on social media analysis.

www.tora-centre.ru/avl3.htm – Avalanche – Internet monitoring and competitive intelligence system.

www.trackur.com – Trackur is a social media monitoring and analysis tool. Allows you to track, for example, the reputation of brands.

trends.google.com – Trackur is a Google tool that shows how often a certain term is searched for in relation to the total volume of search queries.

twitter.com – Twitter is the largest microblogging service.

ukrpatent.org – Ukrainian Institute of Intellectual Property.

visual.ly is a search engine for infographics on the web.

watchthatpage.com is a free service that allows you to automatically collect new information from monitored web resources.

webground.su - Webground is an integrator of Russian-language news.

weibo.com – Sina Weibo (Chinese: 新浪微博) is a microblogging service launched by Sina Corp.

www.worldindustrialreporter.com / solusource – Global supplier Directory by Solusource is a competitive intelligence web interface from Thomas.

worldwidescience.org – World Wide Science – a global scientific portal consisting of scientific databases and portals.

www.yellowpages.kiev.ua - "Yellow Pages" of Kiev.

www.yahoo.com is Yahoo!'s global information retrieval system.

www.yandex.ru is the global information retrieval system Yandex.

www.youscan.io – YouScan is a professional social media monitoring system.

youtube.com – the largest video hosting, providing users with video storage, delivery and display services.

www.youtube.com / playlist ? list = PL -9 OTQQwXf 2 XuDGO _ ElwUOpzUXLDDfcL – OSINT Academy – training course – YouTube

zakon.rada.gov.ua is an information retrieval system for Ukrainian legislation.