ЛАНДЭ Дмитрий Владимирович,

СНАРСКИЙ Андрей Александрович,

БЕЗСУДНОВ Игорь Васильевич

ИНТЕРНЕТИКА

Навигация в сложных сетях: модели и алгоритмы

Москва-2008

УДК 681.3
ББК 32.973.26-018.2.75
Л22

Рекомендовано к изданию
Ученым советом …
(протокол № … от … …..…. 2008 года)

Рецензенты:
А.Н. Новиков - доктор технических наук, профессор,
директор Физико-технического института Национального технического университета Украины "Киевский политехнический институт"
___.___. _____ - доктор физико-математических наук

Л 22 **Ландэ Д.В., Снарский А.А., Безсуднов И.В.**
Интернетика. Навигация в сложных сетях: модели и алгоритмы. –М.: …, 2008. – 300 с.
ISBN ….

Книга посвящена теоретическим и прикладным вопросам нового научного направления – интернетики, охватывающей основы теорий информационного поиска и сложных сетей. Авторы предполагают, что именно на стыке этих двух областей может лежать решение открытой проблемы навигации в современных информационных сетях.

В книге рассматриваются вопросы, относящиеся к информационной структуре веб-пространства, теории сложных сетей, моделям информационного поиска и глубинного анализа текстов, общим закономерностям современный информационных потоков и их моделированию.

Книга рассчитана на широкий круг читателей: специалистов в области информационных технологий, прикладных лингвистов, студентов, аспирантов, аналитиков в различных областях и может служить основой для построения учебных курсов, посвященных вопросам информационного поиска в сетевой среде.

ISBN …

УДК 681.3
ББК 32.973.26-018.2.75

Нашим женам - Наталье, Галине и Екатерине.

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ

Основная идея этой книги – показать связь двух активно развивающихся в настоящее время направлений – теорий информационного поиска и сложных сетей. Именно на стыке этих двух областей может лежать решение открытой проблемы эффективной навигации в современных информационных сетях.

Самое подходящее назание такой интеграции, нового научного направления - Интернетика. Во-первых, это направление является развитием информатики, и, что должно быть созвучно этому термину. Связь с теорией сложных сетей [116] обуславливает наличие корня «нет», однако подразумевается, что исследования в рамках данного направления выйдут за рамки конкретной сети Интернет, анализ которой, безусловно входит в сферу интернетики. Во-вторых, этот термин, хотя уже и встречается, но еще недостаточно устоялся. Известны по меньшей мере две трактовки термина «интернетика». В рамках первой интернетика рассматривается как прикладное научное направление, изучающее свойства и способы использования Интернет преимущественно в аспекте воздействия на социально-экономические процессы [35]. Эта трактовка, по нашему мнению, несколько сужает область исследований (хотя и способствует популярности). Вторая трактовка, автором которой является Дж. Фокс (G. Fox) из Сиракузского университета (США), заключается в том, что интернетика – это развитие информатики в направлении применения современных параллельных сетевых вычислений во всех областях науки, охватывая огромные ресурсы, распределенные в сетевой среде [91, 92]. Вторая трактовка понятия «интернетика», предполагающая использование методов точных наук гораздо ближе авторам, чем первая.

Сегодня структура и объемы информационных потоков, в которых приходится выискивать крупицы необходимой, готовой к непосредственному использованию, обуславливают актуальность самого процесса поиска. Развитие Интернет породило ряд специфических проблем, связанных, в первую очередь, с возрастанием объемов данных в веб-пространстве, в том числе и бесполезных, шумовых. По-видимому, организация поиска необходимой информации в этом

информационном хранилище требует новых подходов. Можно предположить, что современные информационные технологии готовы к подобному пересмотру принципов обеспечения доступа к сетевым данным.

Многие подходы, излагаемые в этой книге, уже стали классическими и широко используются в практике информационного поиска и анализа информации. Авторы попытались дать систематический и вместе с тем достаточно популярный обзор основных моделей, рассматриваемых в рамках теории информационного поиска, научного направления, сформировавшегося в конце XX века. Кроме того, в книге также представлены процедурные основы фрактального анализа, который применяется для исследования информационных потоков.

Сегодня в Интернет существует доступная для экспериментов динамичная информационная база такого объема, который ранее даже трудно было представить. При этом оказалось, что многие задачи, возникающие при работе с сетевым информационным пространством, имеют немало общего, например, с задачами теоретической физики. Это обстоятельство открывает широкие перспективы применения мощного аппарата естественных наук.

Вместе с тем реальный прорыв в области информационного поиска возможен лишь в результате агрегирования различных научных направлений. Излагаемые в книге результаты исследований современного сетевого информационного пространства с нескольких, ранее порой конфликтующих точек зрения, могут представлять интерес как для специалистов в области компьютерной лингвистики, так и для прикладных математиков и физиков, например, в плане аналогового моделирования статистических процессов, в том числе систем с элементами самоорганизации.

Книга ориентирована на достаточно широкий круг читателей: специалистов в области информационного поиска, прикладных лингвистов, студентов, аспирантов; хочется верить, что она будет также полезна и аналитикам, которые при решении задач в различных областях хотят учитывать особенности современного сетевого информационного пространства. Надеемся, что эта книга

окажется также полезной при подготовке учебных курсов по теоретическим и практическим вопросам информационного поиска.

Авторы выражают искреннюю благодарность Сергею Брайчевскому и Александру Дармохвалу за конструктивное обсуждение содержания книги, Александру Снарскому за помощь в обсуждении и редактировании разделов, связанных с технологиями реальных сетей, и Алексею Новикову за конструктивные замечания.

Дмитрий Ландэ, Андрей Снарский,
Игорь Безсуднов

9 мая 2008 г.

ВВЕДЕНИЕ

Nothing`s gonna change my world…
J. Lennon, P. McCartney

Эта книга посвящена новому научному направлению – интернетике. Сегодня в информационных хранилищах, распределенных в сетях, собраны террабайты текстовых данных. Эти данные можно рассматривать, с одной стороны, как сетевую среду реального информационного поиска, а с другой, как объект и полигон для исследований. Учет этих факторов привел к необходимости представить краткий обзор истории и современного состояния инфраструктуры Интернет, остановиться на особенностях гипертекстовых технологий, сети WWW, а также обозначить перспективы.

Именно этим аспектам посвящена первая глава книги. Для обеспечения поиска размещенной в сети информации в настоящее время необходима разработка новых подходов. При этом, безусловно, должны учитываться достоинства и недостатки существующих моделей и алгоритмов информационного поиска, которым посвящена вторая глава. В этой главе уделено внимание также моделям поиска в пиринговых сетях - крупнейших по ресурсам и порождаемому интернет-трафику. В таких сетях отсутствуют выделенные серверы, а каждый узел является как клиентом, так и сервером. Пиринговые сети состоят из узлов, каждый из которых взаимодействует лишь с некоторым подмножеством других узлов. При освещении этой тематики учитывались то, что проблемы поиска и уязвимости в таких сетях до сих пор остаются открытыми. Рассмотрены основные модели поиска, все более широко применяемые в пиринговых сетях, а также проблемы, связанные с распространением подобных сетей.

В третьей главе рассматривается концепция глубинного анализа текстов – Text Mining, которая включила в себя технологические и методологические подходы контент-анализа, компьютерной лингвистики. В частности, в этой главе освещены подходы к решению таких задач, как автоматическе реферирование, анализ взаимосвязей понятий, построение поисковых образов документов.

Классификация информации - это традиционная компонента теории и технологии информационного поиска, лежащая на стыке двух областей - машинного обучения и информационного поиска. При классификации текстов, методы которой детально рассматриваются в четвертой главе, используются различные критерии для построения правил их размещения в заранее определенные категории.

Пятая глава посвящена вопросам кластерного анализа массивов текстовых документов. В отличие от классификации, при кластеризации заранее не фиксируются определенные категории. Результатом кластеризации является автоматическая группировка информации в компактные подгруппы. Алгоритмы кластеризации позволяют автоматически находить «скрытые» признаки и разделять объекты по подгруппам. Кластеризация, как правило, предшествует классификации, поскольку помогает экспертам определять группы объектов - классы. В этой же главе подробно рассмотрены основные алгоритмы ранжирования выдачи информационно-поисковых-систем.

В шестой главе приводятся основные закономерности, присущие документальным потокам в современной сетевой среде. При этом уделяется внимание таким необходимым для понимания этих закономерностей математическим понятиям, как степенные распределения, однородные функции и скейлинг.

Теория информации, которая ранее находила свое основное применение в области передачи данных, становится полезной и для анализа текстовых массивов, динамически порождаемых в сетях. Седьмая глава посвящена таким понятиям, как энтропия и количество информации, которые сегодня находят все большее применение в технологиях информационного поиска.

Восьмая глава посвящена теории сложных сетей (complex networks), в рамках которой рассматриваются характеристики, учитывающие не только их топологию, но и статистические распределения характеристик узлов и связей. Сегодня эта теория особо актуальна в задачах выявления и визуализации различных сетевых кластеров, их внутренних корреляций.

Явления, происходящие в сложных сетях, близки к изучаемым в рамках теории перколяции (протекания), элементы которой излагаются в девятой главе. К задачам теории перколяции и анализа сложных сетей относятся такие, как определение предельного уровня проводимости (пропускной способности), изменения длины пути между узлами и его траектории (извилистости, параллельности) при приближении к порогу протекания, количества узлов, которые необходимо удалить, чтобы нарушить связанность сети.

Математическому моделированию информационных потоков посвящена десятая глава, в которой рассматриваются модели, учитывающие «конкуренцию» реальных тематик. При моделировании этих процессов используются методы нелинейной динамики, теории клеточных автоматов и самоорганизованной критичности.

При моделировании информационных потоков изучаются структурные связи между входящими в них массивами документов. Сегодня при этом все чаще применяется фрактальный анализ, подход, базирующийся на свойствах сохранения внутренней структуры массивов документов при изменениях их размеров или масштабов рассмотрения. Этому посвящена одиннадцатая глава.

Хочется подчеркнуть, что традиционно используемый математический аппарат и инструментальные средства информационного поиска сегодня уже не способны в полной мере удовлетворять потребности пользователей. Изначальная парадигма поисковых систем, сформированная несколько десятилетий тому назад, уже не отвечает реальной ситуации – объемам и динамике информационных потоков, сетевой топологии. Необходим поиск новых принципов, в рамках которых оказалось бы возможным проектирование качественно новых систем обработки больших и динамичных массивов данных. Цель этой книги – систематически изложить состояние существующих теоретических и технологических возможностей, представить читателю возможные перспективы развития, дать импульс новым идеям в области сетевого информационного поиска.

1. СОВРЕМЕННЫЕ ИНФОРМАЦИОННЫЕ СЕТИ

«Дурь быстро множится,
Дурь молится молве…»
Новелла Матвеева

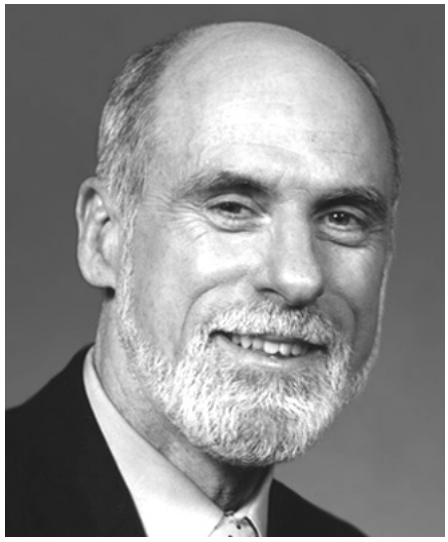
1.1. Интернет – история и протоколы

Интернет – это глобальная компьютерная сеть (или просто, «Сеть»), состоящая из многих тысяч корпоративных, научных, правительственных и домашних компьютерных сетей, части которой взаимосвязаны. Объединение сетей разной архитектуры и топологии стало возможно благодаря протоколам TCP/IP (Transmission Control Protocol/Internet Protocol), единому адресному пространству и принципу маршрутизации пакетов данных. Протокол IP был специально создан независимым относительно физических каналов связи, в результате чего практически любая сеть передачи цифровых данных, проводная или беспроводная, может передавать также трафик Интернет. На стыках сетей специальные маршрутизаторы занимаются автоматической сортировкой и перенаправлением пакетов данных, исходя из IP-адресов получателей этих пакетов. Сети на основе протокола IP образуют единое адресное пространство в масштабах всего мира, но в каждой отдельной сети может существовать и собственное адресное подпространство, выбираемое исходя из класса этой сети. Такая организация IP-адресов позволяет маршрутизаторам однозначно определять дальнейшее направление для каждого мелкого пакета данных.

