IEEE Second International Conference on Data Stream Mining & Processing
August 21-25, 2018, Lviv, Ukraine

Data Science in Open-Access Research On-line Resources

Dmytro Lande
Institute for information recording
National academy of science of Ukraine
Kyiv, Ukraine
dwlande@gmail.com

Valentyna Andrushchenko
State fund for fundamental research of Ukraine
Institute for information recording
National academy of science of Ukraine
valentyna.andrushchenko@gmail.com

Iryna Balagura
Institute for information recording
National academy of science of Ukraine
Kyiv, Ukraine
balaguraira@gmail.com

***Abstract*— The data science methods are widely used in different areas of nowadays life. This paper is dedicated to forming of new approaches, in particular – development of unique model, to provide scientometric research of abstracts from open source pre-print service to process data on subject domains and directions. The objective of research is to analyze degree of presence and importance of data science in different research fields. A new way of working with the information system of the library of the University of Cornelius - the resource of the pre-prints arXiv is proposed in the work. The authors reviewed the abstract information of the resource, which is the result of the search for relevant publications for the given concept. The main attention of the authors was focused on the distribution of publications in the identified scientific areas and the relevant sub-groups provided by the resource. The result of the work is a visual representation and interpretations of the network of subject areas for the concepts - big data, neural networks, deep learning.**

Keywords—scientometric, big data, data science, concept, subject domain, network, scientific papers.

I. INTRODUCTION

Today the problem of recording, processing and storage of information is actual for every field [1]. Big Data has become important for organization everyday wellbeing and using satellite data for forecasting weather, traffic jams, nature disasters [2]. Data science and Big data influence business and sales. Big data could be used for politic companies and for prediction of the stock fluctuating of a certain company [3, 4]. And even farming processes transformed into Smart Farming with machines that are equipped with smart sensors and devices and produce big amounts of data that provide unprecedented decision-making capabilities [5]. It attracts more consumers focused on innovations in goods production and services. Now we have possibilities to use smart transport without drivers, smart houses, Internet of things and Cloud Computing and fill more comfortable with data science development [6]. Data scince is growing but still contain challenges: Data challenges (e.g. data volume, variety, velocity, veracity, volatility, quality, discovery and dogmatism); process challenges; management challenges (privacy, security, governance and ethical aspects) [7]. The Big Data Analytics requires new advanced algorithms such as text analytics, machine learning, predictive analytics, data mining, statistics, and natural language processing [1]. There are no statistical research of data science usage and real state of big data evolving in science [7]. This paper is dedicated to scientometric research of abstracts from open source pre-print service in main fields detection of data science usage. The objective of research is to analyze degree of presence and importance of data science in different research fields. The source of data is open-access on-line recourse arxiv (www.arxiv.org). It includes 1,372,745 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance, Statistics, Electrical Engineering and Systems Science, and Economics. The source allows to design visualization of subfields links for selected concepts with maps using. The concept map of subject domains is an useful instrument for identification of related topics in scientific research, detection of trends in research, definition of its terminology, correct usage and correct application of keywords in scientific works.

II. METHODOLOGY

We propose to visualize data science integration to different research with networks theory, which was created with Euler (1736) the Konigsberg Bridge problem and presented as mathematical notation of nodes and edges [8]. In scientometrics network theory is widely used for mapping of science, among them: co-citation, co-author and co-word networks. Co-word and co-author networks could be used for identification and description of scientific groups and research topics, the most communicative researchers and main principles of science communication [9]. We propose to use maps for fields and subfields connection according to certain concepts which is necessary in datascience.

A. Data input

To provide the correct analysis of obtained information from the point of view of completeness and variety of research fields and direction the open access archive of preprint arXiv was chosen. ArXiv is the largest archive of electronic publications and their open preprints.

The archive was created in 1991. Initially all the publication on archive were allocated in frames of one subject domain - "Physics", but today the resource presumes arrangement of publications within other directions.