Протокол IP был разаработан в рамках проекта ARPANET. Сегодня развитием протоколов Сети занимается организация IETF (Internet Engineering Task Force), название которой можно перевести как «Группа по решению задач проектирования Интернет».

История сети Интернет, которая началась с ARPANET в США, насчитывает около 40 лет. В 1969 году в Министерстве обороны США было принято решение, что на случай войны Америке нужна надежная система передачи информации. Агентство передовых исследовательских проектов (ARPA) предложило

разработать для этого компьютерную сеть. Разработка такой сети была поручена Калифорнийскому университету в Лос-Анжелесе, Стенфордскому исследовательскому центру, университету штата Юта и университету штата Калифорния в Санта-Барбаре.

Изобретатели стека TCP/IP Р. Кан (R. Kahn) и В. Серф (V. Cerf)

Компьютерная сеть была названа ARPANET, в рамках проекта сети были объединены усилия четырех указанных научных учреждений, все работы финансировались за счет Министерства обороны США. ARPANET обеспечивала:

- проведение экспериментов в области компьютерных коммуникаций;
- объединение научного потенциала исследовательских учреждений;
- изучение способов поддержки устойчивой связи в условиях ядерного нападения;
- разработку концепции распределенного управления военными и гражданскими структурами в период ведения войны.

Многие из существующих протоколов Интернета ведут свое начало в ARPANET. Например, протокол обратного поиска DNS до сих пор использует доменное имя верхнего уровня «.arpa»: чтобы найти записи, которые относятся, например, к IP-адресу 1.2.3.4, надо послать запрос на получения адреса 4.3.2.1.in-addr.arpa.

Сеть ARPANET стала активно расти и развиваться, ее начали использовать ученые из разных областей науки. В 1973 году к сети были подключены первые иностранные организации из Великобритании и Норвегии, сеть стала международной. В 1984 году у сети ARPANET появился серьезный соперник в лице Национального фонда науки США (NSF), основавшего большую межуниверситетскую сеть NSFNet, которая имела намного большую пропускную способность (56 Кбит/с), чем ARPANET. В 1990 году сеть ARPANET прекратила свое существование, полностью проиграв конкуренцию NSFNet.

Для того чтобы разные компьютеры Сети могли взаимодействовать, они должны «разговаривать на одном языке». Протоколы Интернет - это тот «язык», который используется компьютерами для обмена данными при работе в этой сети [47]. Систему протоколов Интернет называют стеком TCP/IP.

Роль нормативных документов в сети Интернет выполняют «запросы комментариев» (англ. Request for Comments, RFC), документы из серии пронумерованных информационных документов сети Интернет, к которым относятся технические спецификации и соответствующие Стандарты. Название «Request for Comments» еще можно перевести также как «заявка на обсуждение» или «тема для обсуждения». Сегодня первичной публикацией документов RFC занимается IETF под эгидой открытой организации «Сообщество Интернет» (англ. Internet Society, ISOC). Правами на RFC владеет само Сообщество Интернет, которое обеспечивает свободный доступ к этим документам (см., например, http://www.ietf.org/rfc.html). Сообщество Интернет является организационной основой различных исследовательских и консультативных групп, которые занимаются развитием Сети.

Сегодня почти все стандарты Интернет разрабатываются под эгидой известных научных или интернет-организаций (например W3C, IETF, консорциумами Юникод, Интернет2).

До 1990 -1994 г. наблюдался рост количества ресурсов и Интернет, как сети обмена данными. Затем этот процесс достиг своего естественного насыщения. После 1994 года наблюдается новый этап бурного развития Интернет (рис. 1), связанный с появлением веб-технологий. Назовем лишь некоторые причины,

благодаря которым именно сеть Интернет среди сотен других получила такое развитие:

- высокая технологичность, надежность и стойкость;
- открытость протоколов;
- поддержка пользователями и производителями программного обеспечения;
- способность к саморазвитию, саморасширению;
- постоянное снижение затрат абонентов на работу в Интернет;
- де-факто новый вид интерактивного СМИ.

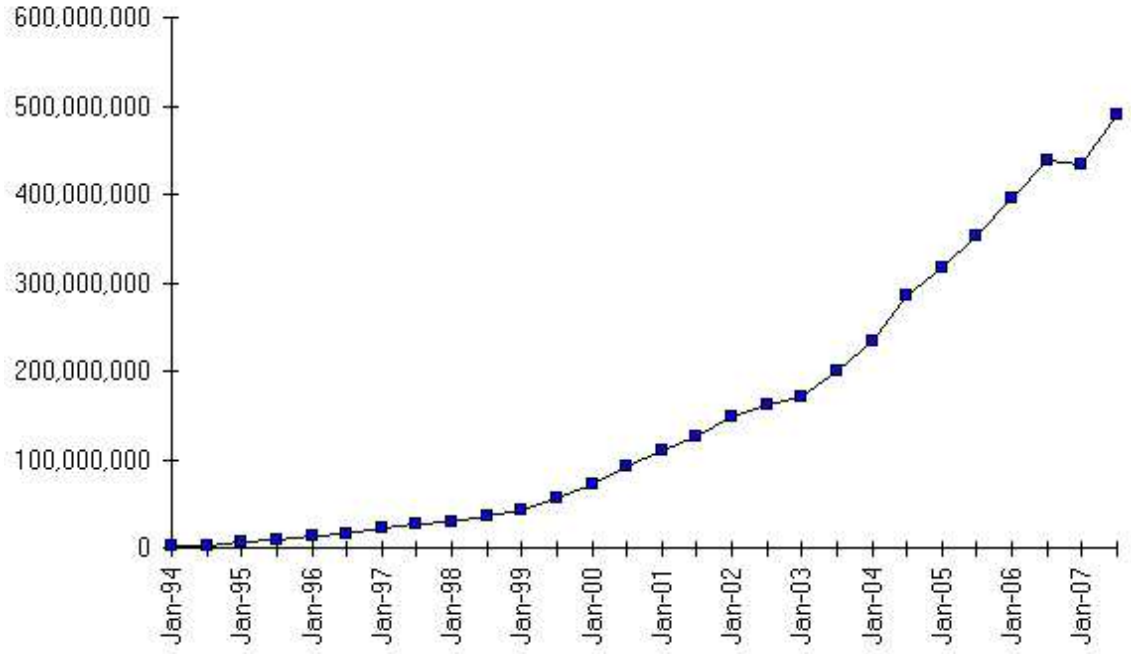


Рис. 1. Динамика развития Интернет как телекоммуникационной сети [Источник: www.isc.org]

1.2. Всемирная паутина - World Wide Web

World Wide Web (или сокращенно, веб) представляет глобальное информационное пространство, основанное на физической инфраструктуре сети Интернет и протоколе передачи данных HTTP.

World Wide Web объединяет миллионы веб-серверов, подключенных к Интернет. В начале существования World Wide Web на небольшом количестве

веб-сайтов публиковалась информация отдельных авторов для относительно большого количества посетителей. Сегодня ситуация резко изменилась. Сами посетители веб-сайтов активно участвуют в создании контента, что привело к резкому росту объемов информации и динамики веб.

Сегодня в веб уже существует свободно доступная для пользователей информационная база такого объема, который ранее трудно было представить. Более того, объемы этой базы превышают на порядки все то, что было доступно десятилетие назад. В августе 2005 года компания Yahoo объявила о том, что проиндексировала около 20 млрд. документов. Достижение компании Google в 2004 году составляло менее 10 млрд. документов, т.е. за один год количество информации. По данным службы Web Server Survey, в авпреле 2008 года количество веб-серверов превысило 166 млн.

Для просмотра информации, полученной от веб-серверов, на компьютерах полязователей используются специальные программы - веб-браузеры, основная функция которых - отображение гипертекста, являющегося основным методом представления информации в веб.

Традиционно под гипертекстом понимается принцип организации текстовых массивов, при котором отдельные информационные элементы связаны между собой ассоциативными отношениями (гиперссылками), обеспечивающими быстрый и удобный поиск необходимой информации и/или просмотр взаимозависимых данных.

Гипертекст, появившийся как форма гиперсвязи между отдельными фрагментами текста, настолько же древнее понятие, как и письменность. Библия, с ее сложным употреблением аннотаций и комментариев, - один из древнейших примеров гипертекста. Словари и энциклопедии также могут рассматриваться как сети из текстовых блоков, соединенных ссылками.

Основные вехи развития современных гипертекстовых технологий следующие:

- в 1945 году Ванневер Буш (Vannevar Bush) создал первую фотоэлектрическую память и приспособление Memex (memory extension), которая представляет собой справочник, реализованный с помощью гиперссылок в

пределах документа. Тед Нельсон (Ted Nelson) в 1965 году ввел термин "гипертекст" и создал гипертекстовую систему Xanadu с двусторонними гиперсвязями;

- в 1980 году Тим Бернерс-Ли (T. Berners-Lee), консультант CERN (Европейская организация ядерных исследований) написал программу, которая разрешает создавать и пересматривать гипертекст, реализующий двунаправленные связи между документами из коллекции;

- в 1990 году для поддержки документации, которая циркулирует в CERN Бернерс-Ли начал работу над графическим интерфейсом пользователя (GUI, Graphical User Interface) для гипертекста. Эта программа была названа "WorldWideWeb". До 1992 года уже были созданные такие GUI как Erwise и Viola.

- в феврале 1993 года M. Андрессен (M. Anderssen) из NCSA (Национальный Центр Суперкомпьютерных приложений США, www.ncsa.uiuc.edu) закончил первую версию программы визуализации гипертекста Mosaic для популярного графического интерфейса Xwindow System под UNIX. Одновременно CERN развивал и улучшал HTML - язык гипертекстовой разметки текстов и HTTP - протокол передачи гипертекста, а также сервер обработки гипертекстовых документов - CERN HTTPD.

Язык HTML (Hypertext Markup Language) представляет собой стандартный язык разметки документов в Интернет, при помощи которого создаются все веб-страницы.

HTML вначале создавался как язык для обмена научной и технической документацией, пригодный для использования людьми, не являющимися специалистами в области верстки. Прародителем HTML является язык SGML (стандартный обобщенный язык разметки), язык HTML является подмножеством SGML, т.е. удовлетворяет международному стандарту ISO 8879.

Язык HTML позволяет размечать текст для форматированного отображения, а также для реализации некоторых элементов интерактивности. Текст с HTML-разметкой интерпретируется специальными программами - браузерами и отображается в виде документа, удобного для восприятия человеком. Браузеры предоставляют пользователю интерфейс для запроса веб-страниц, их просмотра

и, при необходимости, отправки введенных пользователем данных на сервер. Наиболее популярными на сегодня браузерами являются Internet Explorer, Firefox, Opera и Safari.

Последней актуальной версией является HTML 4.01, принятый в 1999 году. В 2000 г. был принят международный стандарт ISO/IEC 15445:2000 (так называемый «ISO HTML», основаный на HTML 4.01 Strict).

В настоящее время Консорциумом W3C разрабатывается пятая версия языка HTML. Черновой вариант спецификации языка появился в Интернете 20 ноября 2007. Параллельно ведётся работа по дальнейшему развитию HTML под названием XHTML (от англ. eXtensible HTML), который, в отличие от предшественника, базирующегося на SGML, основан на XML и в 2000 году был одобрен в качестве Рекомендаций W3C.

Для передачи в сети Интернет гипертекстовой информации используется протокол HTTP (HyperText Transfer Protocol), который вначале использовался исключительно для передачи HTML-документов. В настоящее время с помощью HTTP можно передавать любую информацию, в том числе изображения, звук, видео а также просто абстрактные файлы.

Протокол HTTP определяет простое взаимодействие вида запрос-ответ. Каждое HTTP-взаимодействие состоит из запроса, посылаемого от клиента серверу, и следующего за ним ответа от сервера клиенту. HTTP-запрос состоит из нескольких частей: метода, указывающего на действие (GET, POST, HEAD, PUT), адреса ресурса – его унифицированного указателя (Uniform Resource Locator, URL), а также другой информации, например, такой как тип требуемого документа, аутентификация и разрешение на оплату. URL – это схема указания местонахождения ресурсов в Интернет, которая состоит из трех частей:

- схемы, указывающей название протокола, используемого для доступа к ресурсу (например, ftp);
- адреса сервера, задающего сетевое имя компьютера, на котором ресурс расположен;
- точного адреса объекта, задающего полный путь и имя запрашиваемого объекта на сервере.

Для протокола HTTP формат записи URL имеет вид:

http://[user[:passwd@]host[:port][/path]

где *host* - имя компьютера в Интернет или его IP адрес; *:port* - номер TCP порта для доступа к сервису, если сервер настроен на использование порта, отличного от принятого по умолчанию; *path* - полный путь и имя запрашиваемого объекта; *user* - пользователь; *passwd* - пароль.

На рис. 2 приведена упрощенная схема взаимодействия по протоколу HTTP.

После получения запроса от компьютера пользователя (клиента) сервер выполняет его синтаксический разбор, а затем необходимые действия, определяемые указанным методом. После этого сервер отправляет клиенту ответ, состоящий из строки состояния, указывающей на результат обработки запроса, например, успешно ли он обработан, информации о типе возвращаемого объекта и запрашиваемой информации, а также файл или результаты, сгенерированные серверным приложением.

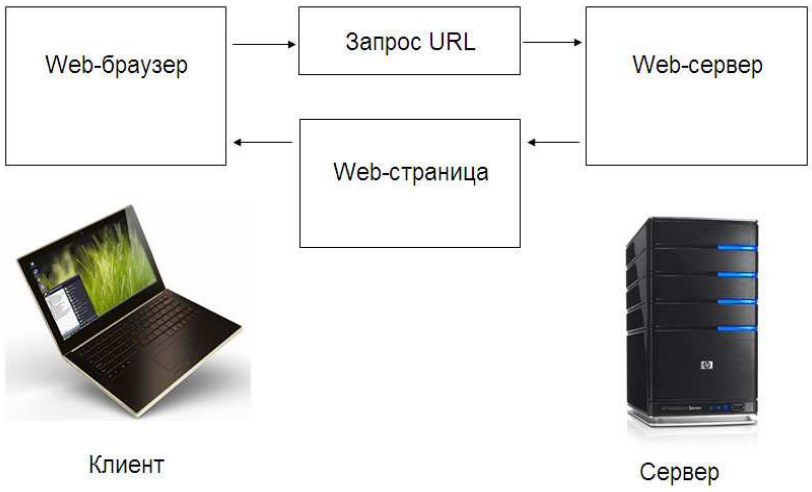


Рис. 2. Схема взаимодействия клиента и сервера по протоколу HTML

В первых версиях протокола HTTP соединение между клиентом и сервером осуществлялось только в промежутке между посылкой запроса и ответом сервера. Сразу после отправки ответа сервер закрывал соединение, что давало ему возможность продолжить обработку ждущих запросов от других клиентов. HTTP называют "протоколом без запоминания состояния", поскольку он не

поддерживал концепцию сеанса связи, обеспечивая лишь доставку запрошенного документа.

В современной версии протокола HTTP (известной как HTTP 1.1, HTTP с устойчивым соединением) TCP-соединение между двумя последовательными операциями остается открытым. Этот метод, называемый "устойчивым соединением" (permanent connection), использует одно и то же TCP-соединение для обслуживания множества HTTP-запросов, при этом исключаются расходы на открытие и закрытие других соединений. Еще одной особенностью протокола HTTP 1.1, влияющей на производительность, является конвейерная обработка запросов, которая позволяет послать сразу много запросов, не ожидая ответа на каждый из них. То есть клиент отправляет множество запросов через TCP-соединение до того, как получит ответ на свои предыдущие запросы.

Таким образом, главными этапами, из которых состоит HTTP-взаимодействие, являются:

- установка соответствия между именем сервера и IP-адресом (с помощью DNS-сервера);
- установка TCP-соединения с сервером;
- передача запроса URL;
- получение ответа (HTML-текста или мультимедиа);
- закрытие TCP/IP-соединения.

Главный недостаток HTML-технологий заключается в том, что HTML был изначально предназначен прежде всего для визуализации данных, исключительно для структурирования содержания сайтов. Несмотря на то, что в последнее время для отображения данных в основном используются специальные средства, расширяющие возможности HTML, в частности, каскадные таблицы стилей CSS (Cascading Style Sheets), он по-прежнему остается неудобным для автоматической обработки информации, в том числе, для организации поиска. То есть, WWW ориентирован на показ пользователям отдельных сайтов и плохо приспособлен для автоматизированного сбора информации, ее классификации и аналитической обработки. Сегодня представление информации на разных сайтах существенным

образом отличается по оформлению и расположению, что усложняет автоматическую обработку.

Так, при необходимости обмена информацией между несколькими веб-сайтами, всегда возникает задача унифицированного представления контента. В противном случае изменение HTML-оформления одного сайта приведет к необходимости одновременной модификации программного обеспечения на всех сайтах, которые принимают его информацию. Аналогичная ситуация возникает при необходимости импортировать информацию на один сайт с нескольких других. Изменение оформления на каждом из сайтов-источников информации будет всегда приводить к необходимости модификации соответствующего программного кода на целевом сайте. Поэтому в настоящее время обновления сайтов предоставляются не в HTML, а в диалектах формата XML, предназначенных для обмена данными и их интеграции.

1.3. Пиринговые сети

В настоящее время WWW не является самой крупной сетью по ресурсам и порождаемому интернет-трафику. Известно, что трафик, объем информационных ресурсов (в байтах), количество узлов пиринговых сетей, если их рассматривать в совокупности, существенно превосходят соответствующие показатели веб. При этом можно отметить, что проблемы поиска и уязвимости в пиринговых сетях, как крупнейшего «белого пятна» современных коммуникаций, пока остаются открытыми.

Пиринговые сети (Peer-to-peer, P2P – равный с равным) - это компьютерные сети, основанные на равноправии участников. В таких сетях отсутствуют выделенные серверы, а каждый узел (peer) является как клиентом, так и сервером. Впервые фраза «peer-to-peer» была использована в 1984 году П. Йохнухуйтсманом (P. Yohnuhuitsman) при разработке архитектуры Advanced Peer to Peer Networking фирмы IBM.

P2P – это сетевой протокол [14], обеспечивающий возможность создания и функционирования сети равноправных узлов, их взаимодействия. Во многих случаях P2P являются наложенными сетями, использующими существующие

транспортные протоколы стека TCP/IP - TCP или UDP. Следует отметить, что на практике пиринговые сети состоят из узлов, каждый из которых взаимодействует лишь с некоторым подмножеством других узлов сети (из-за ограниченности ресурсов). Для реализации протокола P2P используются клиентские программы, обеспечивающие функциональность как отдельных узлов, так и всей пиринговой сети.

Несмотря на то, что все узлы в P2P имеют одинаковый статус, реальные возможности их могут существенно отличаться. На практике большинство пиринговых сетей дополняются выделенными серверами, несущими организационные функции, например авторизацию [51]. В частности, известны библиотечные пиринговые сети, в которых используются выделенные серверы, играющие роль центров авторизации, хеширования и репликации библиографических данных [14].

Архитектура пиринговых сетей принципиально отличается от традиционной централизованной архитектуры «клиент/сервер», подразумевающей, что сеть зависит от центральных узлов (серверов), которая обеспечивает подключенные к сети терминалы (клиенты) необходимыми сервисами. В этой централизованной архитектуре ключевая роль отводится серверам, которые определяют сеть независимо от наличия клиентов, т.е. при падении этих серверов сеть становится нерабочей. Очевидно, что рост количества клиентов сети типа «клиент/сервер» приводит к росту нагрузок на серверную часть, в результате чего она может оказаться перегруженной.

Децентрализованная пиринговая сеть, напротив, становится более производительной при увеличении количества узлов, подключенных к ней. Действительно, каждый узел добавляет в сеть P2P свои ресурсы (дисковое пространство и вычислительные возможности), в результате суммарные ресурсы сети увеличиваются. При этом, конечно, нельзя не учитывать того факта, что на практике большинство P2P-сетей все же зависят от своих центральных узлов, например, одна из самых известных таких сетей BitTorrent зависит от так называемых «трекеров» (tracker), при падении которых эта сеть станет нерабочей.

По сравнению с клиент/серверной архитектурой (например, с веб), архитектура P2P обладает такими преимуществами, как самоорганизованность, отказоустойчивость при потере связи с узлами сети, возможность разделения ресурсов без привязки к конкретным адресам, увеличение скорости копирования информации за счет использования сразу нескольких источников, более эффективное использование полосы пропускания, гибкая балансировка нагрузки.

Первые пиринговые сети, в частности, Gnutella, широко использовали метод BFS, называемый еще методом размножения запросов, который ведет к экспоненциальному росту числа сообщений запросов и, соответственно, ведет к перегрузке сети. Недостатки рассмотренного метода инициировали разработку группы протоколов DHT (Distributed Hash Tables), в частности, протокола Kademlia, который сейчас широко используется в наиболее крупных P2P-сетях. Были введены правила, в соответствии с которыми запросы могут пересылать вверх по дереву только определенные узлы, так называемые, концентраторы, остальные узлы могут лишь посылать им запросы. Эти правила были реализованы в 2003 году как протокол новой сети Gnutella2. В соответствии с этим протоколом у концентратора есть связь с сотнями узлов и десятки соединений с другими концентраторами. Каждый узел пересылает концентратору список идентификаторов ключевых слов, по которым могут быть найдены публикуемые ей ресурсы. Для улучшения качества поиска используются также метаданные файлов - информация о содержании, рейтинги. Допускается возможность «размножения» информации о файле в сети без копирования самого файла.

В настоящее время при реализации пиринговых сетей используются самые различные подходы. В частности, компания Microsoft разработала протоколы для P2P-сетей Scribe и Pastry. Поддержка протокола PNRP (Peer Name Resolution Protocol), также относящегося к P2P-системам, была включена в состав Windows Vista.

Одну из удачных попыток стандартизации протоколов P2P предприняла компания Sun Microsystems в рамках проекта JXTA [28]. Этот проект реализуется с целью унифицированного создания P2P-сетей для различных платформ.

Существует несколько областей применения пиринговых сетей, объясняющих их растущую популярность, назовем некоторые из них:

- *Обмен файлами*. P2P выступают альтернативой FTP-архивам, которые утрачивают перспективу ввиду значительных информационных перегрузок.
- *Распределенные вычисления*. Например, такой проект с элементами P2P, как SETI@HOME, посвященный распределенному поиску внеземных цивилизаций, продемонстрировал высокий вычислительный потенциал для распараллеливаемых задач. Вместе с тем, этому проекту свойствена централизованая раздача и прием данных.
- *Обмен сообщениями*. Как известно, ICQ – это P2P-проект. Эта сеть также обладает элементами централизации, в частности, очень зависит от состояния сервера login.icq.com.
- *Интернет телефония*. Сегодня одной из самых популярных служб Интернет-телефонии является Skype (www.skype.com), созданная в 2003 г. Н. Зеннстромом и Я. Фриисом, авторами известной пиринговой сети KaZaA. Построенная в архитектуре P2P служба Skype охватывает свыше 10 млн. пользователей.
- *Групповая работа*. Сегодня реализованы такие сети групповой работы, как Groove Network (защищенное пространство для коммуникаций) и OpenCola (поиск информации и обмен ссылками).

Отдельного рассмотрения заслуживают файлообменные P2P-сети, которые в начале 2008 года охватывали уже более 150 млн. узлов. Сегодня в Интернет более половины всего трафика приходится на файлообменные P2P-сети. Наиболее популярные из них - это Bittorrent, Gnutella2 и eDonkey2000.

Сеть BitTorrent была создана в 2001 году. В соответствии с протоколом BitTorrent файлы передаются не целиком, а частями, причем каждый клиент, закачивая эти части, в это же время отдает их другим клиентам, что снижает нагрузку и зависимость от каждого клиента-источника и обеспечивает избыточность данных. С целью инициализации узла в сети трекеров (www.bittorrent.com) клиентская программа обращается к выделенному серверу,

предоставляющему информацию о файлах, доступных для копирования, а также содержащем статистическую и маршрутную информацию об узлах сети. Если узел “хочет” опубликовать файл, то программа разделяет этот файл на части и создает файл метаданных (torrent file) с информацией о частях файла, их местонахождении и узла, который будет поддерживать распространение этого файла.

В 2000 г. была создана одна из первых пиринговых сетей Gnutella (www.gnutella.com). Сегодня наиболее популярна более поздняя версия этой сети - Gnutella2 (www.gnutella2.com), созданная в 2003 году, которая реализует открытый файлообменный P2P-протокол.