ArXiv is an information tool for hundreds and thousands of scholars. Among the users are more than 50 Nobel laureates, winners of prestigious scientific awards. Resource is an actual tool for users from countries with limited access to scientific information. Today, the archive contains 8 sections, where you can post your own materials: Computer Science (42 areas), Economics (1 direction), Electrical Engineering and System Science (3 courses), Mathematics (32 directions), Physics (13 units), Quantitative Biology (10

directions), Quantitative Finance (9 destinations), Statistics (6 destinations).

We will use abstract information which contains the following data:

- Number of article, identifier in the system, in the form:
  arXiv: XXXX.XXXXX [***], where the HTML code is the publication number in the system, *** - the list of available file formats for download;
- Topic of the publication;
- Author (s);
- Comments - contains information about the number, pages of publication, number of drawings and other items (not urgent);
- Journal-ref – contains information about the paper (available for publications that have already been published);
- Subject - the subject area or specific information on the scientific direction within the scope of the subject area (according to how the author of the publication noted during the presentation of the publication to place it on the resource).

B. Algorithm

Under the concept we will understand the meaningful verbal unit, or a combination of units, which defines the framework of scientific perception of the meaning of a particular notion that is appropriate to one or more subject areas [10]. The network of subject areas is a way of presenting a model of subject areas by defining generalized descriptions of the domain, represented by their proper name and the names of subordinate units of it, the scientific directions that more specifically describe the subject area defined by the information system on the basis of which the given network is constructed or on the basis of the proposed systematization of subject areas [11]. The search is provided for the concept which can be represented as a word or word combination through the array of resource publications. The algorithm for constructing a subject areas network for a given concept involves the definition of subject areas and scientific directions for which the given concept is appropriate. The implementation of the algorithm is realized by processing search results. We use abstract of paper, key-words and sub-fields. So, we will define nodes as fields and sub-fields and edges as co-occurrence of sub-fields in one paper (Fig.1).

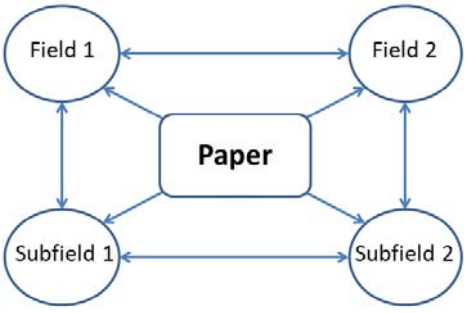

Fig. 1. Example of concept network building and fields connections

The algorithm of subject domain network building consist of next stages:

- The concept definition for the search.
- Extraction of abstract information.
- The scientific direction detection (sub-field) and subject domain (field), which is indicated in the abstract information.
- The scientific direction is the next node graph and connection with appropriate subject domain.
- If a scientific direction node has already been constructed - the name of the subject area, then only the node is constructed - the name of the scientific direction, which is connected with the node - the corresponding subject area.
- If a corresponding node has already been built for the name of the scientific direction, then the transition to upper steps. If the name of the scientific direction nodes has not yet been made, then upper steps (Fig.2)
- If there is no suched results, the network is considered to be built.

We use networks characteristics for networks analysis. Number of nodes, edges and density of network could be applied for detection of widely used terms in different fields. The density of network is the ratio of existing links to the total number of possible links. For a network of N nodes, the network link density is

$$\rho = \frac{2e}{n(n-1)} \quad (1)$$

where e – number of edges, n – number of nodes.

The (maximal) link density of a completely connected network is 1. We will admit that the lower the density of the network - the higher the polythematism.