Сеть EDonkey2000 была создана в 2000 году. Информация о наличии файлов в ней публикуется клиентом на серверах в виде так называемых ed2k-ссылок, использующих уникальный ID ресурса. В сети EDonkey2000 выделенные серверы обеспечивают поиск узлов и информации. Существует около 200 серверов и порядка миллиарда файлов. Число пользователей EDonkey2000 составляет более 10 млн. человек. При работе каждый клиент EDonkey2000 связан с одним из серверов. Клиент сообщает серверу, какие файлы он предоставляет в общий доступ. Каждый сервер поддерживает список всех общих файлов клиентов, подключенных к нему. Когда клиент что-то ищет, он посылает поисковый запрос своему основному серверу. В ответ сервер проверяет все файлы, которые ему известны, и возвращает клиенту список файлов, удовлетворяющих его запросу.

Существует много областей, где успешно применяется P2P-технология, например, параллельное программирование, кэширование данных, резервное копирование данных.

Благодаря таким характеристикам, как живучесть, отказоустойчивость, масштабируемость, пиринговые сети находят все большее применение в системах управления производствами и организациями (например, P2P-технология сегодня применяется в Государственном Департаменте США). В данном случае возможный выход из строя части узлов или серверов не существенно влияют на управляемость всей системы. Общеизвестно, что система доменных имен (DNS) в

сети Интернет также фактически является сетью обмена данными, построенной по принципу P2P [115].

Реализацией технологии P2P является также популярная в настоящее время система распределенных вычислений GRID. Еще одним примером распределенных вычислений может служить проект distributed.net, участники которого занимаются легальным взломом криптографических шифров, чтобы проверить их надежность.

Помимо названных выше преимуществ пиринговых сетей, им присущ также ряд недостатков, первая группа которых связана со сложностью управления по сравнению с клиент-серверными системами. Приходится тратить значительные усилия на поддержку стабильного уровня их производительности, резервное копирование данных, антивирусную защиту, защиту от информационного шума и других злонамеренных действий пользователей.

Большая проблема – это легитимность контента, передаваемого в P2P-сетях. Неудовлетворительное решение этой проблемы привело уже к скандальному закрытию многих таких сетей (например, Napster в июле 2001 года). Есть и другие проблемы, имеющие социальную природу. Так в системе Gnutella, например, 70% пользователей не добавляют вообще никаких файлов в сеть. Более половины ресурсов в этой сети предоставляется одним процентом пользователей, т.е. сеть эволюционирует в направлении клиент-серверной архитектуры.

Еще одна проблема P2P-сетей связана с качеством и достоверностью предоставляемого контента. Серьезной проблемой является фальсификация файлов и распространение фальшивых ресурсов. Еще одной проблемой является возможность фальсификации ID узлов. Защита распределенной сети от хакерских атак, ботнетов, вирусов и «троянских коней» является весьма сложной задачей. Зачастую информация с данными об участниках P2P-сетей хранится в открытом виде, доступном для перехвата.

1.4. Проблемы развития интернет-контента

Сегодня существует несколько глобальных проблем, связанных с развитием интернет-контента, среди которых две главнейшие (первая из которых пародоксальная):

- прогресс в области производства информации ведет к снижению уровня информированности людей;
- интенсивность роста объемов шумовой информации во много раз превышает интенсивность роста объемов полезной информации.

Новый уровень развития сетевого информационного пространства обуславливает необходимость создания и развития адекватных моделей информационного пространства, информационных потоков, сетевого поиска. В этой связи возникает интерес к подходам, основанным на понимании информации как меры упорядоченности некоторой системы и, соответственно, к статистическим методам ее обработки. Для организации эфективной коммуникации в сетях сегодня приходится постоянно возвращаться к истокам теории информации, понятиям энтропии, теории Шеннона, уравнениям Больцмана и др., что обуславливает широкие перспективы применению мощного аппарата математики и физики в решении теоретико-информационных задач.

2. ИНФОРМАЦИОННЫЙ ПОИСК

«Ну, любимый, смущайся и ахай:
ты лишь кликнул, а я уже тут.»
Майя Борисова

Доступ пользователей к современным информационным сетям, эффективное удовлетворение их информационных потребностей возможно только с помощью развитых средств навигации в этих сетях. Основным инструментом при этом выступают информационно-поисковые системы, обеспечивающие поиск в гигантских объемах текстовой информации.

Первые реально функционирующие полнотекстовые информационно-поисковые системы (Retrieval Systems, ИПС) появились в начале компьютерной эры. Назначением этих систем был поиск в библиотечных каталогах, архивах, массивах документов, таких как статьи, нормативные акты, рефераты, брошюры, диссертации, монографии.

Основными функциями информационно-поисковых систем изначально были:

- хранение больших объемов информации;
- быстрый поиск необходимой информации;
- добавление, удаление и изменение хранимой информации;
- вывод информации в удобном для пользователя виде.

В 1966 году 16-ю американскими библиотеками для установления стандартного формата для электронных каталогов была начата реализация проекта MARC (http://www.loc.gov/marc/), обеспечившего переход к унифицированному обмену электронными данными, что способствовало эффективной организации электронных каталогов. Внедрение стандартного библиографического формата позволило библиотекам объединить усилия. В 1972 году получил международное признание стандарт MARC-2 [67, 32], на основе которого были созданы многие национальные стандарты.

В начале 1970-х годов коммерческие компьютерные службы уже предоставляли возможность интерактивного поиска в тематических базах данных Национальной медицинской библиотеки и Министерства образования США. При

этом некоторые из этих служб существуют и сегодня: основанная еще в 1965 году система Dialog (http://www.dialog.com/), входящая в настоящее время в корпорацию Thomson, сегодня обеспечивает своим клиентам доступ к сотням базам данных.

В начале 1990-х годов для унификации информационных систем был разработан международный стандарт Z39.50 - информационно-поисковый протокол для библиографических систем. В 1994 университет Джорджии запустил пилотный проект "Галилей" (http://www.usg.edu/galileo/) с использованием Site-Search - пакета программ Огайского центра, соответствующий стандарту Z39.50. Стандарт Z39.50 также был положен в основу исторически первой службы поиска распределенной информации в Интернет - WAIS (Wide Area Information Service) [127], в настоящее время уже утратившей свою актуальность.

В настоящее время информационные ресурсы только веб-пространства составляют свыше двадцати миллиардов документов, к которым возможен свободный доступ любого пользователя. Естественно, для того, чтобы найти необходимую информацию и этой крупнейшей распределенной полнотекстовой базе данных необходимо использовать самые мощные ИПС. Такие системы существуют и конкурируют друг с другом. Сегодня миллионам пользователей Интернет известны такие информационно-поисковые системы, как Google, Yahoo, AltaVista, AllTheWeb, MSN, Яndex, Rambler, которые охватывают миллиарды веб-документов. В основу работы всех подобных систем положены специальные алгоритмы, являющиеся модификациями основных подходов - моделей поиска [68].

В основу традиционных методов положены три главных подхода, первый из которых базируется на теории множеств (булева модель), второй - на векторной алгебре (векторно-пространственная модель), а третий - на теории вероятностей (вероятностная модель). Эти подходы могут применяться на практике и в каноническом виде, однако у них есть общий недостаток, обусловленный предположением, что содержание документа определяется множеством слов и устойчивых словосочетаний – термов (англ. - Terms), которые входят в него без

учета взаимосвязей, как «мешок со словами» (от англ. Bag of Words), и, более того, считаются независимыми. Конечно же, такое предположение ведет к потере содержательных оттенков, тем не менее оно позволяет реализовать поиск и группирование документов по формальным признакам. Известны такие основные недостатки традиционных моделей:

- Булева модель - невысокая эффективность поиска, отсутствие контекстных операторов, невозможность ранжирования результатов поиска.
- Векторно-пространственная модель связана с расчетом массивов высокой размерности и в каноническом виде малопригодна для обработки больших массивов данных.
- Вероятностная модель характеризуется низкой вычислительной масштабируемостью (т.е. резким снижением эффективности при росте объемов данных) , необходимостью постоянного обучения системы.

Системы, построеные на “рафинированных” поисковых моделях, недостаточно оперативны и обладают слабо развитыми поисковыми возможностями и средствами обобщения данных.

Кроме представленных ниже, существуют и другие модели поиска, например, семантические, в рамках которых делаются попытки организации смыслового поиска за счет анализа грамматики текста, использования баз знаний, тезаурусов, онтологий, которые реализуют семантические связи между отдельными словами и их группами. Вместе с тем, эффективноть систем, базирующихся на таких подходах пока, остается невысокой.

Перед рассмотрением отдельных моделей сформулируем некоторые допущения и понятия.

Пусть i - индекс терма t_i из словаря T $(i = 1,...,M)$, $d^{(j)}$ - документ, принадлежащий множеству документов D, а $w_i^{(j)} \geq 0$ - вес, ассоциированный с парой $\left(t_i, d^{(j)}\right)$.

Для каждого терма t_i, который не входит в документ $d^{(j)}$, его вес равен нулю: $w_i^{(j)} = 0$. Документ $d^{(j)}$ будем рассматривать как вектор $d^{(j)} = \left(w_1^{(j)}, w_2^{(j)}, ..., w_M^{(j)}\right)$.

Введем также в рассмотрение инверсную функцию g_i, соответствующую индексу терма t_i, которая определяется следующим образом: $g_i(d^{(j)}) = w_i^{(j)}$.

2.1. Булева модель поиска

2.1.1. Классическая булева модель

Булева модель базируется на теории множеств и математической логике. Популярность этой модели связана прежде всего с простотой ее реализации, которая позволяет индексировать и выполнять поиск в больших документальных массивах.

В рамках булевой модели документы и запросы представляются в виде множества термов - ключевых слов и устойчивых словосочетаний. Каждый терм представлен как булева переменная: 0 (терм из запроса не присутствует в документе) или 1 (терм из запроса присутствует в документе). При этом весовые значения терма в документе принимает лишь два значения: $w_i^{(j)} \in \{0,1\}$.

В булевой модели запрос пользователя представляет собой логическое выражение, в котором термы связываются логическими операторами конъюнкции (AND, $\wedge$), дизъюнкции (OR, $\vee$) и отрицания (NOT, $\neg$). Известно, что любое логическое выражение можно представить дизъюнкцией некоторых выражений, соединенных между собой операцией конъюнкции (дизъюнктивной нормальной формой, ДНФ - dnf).

Покажем, как на практике запрос, составленный из логических операторов и скобок, определяющих приоритет операторов, приводится к ДНФ. Рассмотрим пример из [68] – запрос: $q = a \wedge (b \vee \neg c)$. Если сопоставить множествам документов, содержащих термы из запроса a, b и c соответствующие диаграммы Эйлера (рис. 3), то легко можно видеть, что исходный запрос

эквивалентен запросу $(a \wedge b \wedge c) \vee (a \wedge b \wedge \neg c) \vee (a \wedge \neg b \wedge \neg c)$, т.е. дизъюнктивная нормальная форма запроса примет вид: $q_{dnf} = (1,1,1) \vee (1,1,0) \vee (1,0,0)$, где каждая из двоичных компонент ассоциируется с a, b или c (или их отрицаниями), а запятая между двоичными компонентами – с операциями конъюнкции.

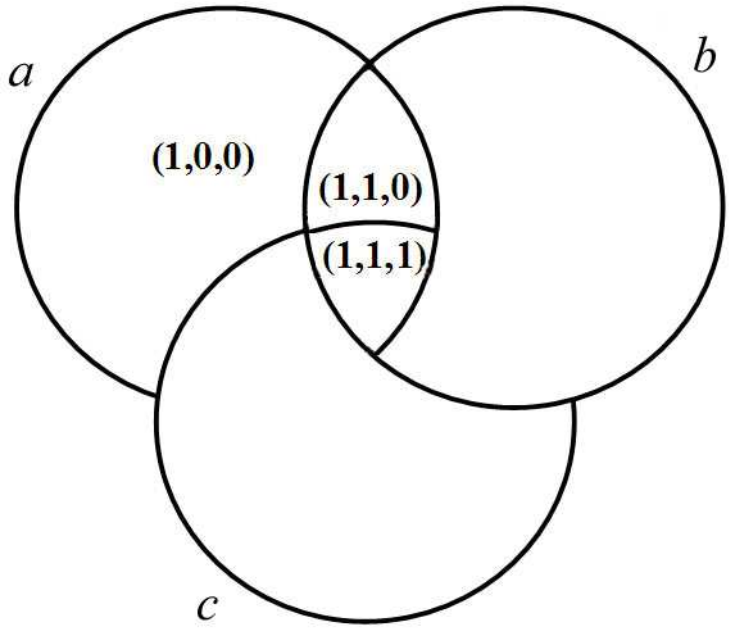


Рис. 3. Три конъюнктивных компоненты запроса $q = a \wedge (b \vee \neg c)$

В булевой модели запрос – это булево выражение. Так как оно приводится к дизъюнктивной нормальной форме, то можно записать:

$$q \equiv q_{dnf} = \bigvee_{i=1,..,N} q_{cc}^{(i)},$$

где $q_{cc}^{(i)}$ - i-я конъюнктивная компонента формы запроса q_{dnf}. Тогда мера близости документа $d^{(j)}$ и запроса q - $sim(d^{(j)}, q)$ (от англ. – similarity, близость) в булевой модели определяется выражением:

$$sim(d^{(j)}, q) = \begin{cases} 1, & \text{если } \exists q_{cc}^{(i)} : \ (q_{cc}^{(i)} \in q_{dnf}) \wedge (\forall k, \ g_k(q_{cc}^{(i)}) = g_k(d^{(j)})), \\ 0, & \text{иначе.} \end{cases}$$

То есть $sim(d^{(j)}, q)$ принимает значение 1, если существует такая конъюнктивная компонента $q_{cc}^{(i)}$, входящая в дизъюнктивную нормальную форму q_{dnf}, что инверсная функция каждого терма k данной конъюнктивной

компоненты совпадает с этой же инверсной функцией для документа $d^{(j)}$. В противном случае $sim(d^{(j)}, q)$ оказывается равной 0.