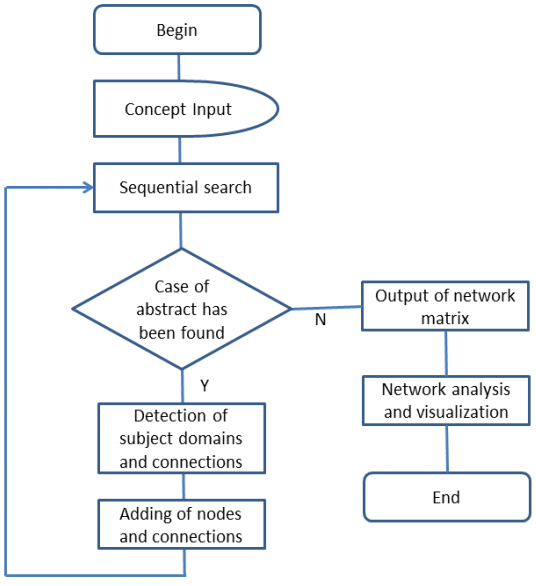

Fig 2. The algorithm of subject domain network building

III. MAIN RESULTS

Developed software and the algorithm which was proposed above we built the networks for the concepts: big data, neural networks, deep learning, and complex network. Main parameters of the networks are shown in table 1. The largest number of nodes and smallest density among proposed networks are for concept "neural network". It confirms the prevalence of the method in different fields. The visualizations were provided with the Gephi software (gephi.org). The concept "big data" refers mostly to computer science (CS) and statistics (Stat) fields, few articles in mathematics (math) and physics (fig.3.). Main connected sub-fields: Mashine learning (CS.LG); Distributed, Parallel, and Cluster Computing (CS.DC); Data Structures and Algorithms (CS.DS); Computational Engineering, Finance, and Science (CS.CE); Social and Information networks (CS.SI);Artificial Intelligence (CS.AI); Information retrieval (CS.IR);Networking and Internet architecture (CS.NI); Computation in statistics (Stat.CO); Methodology in statistics (Stat.ME) and others. The amount of sub-field is not low – 35. But it was expected to observe "big data" concept applied to such a wide spread fields as biology, finance, astronomy research. We can assume the reasons of absence of such research directions could be caused by author key-words missing, actively development of data science theoretical and fundamental laws, insufficiency of data base or practical industry usage etc.

The concept "neural networks" is represented in different sub-fields among them are: computer science (CS); statistics (Stat); Physics; Mathematics (Math); Quantitative Biology (q-bio); Electrical Engineering and Systems Science (eess); Quantitative Finance (q-fin); Econometrics (Econ.EM). So we can draw a clear conclusion that concept "neural networks" is widely used instrument in different fields. "Neural networks" and "deep learning" currently provide the best solutions in image recognition, speech recognition, and natural language processing. We determined the range of scientific directions related to the given concept by scanning the largest resource in the global network of preprints, containing a large amount of publications both prepared for printing and placed in the leading scientific publications. Developed applications by the proposed algorithm will allow using the network of subject areas as an additional tool for finding collaborators, expanding the use of the concept within different scientific areas and thus obtaining the opportunity for expanding collaborations and attracting specialists from various scientific fields.