Таким образом, если $sim(d^{(j)}, q) = 1$, то в соответствии с булевой моделью документ $d^{(j)}$ считается релевантным (соответствующим) запросу q. В противном случае документ не является релевантным. Весовых различий, необходимых для ранжирования документов по уровню соответствия запросу в булевой модели не предусмотрено, что является существенным недостатком данной модели. Ниже будет рассмотрена расширенная булева модель, в которой этот недостаток преодолен.

Существует несколько подходов к формированию архитектуры поисковых систем, соответствующих булевой модели и нашедших свое воплощение в реальных информационно-поисковых системах. Одной из реализаций такой модели была некогда популярная система STAIRS корпорации IBM. База данных этой, уже ставшей классической, системы состоит из следующих основных таблиц:

- текстовой, содержащей текстовую часть всех документов;
- указателей текстов, которая включает указатели на местонахождение документов в текстовой таблице;
- словарной, содержащей все уникальные слова, встречающиеся в документах, то есть те слова, по которым может осуществляться поиск;
- инверсной, содержащей списки номеров документов и координаты отдельных слов в документах.

Поиск по слову в базе данных системы такой архитектуры осуществляется в соответствии с алгоритмом:

1. Происходит обращение к словарной таблице, по которой определяется, входит ли слово в состав словаря базы данных, и если входит, то определяется ссылка в инверсной таблице на цепочку появлений этого слова в документах.

2. Происходит обращение к инверсной таблице, по которой определяются номера документов, содержащих данное слово, и координаты всех вхождений слова в текстах базы данных.

3. По номеру документа происходит обращение к записи таблицы указателей текстов. Каждая запись этого файла соответствует одному документу в базе данных.

4. По номеру документа происходит прямое обращение к фрагменту текстовой таблицы – документу, после чего следует вывод найденного документа.

Приведенный алгоритм охватывает случай, когда запрос состоит из одного слова. Если же в запрос входит не одно слово, а некоторая их комбинация, то в результате выполнения поиска по каждому из этих слов запроса формируется массив записей, которые соответствуют вхождению этого слова в базу данных. После окончания формирования массивов результатов поиска происходит выявление релевантных документов путем выполнения теоретико-множественных операций над записями этих массивов в соответствии с правилами булевой логики.

2.1.2. Расширенная булева модель

Недостаток классической булевой модели связан с отсутствием весовых значений термов, а значит и нивелированием значимости отдельных термов. Это приводит к невозможности ранжирования результатов поиска по уровню их соответствия информационным запросам.

Для того чтобы устранить этот недостаток и вместе с тем использовать вычислительные преимущества булевой модели, Г. Солтоном (G. Salton), Э.А. Фоксом (E. Fox) и Г. Ву (H.Wu) в 1983 году была предложена расширенная булева модель [130].

В соответствии с этой моделью, каждому терму приписывается вес - значение из интервала [0, 1]. Пусть D - массив документов, в документ $\vec{d} \in D$ входят термы x и y. Тогда парам $\left[x, \vec{d}\right]$ и $\left[y, \vec{d}\right]$ ставятся в соответствие весовые значения $\hat{x}$ и $\hat{y}$. Эти значения могут быть определены, например, по формуле:

$$\hat{x} = f_x \frac{idf_x}{\max\ idf}, \quad \hat{y} = f_y \frac{idf_y}{\max\ idf},$$

где f_x - нормализованная частота терма x в документе $\vec{d}$, а idf_x - величина, обратная нормированному количеству документов во всем массиве (инверсная частота), содержащих терм x.

Герхард Солтон (1927-1995)

Вместо документа в рамках модели рассматривается вектор из термов (в простейшем случае – двух) $\vec{d} = (\hat{x}, \hat{y})$, который также можно рассматривать как точку в квадрате [0, 1] x [0, 1]. Близость между документами можно интерпретировать как некоторое нормированное расстояние между соответствующими точками.

В соответствии с расширенной булевой моделью вводятся меры близости между документом и запросом для двух типов запросов: $q_{or} = x \vee y$ и $q_{and} = x \wedge y$ следующим образом:

$$sim(q_{or}, d) = \sqrt{\frac{\hat{x}^2 + \hat{y}^2}{2}};$$

$$sim(q_{and}, d) = 1 - \sqrt{\frac{(1-\hat{x})^2 + (1-\hat{y})^2}{2}}.$$

Если еще более расширить модель и предположить, что соответствующие термы могут иметь вес в запросе, соответственно, *a* и *b*, то определяется:

$$sim(q_{or},d)=\sqrt{\frac{a^2\hat{x}^2+b^2\hat{y}^2}{a^2+b^2}};$$

$$sim(q_{and},d)=1-\sqrt{\frac{a^2(1-\hat{x})^2+b^2(1-\hat{y})^2}{a^2+b^2}}.$$

В рассмотренных выше случаях использовалась модель степени 2, которая легко обобщается до модели степени p, где $1\le p<\infty$. В этом случае предполагается, что запросы состоят из m термов, соединенных обобщенными операторами дизъюнкции ($\vee^p$) и конъюнкции ($\wedge^p$) степени p: $q_{or}=x_1\vee^p x_2\vee^p ...\vee^p x_m$ и $q_{and}=x_1\wedge^p x_2\wedge^p ...\wedge^p x_m$. При этом вводится такое определение близости этих запросов и документа:

$$sim(q_{or},d)=\left(\frac{\hat{x}_1^p+\hat{x}_2^p+...+\hat{x}_m^p}{m}\right)^{\frac{1}{p}};$$

$$sim(q_{and},d)=1-\left(\frac{(1-\hat{x}_1)^p+(1-\hat{x}_2)^p+...+(1-\hat{x}_m)^p}{m}\right)^{\frac{1}{p}}.$$

Очевидно, что для $p=1$ справедливо:

$$sim(q_{or},d)=sim(q_{and},d)=\frac{\hat{x}_1+\hat{x}_2+...+\hat{x}_m}{m}.$$

При $p\to\infty$ будет выполняться:

$sim(q_{or},d)\to\max(\hat{x}_i)$;

$sim(q_{and},d)\to\min(\hat{x}_i)$.

Если запрос является более сложным, то близость документа и запроса может быть вычислена путем использования двух основных правил (для дизъюнкции и конъюнкции).

Например, в случае, если запрос представляет собой выражение, $q=(x_1\vee^p x_2)\wedge^p x_3$, то выполняется:

$$sim(q,d)=1-\left(\frac{1}{2}\left(\left(1-\left(\frac{1}{2}\left(\hat{x}_1^p+\hat{x}_2^p\right)\right)^{\frac{1}{p}}\right)^p+(1-\hat{x}_3)^p\right)\right)^{\frac{1}{p}}.$$

2.1.3. Модель нечеткого поиска

Теория нечетких множеств (Fuzzy set theory), предложенная Л.А. Заде (L.A. Zade), расширяющая классическую теорию множеств, основывается на той идее, что функция принадлежности элемента множеству может принимать произвольные значения в интервале [0, 1], а не только 0 или 1 [18].

Лотфи А. Заде

По определению, нечетким множеством (fuzzy set) A на универсальном множестве U (любой природы) является совокупность пар $(\mu_A(U),U)$, где $\mu_A(U)$ – степень принадлежности элемента $u \in U$ нечеткому множеству A. Степень принадлежности - это число из диапазона [0, 1]. Чем выше степень принадлежности, тем в большей мере элемент $u \in U$ соответствует свойствам нечеткого множества. Функцией принадлежности называется функция, которая позволяет вычислить степень принадлежности произвольного элемента универсального множества к нечеткому множеству.

Пусть A и B - два нечетких множества над универсальным множеством U, а $\bar{A}$ – дополнение A до U и $u \in U$. Функцию принадлежности μ можно определить, например, следующим образом:

$$\mu_{\bar{A}}(u) = 1 - \mu_A(u);$$

$$\mu_{A \cup B}(u) = \max(\mu_A(u), \mu_B(u));$$

$$\mu_{A \cap B}(u) = \min(\mu_A(u), \mu_B(u)).$$

Л. Заде ввел также понятия носителя, высоты и точки перехода нечеткого множества [18]. Носителем нечеткого множества A называется множество таких точек $u \in U$, для которых величина $\mu_A(u)$ положительна. Высотой нечеткого множества A называется величина $\sup_{u \in U} \mu_A(u)$. Точкой перехода нечеткого множества A называется такой элемент множества U, степень принадлежности которого множеству A равна 1/2.

В качестве примера, также предложенного в [18], рассматривается универсальное множество U, представляющее собой интервал $[0, 100]$, и переменная u, принимающая значения из этого интервала, интерпретируемая как «возраст». При этом нечеткое подмножество универсального множества U, обозначаемое термином «старый», можно задать функцией принадлежности вида (рис. 4):

$$\mu_A(u) = \begin{cases} 0, & (0 \le u \le 50), \\ \left(1 + \left(\dfrac{u - 50}{5}\right)^{-2}\right)^{-1}, & (50 \le u \le 100). \end{cases}$$

В этом примере носителем нечеткого множества «старый» является интервал [50, 100], высота множества «старый» близка к 1, а точкой перехода является значение $u = 55$ (рис. 4). Таким образом, говоря простым языком, если возраст человека составляет 55 лет, то его соответствие понятию «старый» составляет 0.5, а если 70, то близко к 0.9. Человека же моложе 50 лет ($\mu_A(u) = 0$) вообще нельзя назвать старым. Заметим, что функция принадлежности выбиралась из субъективных представлений о том, с какого возраста начинается старость.

В модели нечеткого поиска каждый терм запроса рассматривается как нечеткое множество, а каждый документ рассматривается по степени принадлежности этому множеству. При этом рассматривается матрица

взаимосвязей термов $\hat{c}$, элемент которой c_{il} определяет взаимосвязь между термами с индексами i и l:

$$c_{il} = \frac{n_{il}}{n_i + n_l - n_{il}},$$

где n_i - количество документов, содержащих терм t_i, n_l - количество документов, содержащих терм t_l, а n_{il} - количество документов, содержащих оба терма.

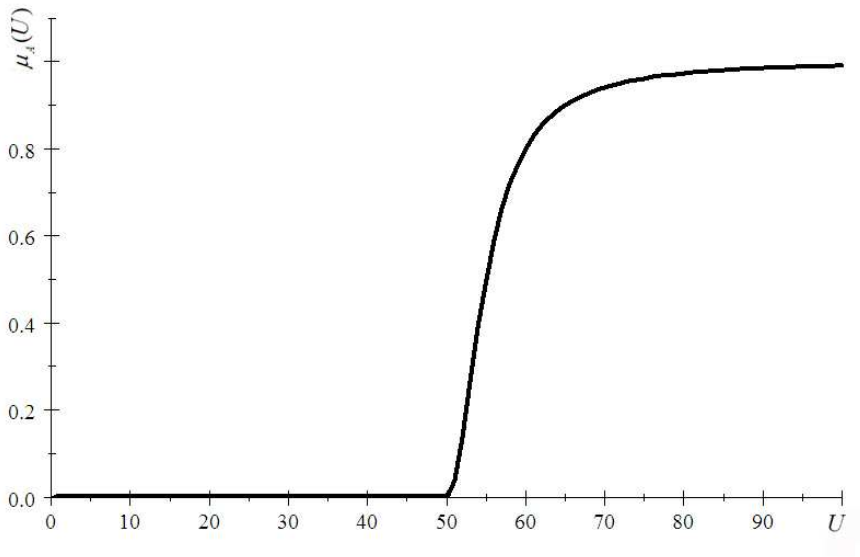


Рис. 4. Функция принадлежности нечеткого множества, определяемого значением «старый»

В рамках модели нечетких множеств, документ d_j обладает степенью принадлежности μ_{ij}, рассчитываемой следующим образом:

$$\mu_{ij} = 1 - \prod_{x_l \in d_j} (1 - c_{il}).$$

Документ $d^{(j)}$ ассоциируется с $i-$м термом t_i, если его собственные термы ассоциируются с t_i. Если хотя бы один терм t_l из $d^{(j)}$ строго ассоциируется с термом с индексом t_i (т.е. $c_{il} \sim 1$), то $\mu_{ij} \sim 1$ и терм t_i является «хорошим» нечетким индексом документа $d^{(j)}$. Если же это не так и можно считать, что $c_{il} << 1$, то μ_{ij} заменяется алгебраической суммой c_{il} (что вполне оправдано для большого словаря). Кроме того, вместо функции max при вычислении функции принадлежности в случае реализации оператора дизъюнкции между разными

термами также применяется суммирование, что также является некоторым огрублением изначально определенной модели. Для вычисления конъюнкции, соответственно, используется произведение элементов c_{il}.

Для поиска в модели нечетких множеств используется запрос, подобный обычному булевскому выражению. Точно так же, как и в случае булевой логики, запрос может быть представлен в дизъюнктивной нормальной форме, а именно:

$$q_{dnf} = cc_1 \vee cc_2 \vee ... \vee cc_p,$$

где cc_i - i-я конъюнктивная компонента, p - количество конъюнктивных компонент q_{dnf}.

Соответственно, функция принадлежности μ_{qj} документа $d^{(j)}$ нечеткому множеству D_q, соответствующему запросу q (и рассматриваемая как параметр для ранжирования релевантных документов), вычисляется следующим образом:

$$\mu_{qj} = 1 - \prod_{i=1}^{p} \left(1 - \mu_{cc_i j}\right).$$

2.2. Векторно-пространственная модель поиска

Многие из известных информационно-поисковых систем базируются на векторно-пространственной модели описания данных (Vector Space Model), предложенной Г. Солтоном в 1975 г. и впервые примененной в системе SMART [131]. Данная модель является классической алгебраической. В рамках этой модели документ описывается вектором в евклидовом пространстве, в котором каждому терму, использующемуся в документе, ставится в соответствие его весовое значение, определяемое на основе статистической информации о его появлении как в отдельном документе, так и во всем документальном массиве. Описание запроса, соответствующего необходимой пользователю тематике, также представляет собой вектор в том же евклидовом пространстве термов. Для оценки близости запроса и документа используется скалярное произведение соответствующих векторов запроса и документа.

В рамках этой модели каждому терму t_i в документе $d^{(j)}$ соответствует некоторый неотрицательный вес $w_i^{(j)}$.

Каждому запросу q, который представляет собой также множество термов, не соединенных между собой никакими логическими операторами, также соответствует вектор весовых значений w_i^q.

Таким образом, каждый документ и запрос могут быть представлены в виде $n-$мерного вектора, где n – общее количество термов в словаре модели. В соответствии с рассматриваемой моделью, близость документа $d^{(j)}$ к запросу q, которые как и в предыдущих моделях рассматриваются как информационные векторы $\vec{d}_j = (w_1^{(j)}, w_2^{(j)}, ..., w_n^{(j)})$ и $\vec{q} = (w_1^q, w_2^q, ..., w_n^q)$ оценивается как их скалярное произведение. При этом вес отдельных термов можно вычислять разными способами. Один из возможных простейших подходов - использовать как вес терма $w_i^{(j)}$ в документе нормализованную частоту $freq_i^{(j)}$ его встречаемости в данном документе, то есть:

$$w_i^{(j)} = freq_i^{(j)} / \max_{1 \le k \le n} freq_k^{(j)}.$$

Вычисленный таким образом вес терма в документе принято обозначать аббревиатурой $tf_i^{(j)}$ или просто *TF* (от англ. Term Frequency – частота термина).

Однако этот подход не учитывает, насколько часто рассматриваемый терм используется во всем массиве документов, так называемую, дискриминационную силу терма. Поэтому в случае, когда доступна статистика использования термов во всем документальном массиве, более эффективно следующее правило вычисления веса:

$$w_i^{(j)} = tf_i^{(j)} \cdot \log \frac{N}{n_i},$$

где n_i - количество документов, в которых используется терм t_i, а $N-$ общее количество документов в массиве. Например, если некоторое слово встречается в каждом документе массива, то его использование в запросе, очевидно,

бесполезно. Соответственно, в этом случае $n_i = N$, и следовательно, $w_i^{(j)} = tf_i^{(j)} \cdot \log\frac{N}{N} = 0.$

Следует отметить, что приведенная выше формула многократно уточнялась с целью наиболее точного соответствия выдаваемых документов запросам пользователей. В 1988 году Солтоном был предложен такой вариант для вычисления веса терма t_i из запроса в документе:

$$w_i^q = \left(0.5 + \frac{freq_i^q}{\max\limits_{1 \le l \le n} freq_l^q}\right) \cdot \log\frac{N}{n_i},$$

где $freq_i^q$ - частота терма t_i из запроса в тексте этого документа.

Обычно весовые значения $w_i^{(j)}$ нормируются путем деления на их общую сумму. Такой метод взвешивания термов имеет стандартное обозначение - $TF \cdot IDF$, где TF указывает на частоту появления терма в документе, а $IDF -$ на величину, обратную количеству документов в массиве, содержащих данный терм (от англ. - inverse document frequency).

Когда возникает задача определения тематической близости двух документов или документа и запроса, в этой модели используется простое скалярное произведение $sim(d^{(1)}, d^{(2)})$, двух соответствующих векторов весовых значений $(w_1^{(1)}, w_2^{(1)}, ..., w_n^{(1)})$ и $(w_1^{(2)}, w_2^{(2)}, ..., w_n^{(2)})$ которое соответствует косинусу угла между векторами - образами документов $d^{(1)}$ и $d^{(2)}$. Очевидно, $sim(d^{(1)}, d^{(2)})$ принадлежит диапазону [0, 1]. Чем больше величина $sim(d^{(1)}, d^{(2)})$ - тем более близки документы $d^{(1)}$ и $d^{(2)}$. Для любого документа d имеем $sim(d, d) = 1$. Аналогично мерой близости документа $d^{(j)}$ и запроса q является величина:

$$sim(d_j, q) = \frac{\vec{d}^{(j)} \cdot \vec{q}}{|\vec{d}^{(j)}| \cdot |\vec{q}|} = \frac{\sum\limits_{i=1}^{n} w_i^{(j)} w_i^q}{\sqrt{\sum\limits_{i=1}^{n} (w_i^{(j)})^2} \sqrt{\sum\limits_{i=1}^{n} (w_i^q)^2}}.$$

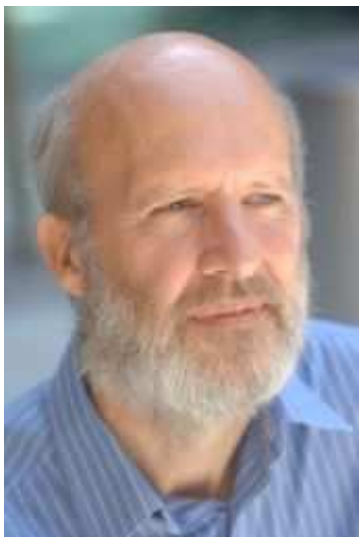
Векторно-пространственная модель представления данных обеспечивает системам, построенным на ее основе, такие возможности, как:

- обработку запросов без ограничений их длины;
- простоту реализации режима поиска подобных документов (каждый документ может расматриваться как запрос);
- сохранение результатов поиска с возможностью выполнения уточняющего поиска.

Вместе с тем в векторно-пространственной модели не предусмотрено использование логических операций в запросах, что существенно ограничивает ее применимость.

2.3. Вероятностная модель поиска

В 1977 году С. Э. Робертсон (S.E. Robertson) и К. Спарк-Джонс (K. Sparck Jones) обосновали и реализовали вероятностную модель, предложенную еще в 1960 году. В данной модели поиска вероятность того, что документ релевантен запросу основывается на предположении, что термы запроса по-разному распределены среди релевантных и нерелевантных документов. При этом используются формулы расчета вероятности, базирующиеся на теореме Байеса.



С. Робертсон (Microsoft Research Laboratory)

Основной вопрос, который решается с помощью модели: как велика вероятность того, что документ d релевантен запросу q? Релевантность при этом рассматривается как вероятность того, что данный документ может оказаться

интересным пользователю. Функционирование модели базируется как на экспертных оценках, получаемых в результате обучения модели, которые признают документы из учебной коллекции релевантными/нерелевантными, так и на последующих оценках вероятности того, что документ является релевантным запросу исходя из состава его термов.

Если для запроса известны эти оценки вероятностей для всех документов, то документы можно сортировать по ним и выводить пользователям в нисходящем порядке. То есть вероятностная модель поиска предусматривает определение вероятностей соответствия запросу для документов, сортировку и предоставление документов с ненулевой вероятностью пользователю.

С самого начала в вероятностной модели использовалось упрощение, которое допускает независимость вхождения в документ любой пары термов (поэтому такой подход называется «наивным» байесовским).

При этом в вероятностной модели поиска предполагается наличие учебных наборов релевантных и нерелевантных документов, выбранных пользователем или полученных автоматически при каком-то начальном предположении. Вероятность того, что поступивший документ является релевантным, рассчитывается на основании соотношения появления термов в релевантном и нерелевантном массиве документов.

В случае применения экспертных оценок процесс поиска является итерационным (в реальных системах, использующих элементы вероятностой модели, как экспертные оценки могут рассматриваться, например, предпочтения пользователейпри выборе интересующих их документов). На каждом шаге итерации, благодаря режиму обратной связи, определяется множество документов, отмеченных пользователем как удовлетворяющих его информационным потребностям.

Рассмотрим основу модели, а именно байесовский подход, более детально. Пусть X, $Y-$ два независимых события, X, $Y \subset G$, $G-$базовое вероятностное пространство.

Вероятность X при условии Y определяется таким образом:

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}.$$

Известно, что из этого соотношения следует формула Байеса:

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}; \;\; P(Y \mid X) = \frac{P(Y \cap X)}{P(X)};$$

$$P(Y \cap X) = P(X \cap Y) \Rightarrow P(X \mid Y) = \frac{P(Y \mid X) \cdot P(X)}{P(Y)}.$$

Рассмотрим условные вероятности двух событий, а именно того, что документ релевантен (R) запросу - $P(R \mid q,d)$, где $q-$ запрос, $d-$документ, а также того, что документ нерелевантен ($\overline{R}$) запросу - $P(\overline{R} \mid q,d)$.

В рамках вероятностной модели вводится понятие квоты релевантности как меры близости документа запросу - $O(R)$:

$$O(R) = \frac{P(R)}{P(\overline{R})} = \frac{P(R)}{1 - P(R)}.$$

Очевидно, что квота меньше, чем 1 для вероятности $P(R) < 0.5$ и больше 1 для вероятности $P(R) > 0.5$.