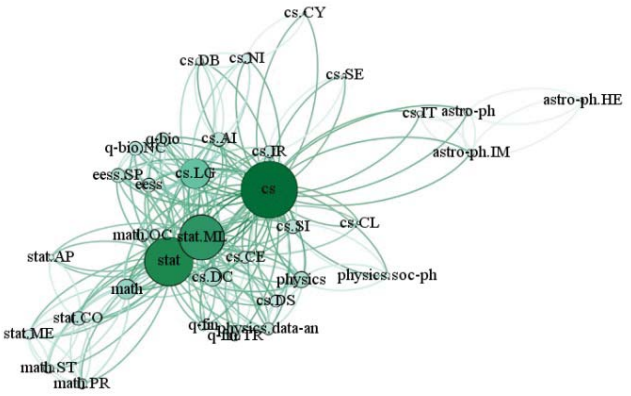

Fig. 3. Subject domain network for Big data concept

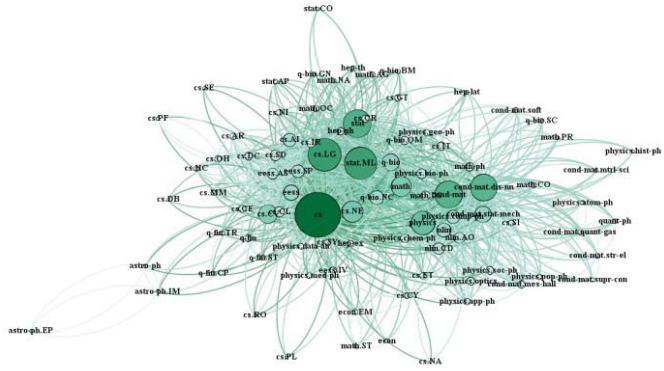

Fig. 4. Subject domain network for Neural network

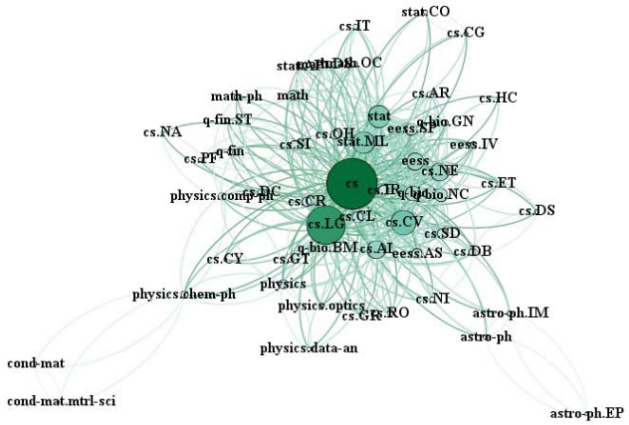

Fig. 5. Subject domain network for Deep learning

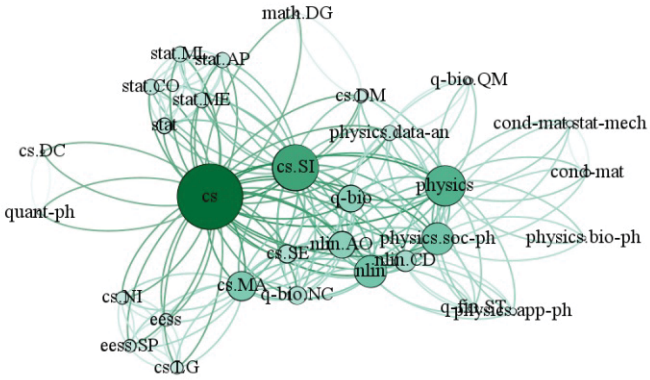

Fig. 6. Subject domain network for Complex networks

TABLE I. MAIN PARAMETERS OF THE SUBJECT DOMAIN NETWORKS

Concept	Density of nodes	Number of nodes	Number of edges
Neural network	0.136	90	1090
Deep learning	0.159	54	454
Big data	0.213	35	254
Complex network	0.234	31	218

For the development of the proposed approaches for the search, processing and interpretation of scientific information through the implementation of these algorithms, it is possible to construct a more developed network by grouping the names of scientific areas within dictionaries, as well as calculating network parameters.

IV. Conclusions

The algorithms, which based on subject domain mapping for certain concept are proposed. In maps we used nodes as fields and sub-fields and edges as co-occurrence of sub-fields in one paper. We offer to use the concept map of subject domains for identification of related topics in scientific research, detection of trends in research, searching for the ambiguity of terminology, correct usage of terms and describing science structure. Provided algorithm isn't strongly connected to the field of research or source of data and could be continued with other examples.

This paper is dedicated to forming of new approaches, such as new unique models for providing the scientometric research based on the open access archive of preprint to isolate and process the data connected to the subject domains of the publications and appropriate research directions. The objective of research is to analyze the level of representativeness and importance of data science in different research fields. We showed that the data science integration to different research with example of concepts "big data", "deep learning", "neural networks", "complex networks" using one of the biggest open access archive. We used density of complex network for estimation of the widest in the sense of concepts usage. Main subject area for selected concepts is computer science. The most common concept is "neural networks" which is used in 90 different sub-fields, other concepts mostly used in computer science. The theory of complex networks decreased inherency in comparing with neural networks and contains in 31 research fields.