Определим квоту того события, что документ релевантен запросу:

$$O\left(R \mid q,d\right) = \frac{P(R \mid q,d)}{P(\overline{R} \mid q,d)}.$$

Для числителя этой формулы справедливо:

$$P(R \mid d,q) = \frac{P(R \cap q \cap d)}{P(q \cap d)}.$$

Величина $P(R \cap q \cap d)$ в приведенном выражении интерпретируется как вероятность события, заключающегося в том, что документ d релевантен запросу q, а величина $P(q \cap d)$ - вероятность того, что по запросу q будет выдан документ d. аюИспользуя формулу Байеса для числителя и знаменателя получаем:

$$P(R \mid d,q) = \frac{P(d \mid R \cap q) \cdot P(R \cap q)}{P(d \mid q) \cdot P(q)} = \frac{P(d \mid R \cap q) \cdot P(R \mid q)}{P(d \mid q)}.$$

Подставляя выражения $P(R \mid d,q)$ и $P(\overline{R} \mid d,q)$ в числитель и знаменатель формулы для квоты релевантности, получаем:

$$O(R \mid d,q) = \frac{\dfrac{p(d \mid R \cap q) \cdot p(R \mid q)}{p(d \mid q)}}{\dfrac{p(d \mid \bar{R} \cap q) \cdot p(\bar{R} \mid q)}{p(d \mid q)}} = \frac{p(R \mid q)}{p(\bar{R} \mid q)} \times \frac{p(d \mid R \cap q)}{p(d \mid \bar{R} \cap q)}.$$

Перейдем к рассмотрению документа как вектора термов. Пусть $T = \{t_1, \ ..., \ t_n\}$ – множество термов, которые содержатся в корпусе документов D. Документ рассматривается как вектор из бинарных значений весов входящих в него термов $\vec{d} = (w_1, ..., w_n)$, где:

$$w_i = \begin{cases} 1, & t_i \in d; \\ 0, & t_i \notin d. \end{cases}$$

Тогда, предполагая независимость термов в рамках рассматриваемой «наивной» байесовской модели, получаем:

$$p(d \mid R,q) = p(\vec{d} \mid R,q) = \prod_{i=1}^{n} p(t_i \mid R,q).$$

В результате квота релевантности принимает вид:

$$O(R \mid q,d) = \frac{p(R \mid q)}{p(\bar{R} \mid q)} \cdot \frac{p(d \mid R \cap q)}{p(d \mid \bar{R} \cap q)} = O(R \mid q) \cdot \prod_{i=1}^{n} \frac{p(t_i \mid R \cap q)}{p(t_i \mid \bar{R} \cap q)}.$$

Здесь $O(R \mid q)$ - квота релевантности для запроса без учета документов. Модель предусматривает еще одно упрощение, а именно то, что для термов, не входящих в запросы (для $t_i \in T \setminus q$), предполагается одинаковая вероятность их появления в релевантных и нерелевантных документах, т.е.:

$t_i \in T \setminus q : p(t_i \mid R,q) = p(t_i \mid \bar{R},q).$

Разложим произведение в формуле квоты релевантности следующим образом:

$$O(R \mid q,d) = O(R \mid q) \cdot \prod_{t_i \in q \cap d} \frac{p(t_i \mid R \cap q)}{p(t_i \mid \bar{R} \cap q)} \cdot \prod_{t_i \in q \setminus d} \frac{p(t_i \mid R \cap q)}{p(t_i \mid \bar{R} \cap q)} \cdot \prod_{t_i \notin q} \frac{p(t_i \mid R \cap q)}{p(t_i \mid \bar{R} \cap q)}.$$

В приведенных обозначениях под знаком произведения $q \cap d$ означает множество общих термов в запросе и документе, $q \setminus d$ - множество слов, входящих в запрос, но отсутствующих в документе, q - множество слов, входящих в запрос.

Последний сомножитель равен единице ввиду вышеприведенного предположения. Введем обозначения для вероятностей того, что слово присутствует в документе, при условии того, что документ релевантен или нерелевантен запросу:

$$r_i = p(w_i = 1 | R, q);$$
$$n_i = p(w_i = 1 | \bar{R}, q).$$

В этих обозначениях выполняется:

$$O(R | q, d) = O(R | q) \cdot \prod_{t_i \in q \cap d} \frac{r_i}{n_i} \cdot \prod_{t_i \in q \setminus d} \frac{1 - r_i}{1 - n_i}.$$

Учитывая то, что:

$$\prod_{t_i \in q \cap d} \frac{(1 - r_i)(1 - n_i)}{(1 - n_i)(1 - r_i)} = 1,$$

получаем:

$$O(R | q, d) = O(R | q) \cdot \prod_{t_i \in q \cap d} \frac{r_i (1 - n_i)}{n_i (1 - r_i)} \cdot \prod_{t_i \in q} \frac{1 - r_i}{1 - n_i}.$$

Для исследования релевантной последовательности элементов достаточно учитывать только второй сомножитель, так как только в нем присутствуют признаки, связанные с документом. При значение этого сомножителя можно прологарифмировать (логарифм - монотонная функция, которая не меняет рангов документов). То есть можно анализировать сумму:

$$\sum_{t_i \in q \cap d} \log \frac{r_i (1 - n_i)}{n_i (1 - r_i)} = \sum_{t_i \in q \cap d} \left[\log \frac{r_i}{n_i} + \log \frac{1 - n_i}{1 - r_i} \right].$$

Рассмотрим приближенные значения, полученные на основе анализа некоторой предварительно полученной учебной выборки:

$$\tilde{r}_i = \frac{rel_i}{rel}; \quad \tilde{n}_i = \frac{nrel_i}{nrel},$$

где rel_i – количество релевантных документов, которое содержит терм с индексом i; $nrel_i$ – соответственно, количество нерелевантных документов.

То есть можно анализировать сумму, называемую поисковым статусом:

$$SV = \sum_{t_i \in q \cap d} \log \frac{\tilde{r}_i (1-\tilde{n}_i)}{\tilde{n}_i (1-\tilde{r}_i)} = \sum_{t_i \in q \cap d} \log \frac{\frac{rel_i}{rel} \cdot \left(1 - \frac{nrel_i}{nrel}\right)}{\frac{nrel_i}{nrel} \cdot \left(1 - \frac{rel_i}{rel}\right)}.$$

Проведя элементарные преобразования, получаем:

$$SV = \sum_{t_i \in q \cap d} SV_i = \sum_{t_i \in q \cap d} \log \frac{rel_i (nrel - nrel_i)}{nrel_i (rel - rel_i)}.$$

В качестве примера рассмотрим массив документов, состоящий из двух частей: учебной выборки - документов $d^{(1)}$,..., $d^{(6)}$ (Табл. 1) и новых документов - $d^{(7)}$,..., $d^{(9)}$ (Табл. 2), для которых необходимо оценить уровень релевантности. Предположим, что запрос состоит из четырех термов - t_1, t_2, t_3, t_4 (соответственно, это та часть словаря, которая существенна для анализа). В таблице отдельным столбцом приведена некоторая экспертная оценка R релевантности для документов из учебной выборки.

Табл. 1

	t_1	t_2	t_3	t_4	R
$d^{(1)}$	1	0	0	1	1
$d^{(2)}$	1	1	0	1	1
$d^{(3)}$	0	1	1	0	1
$d^{(4)}$	0	0	1	1	0
$d^{(5)}$	0	0	1	1	0
$d^{(6)}$	1	1	0	0	0
rel_i	2	2	1	2	$rel = 3$
$nrel_i$	1	1	2	2	$nrel = 3$
$\exp(SV_i)$	4	4	1/4	1	

По этим данным рассчитываются значения rel_i и $nrel_i$, а также экспоненты от соответствующей составляющей статуса релевантности $\exp(SV_i)$, которые приведены в последних трех строках Табл. 1.

Предположим, что необходимо проанализировать новые документы $d^{(7)}$, $d^{(8)}$, $d^{(9)}$, встречаемость терминов t_1, t_2, t_3, t_4 для которых приведена в соответствующих ячейках Табл. 2. Для новых документов статус релевантности рассчитывается в соответствии с вышеприведенной формулой. Результаты, также

приведенные в Табл. 2, свидетельствуют, в частности, о значимом уровне релевантности документа $d^{(8)}$, рассчитанного в соответствии с вероятностной моделью.

Табл. 2

	t_1	t_2	t_3	t_4	Статус (SV)
$d^{(7)}$	0	1	0	1	$2 = \log 4 + \log 1$
$d^{(8)}$	1	1	0	0	$4 = \log 4 + \log 4$
$d^{(9)}$	1	0	1	1	$0 = \log 4 + \log 1/4 + \log 1$

2.4. Алгоритмы поиска в пиринговых сетях

Главная задача информационного поиска в децентрализованных пиринговых сетях (сетях P2P) – быстрое и эффективное нахождение наиболее релевантных откликов на запрос, передаваемый от узла ко всей сети. В частности, как всегда актуальна задача получения качественного результата при общем уменьшении сетевого трафика [104].

2.4.1. Алгоритм поиска ресурсов по ключам

В большинстве пиринговых сетей, ориентированных на обмен файлами, используется два вида сущностей, которым приписываются соответствующие идентификаторы (ID): узлы и ресурсы (например, файлы), характеризующиеся ключами (Key), т.е. сеть может быть представлена двумерной матрицей размерности MN, где M – количество узлов, N – количество ресурсов. В данном случае задача поиска сводится к нахождению ID узла, на котором хранится ключ ресурса. На рис. 4. представлен процесс поиска ресурса, запускаемый с узла с ID 0 [155, 156]. В данном случае с узла с ID 0 запускается поиск ресурса с ключом 14. Запрос проходит определенный маршрут и достигает узла, на котором находится ключ 14. Далее узел с ID 14 пересылает узлу с ID 0 адреса всех узлов, обладающих ресурсом, соответствующим ключу 14.

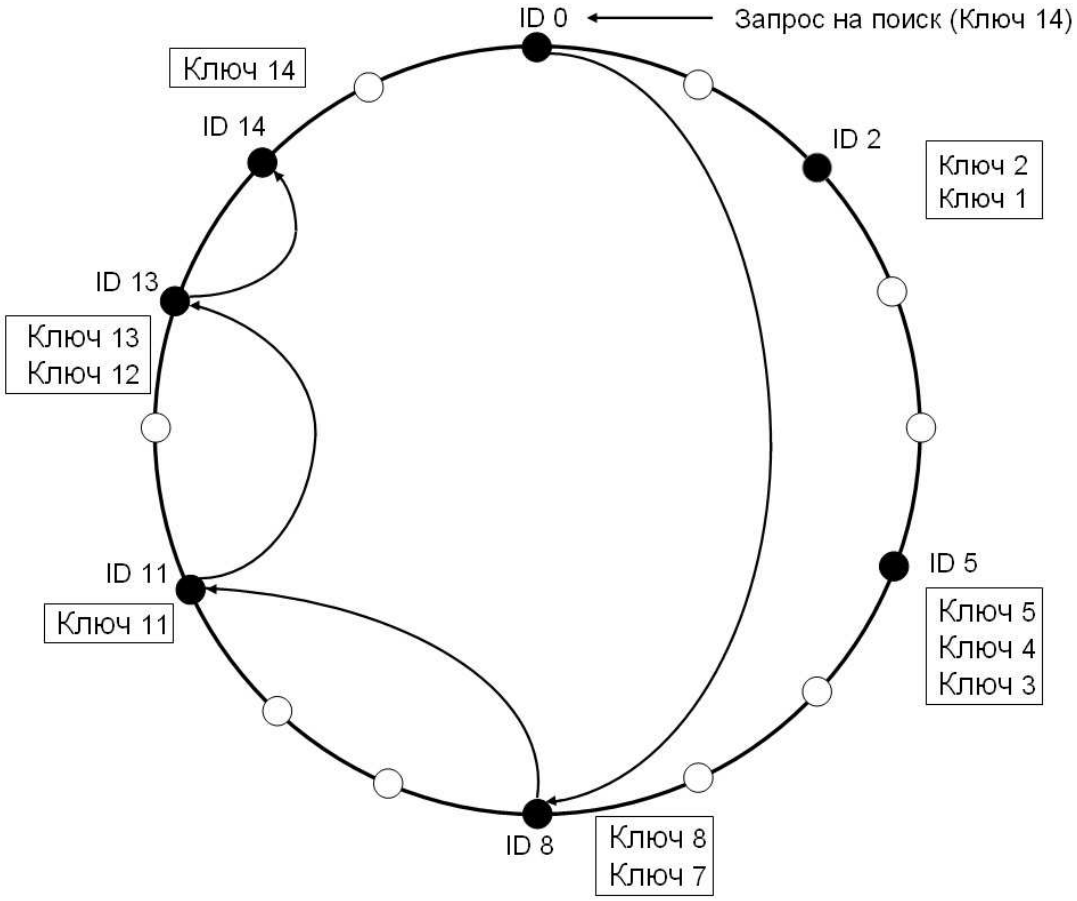


Рис. 4. Модель поиска ресурса по ключу

Рассмотрим алгоритмы поиска в пиринговых сетях, ограничившись основными методами поиска по термам.

2.4.2. Метод широкого первичного поиска

Метод широкого первичного поиска (Breadth First Search, BFS) [104] широко используется в реальных файлообменных сетях P2P, таких как, например, Gnutella (www.gnutella.com). Метод BFS (рис. 5) в сети P2P размерности N реализуется следующим образом. Узел q генерирует запрос, который адресуется ко всем соседям (ближайшим по некоторым критериям узлам). Когда узел p получает запрос, выполняется поиск в его локальном индексе. Если некоторый узел r принимает запрос (Query) и обрабатывает его, то он генерирует сообщение-отклик (QueryHit), чтобы возвратить результат. Сообщение QueryHit включает информацию о релевантных документах, которая доставляется по сети запрашивающему узлу (рис. 5).