References

[1] M.M. Najafabadi, F. Villanustre, T.M. Khoshgoftaar et al., "Deep learning applications and challenges in big data analytics," Journal of Big Data, Springer, vol. 2 , iss. 1, pp. 1-21, 2015. https://doi.org/10.1186/s40537-014-0007-7

[2] P. Thakuria, N. Y. Tilahun, and M. Zellner, "Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery," Seeing Cities Through Big Data, Springer, NY, pp.11-45, 2017.

[3] Q. Li, Y. Chen, J. Wang, Y. Chen, and H. Chen, "Web Media and Stock Markets : A Survey and Future Directions from a Big Data Perspective," IEEE Transactions on Knowledge and Data Engineering, vol. 30, pp.381-399, Feb. 1 2018.

[4] Eitan D. Hersh, and Brian F. Schaffner, "Targeted Campaign Appeals and the Value of Ambiguity," The Journal of Politics, The University of Chicago Press, vol. 75, iss. 2, pp.520-534, April 2013.

[5] Sjaak Wolfert, Lan Ge, Cor Verdouw, and Marc-Jeroen Bogaardt, "Big Data in Smart Farming – A review," Agricultural Systems, vol.153, pp. 69-80, 2017.

[6] M. Mazhar Rathore, Awais Ahmad, Anand Paul, and Seungmin Rho, "Urban planning and building smart cities based on the Internet of Things using Big Data analytics," Computer Networks, vol. 101, pp. 63-80, 2016.

[7] U. Sivarajah, M. Mustafa Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," Journal of Business Research, vol. 70, pp. 263-2862, 2017.

[8] V. Kale, Big data computing: A Guide for Business and Technology Managers. Taylor and Francis Group, CRC Press, 2017.

[9] D. V. Lande, I. V. Balagura, and V. B. Andrushchenko, "The detection of actual research topics using co-word networks," Open Semantic Technologies for Intelligent Systems : proceedings, Minsk, BNUIR, pp. 207-210, 2018.

[10] J. C. Hayes, and D. J. M. Kraemer, "Grounded understanding of abstract concepts: The case of STEM learning." Cognitive Research, vol. 2, iss.1, 2017. doi:10.1186/s41235-016-0046-z.

[11] D. V. Lande, V. B. Andrushchenko, and I. V. "Balagura Formation of the Subject Area on the Base of Wikipedia Service," Open Semantic Technologies for Intelligent Systems : proceedings, Minsk, BNUIR, pp. 211-214, 2017.

Data Stream Mining & Processing

PROCEEDINGS of the

2018 IEEE Second International Conference on

Data Stream Mining & Processing (DSMP)

IEEE Ukraine Section (Kharkiv)
SP/AP/C/EMC/COM
Societies Joint Chapter

IEEE Ukraine Section (West)
AP/ED/MTT/CPMT/SSC
Societies Joint Chapter

August 21–25, 2018

Lviv, Ukraine

Proceedings of the

2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)

Organized by

IEEE Ukraine Section

IEEE Ukraine Section (Kharkiv) SP/AP/C/EMC/COM Societies Joint Chapter

IEEE Ukraine Section (West) AP/ED/MTT/CPMT/SSC Societies Joint Chapter

IT Step University

Ukrainian Catholic University

Lviv Polytechnic National University

Kharkiv National University of Radio Electronics

Lviv, Ukraine
August 21-25, 2018

Additional copies may be ordered from:

IEEE Conference Operations
445 Hoes Lane, P.O. Box 1331, Piscataway, NJ
08855-1331 USA

DSMP'2018 Organizing Committee
IT Step University,
83a Zamarstynivs'ka st., 79019, Lviv, Ukraine

E-mail: dsmp.conference@gmail.com

IEEE Catalog Number: CFP18J13-CDR

ISBN: 978-1-5386-8175-6

Table of Contents