Когда узел q получает QueryHits от более чем одного узла, он может загрузить файл с наиболее доступного ресурса. Сообщения QueryHit возвращаются тем же путем, что и первичный запрос.

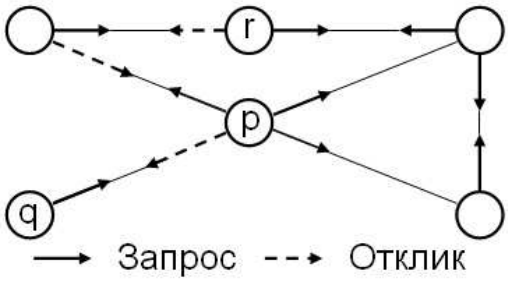


Рис. 5. Метод BFS

В BFS каждый запрос вызывает чрезмерную нагрузку сети, так как он передается по всем связям (в том числе и узлам с высоким временами ожидания). Поэтому узел с низкой пропускной способностью может стать узким местом. Одним из методов, позволяющий избежать перегрузки всей сети сообщениями заключается в приписывании каждому запросу параметра времени жизни (Time-to-live, TTL). Параметр TTL определяет максимальное число переходов, по которым можно пересылать запрос (очень важно, что именно число переходов, а не физическое время, измеряемое, например, в секундах). При типичном поиске начальное значение для TTL составляет 5-7 и уменьшается каждый раз, когда запрос пересылается. Когда TTL становится равным 0, сообщение больше не передается. BFS гарантирует высокий уровень качества поиска за счет большого числа переданых сообщений.

2.4.3. Метод случайного широкого первичного поиска

Метод случайного широкого первичного поиска (Random Breadth First Search, RBFS) был предложен как улучшение «наивного» подхода BFS [104]. В методе RBFS (рис. 6) узел q пересылает поисковое предписание только части узлов сети, выбранной в случайном порядке. Какая именно часть узлов – это параметр метода RBFS.

Преимущество RBFS заключается в том, что не требуется глобальной информации о состоянии контента сети; узел может получать локальные решения так быстро, как это потребуется. С другой стороны, этот метод вероятностный. Поэтому некоторые большие сегменты сети могут оказаться недостижимыми.

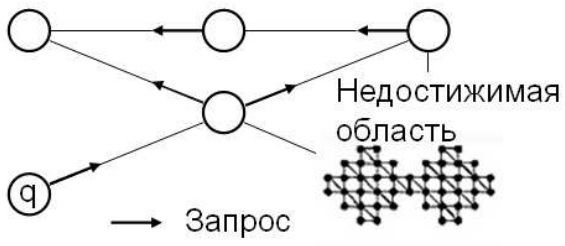


Рис. 6. Метод RBFS

2.4.4. Интеллектуальный поисковый механизм

Интеллектуальный поисковый механизм (Intelligent Search Mechanism, ISM) является относительно новым методом поиска в сетях P2P (рис. 7). Улучшение скорости и эффективности поиска информации с помощью данного метода достигается за счет минимизации затрат на связи, то есть на число сообщений, передающихся между узлами, и минимизации количества узлов, которые опрашиваются для каждого поискового запроса. Чтобы достичь этого, для каждого запроса оцениваются лишь те узлы, которые наиболее соответствуют данному запросу.

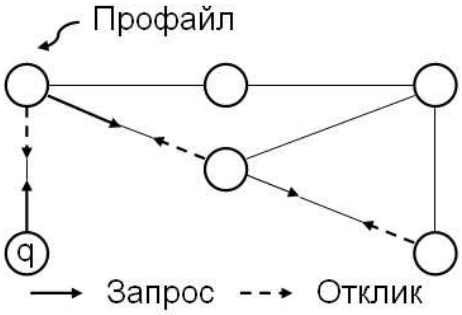


Рис. 7. Метод ISM

Интеллектуальный поисковый механизм состоит из двух компонент – профайла (profile) и способа его ранжирования, так называемого ранга

релевантности. Каждый узел сети строит информационный профайл для каждого из соседних узлов. Профайл содержит последние ответы каждого из узлов. С помощью ранга релевантности осуществляется ранжирование профайлов узлов для выбора тех соседних, которые будут давать наиболее релевантные документы по запросу. Механизм профайлов используется для того, чтобы сохранять последние запросы, а также количественные характеристике результатов поиска.

При реализации модели ISM используется единый стек запросов, в котором сохраняется по T запросов для N узлов. Как только стек заполняется, происходит замена «того запроса, который не использовался дольше всего» (Least Recently Used, LRU) для сохранения последних запросов. Функция «ранг релевантности» (Relevance Rank, RR) используется узлом P_l, чтобы выполнять оперативную классификацию его соседей для определения тех, которые следует опрашивать первыми по запросу q. Для вычисления ранга релевантности RR каждого узла P_i, P_l сравнивает q со всеми запросами в структуре профайла, для которого известен список ответов на предыдущие запросы, и вычисляет $RR(P_i, q)$:

$$RR(P_i, q) = \sum_{j \in Q} Sim(q_j, q)^{\alpha} \cdot S(P_i, q_j),$$

где α - параметр, задающий вес запросов. В этой формуле Q - множество запросов, на которые был ответ у узла P_i; $S(P_i, q_j)$ – количество результатов, возвращаемых узлом P_i по запросу q_j; метрика Sim рассчитывается по правилу, рассмотренному в векторно-пространственной модели поиска:

$$Sim(q_j, q) = \frac{q_j \cdot q}{|q_j||q|}.$$

Ранг релевантности RR обеспечивает более высокий ранг узла, который возвращает больше результатов. Кроме того, используется параметр α, который позволяет увеличивать вес запросов, наиболее подобных исходному. В случае, когда α большое, запросы с большим подобием $Qsim(q_j, q)$ доминируют в приведенной выше формуле. Рассмотрим ситуацию, когда узел P_1 соответствует запросам q_1 и q_2 со значениями подобия для запроса q: $Qsim(q_1, q) = 0.5$ и

$Qsim(q_2, q) = 0.1$, а узел P_2 соответствуют запросам q_3 и q_4 со значениями $Qsim(q_3, q) = 0.4$ и $Qsim(q_4, q) = 0.3$. Если выбрать $\alpha = 10$, то $Qsim(q_1, q)^{10}$ доминирует, так как $0.5^{10} + 0.1^{10} > 0.4^{10} + 0.3^{10}$.

Однако для $\alpha = 1$ все запросы весят одинаково, и P_2 дает более высокую релевантность. При $\alpha = 0$ учитывается только количество результатов, возвращенных каждым узлом.

Метод ISM эффективно работает в сетях, где узлы содержат некоторые специализированные сведения. В частности, исследование сети Gnutella показывает, что качество поиска очень зависит от «окружения» узла, с которого задается запрос. Большая проблема в методе ISM состоит в том, что поисковые сообщения могут циклично проходить одни и те же узлы сети, не достигая некоторых ее частей. Чтобы разрешить эту проблему для охвата большей части сети в [156] был описан подход, при котором для каждого запроса выбиралось небольшое подмножество случайных узлов, добавляляющееся к набору релевантных узлов.

Существуют и другие подходы решения этой проблемы, например, применяемый в протоколе BGP4 (RFC 1771), где каждый запрос сохраняет «историю» - список узлов, через которые он уже прошел.

2.4.5. Методы «большинства результатов по прошлой эвристике»

В [151, 152] был представлен метод «большинства результатов по прошлой эвристике» «>RES» (рис. 8), в котором каждый узел пересылал запрос подмножеству своих узлов, образованному на основании некоторой обобщенной статистики.

Запрос в методе >RES является удовлетворительным, если выдается Z или больше результатов (Z – некоторая постоянная). В методе >RES узел q пересылает запросы к k узлам, выдавшим наибольшие результаты для последних m запросов. В экспериментах k изменялось от 1 до 10 и таким путем метод >RES варьировался от BFS до метода, называемого глубинным первичным поиском (Depth-first-search).

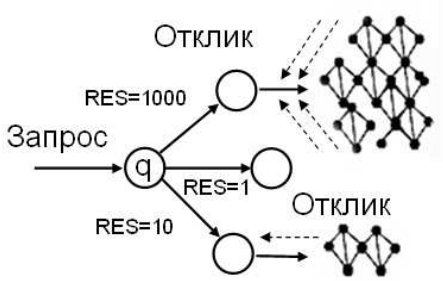


Рис. 8. Метод >RES

Метод >RES подобен методу ISM, но использует более простую информацию об узлах. Его главный недостаток по сравнению с ISM – отсутствие анализа параметров узлов, содержание которых связано с запросом. Поэтому метод >RES характеризуется скорее как количественный, а не качественный подход. Из опыта известно, что >RES хорош тем, что он маршрутизирует запросы в большие сегменты сети (которые возможно, также содержат более релевантные ответы). Он также захватывает соседей, которые менее перегружены, начиная с тех, которые обычно возвращают больше результатов.

2.4.6. Метод «случайных блужданий»

В [109] представлен алгоритм «случайных блужданий» (Random Walkers, Algorithm, RWA). Ключевая идея метода заключается в том, что каждый узел случайным образом пересылает сообщение с запросом, именуемое «посылкой» одному из соседних узлов. Чтобы сократить время, необходимое для получения результатов, вместо одной «посылки» рассматривается k независимых посылок, последовательно запущенных с поискового узла. Ожидается, что «k-посылок» после T шагов достигнет тех же результатов, что и одна посылка за kT шагов. Этот алгоритм напоминает метод RBFS, но в RBFS каждый узел пересылает сообщение запроса части соседей. К тому же, в RBFS предполагается экспоненциальное увеличение пересылаемых сообщений, а в методе случайных блужданий – линейное. Оба метода - и RBFS, и RWA не используют никаких явных правил для того чтобы адресовать поисковый запрос к наиболее релевантному содержанию.

Еще одной методикой, подобной RWA, является «адаптивный вероятностный поиск» (Adaptive Probabilistic Search, APS) [156]. В APS каждый узел развертывает на своих ресурсах локальный индекс, содержащий значения условных вероятностей для каждого соседа, который может быть выбран для обработки следующего запроса. Главное отличие от RWA в данном случае - это то, что в APS узел использует в качестве обратной связи результаты предыдущих поисков (в виде условных вероятностей) вместо полностью случайных переходов. Поэтому метод APS часто дает лучшие результаты, чем RWA.

2.5. Информационно-поисковые языки

Информационно-поисковые языки являются основными компонентами информационно-поисковых систем, с помощью которых, в частности, реализуются интерфейсы между пользователями и системами.

В отличие от реляционных СУБД, у систем полнотекстового поиска не существует стандартизированного языка запросов. У каждой системы этого типа существует свой способ задания критериев поиска.

Очень часто языки запросов ИПС приближенны к SQL, однако каждой из поисковых систем присущий ряд индивидуальных особенностей, связанных с такими моментами, как:

- интерпретация операций, задающих порядок расположения слов в тексте (операций контекстной близости);
- вычисление уровня релевантности найденных документов запросам для представления результатов поиска;
- применение нестандартных для реляционных СУБД функций, например, таких как нахождение документов по принципу подобия содержания, построение дайджестов из фрагментов документов, сниппетов (от англ. snippet – фрагмент, отрывок), включаемых поисковыми системами в списки найденных документов и т.п.

В различных полнотекстовых информационно-поисковых системах применяются различные архитектурные решения, охватывающие структуры

данных, алгоритмы их обработки, методы организации поиска. Вместе с тем, у современных информационно-поисковых систем много общих свойств, например, все из них обеспечивают поиск хотя бы по одному слову, большинство подобных систем реализуют грамматический поиск как результат применения лингвистического анализа (например, в русскоязычных системах Апорт, Яндекс и Рамблер по терму из запроса «человек» находятся не только словоизменения «человека», «человеку», но и множественное число – «люди»). Большинство из современных систем способны реализовывать контекстный поиск фразы, заключенной в кавычки (Google, Alltheweb, AltaVista, Яндекс и т.п.), поиск с использованием булевых операторов AND, OR и NOT, а также возможностью указания скобок для группирования термов и операторов. Функции контекстной близости в свое время получили наибольшее развитие в системе Lycos, где были реализованы с помощью четырех операторов: ADJ, NEAR, FAR и BEFORE.

В самой популярной в мире системе Google используется достаточно лаконичный набор операторов (http://www.googleguide.com/), основные из которых - это конъюнкция (подразумевается по умолчанию, система выдает документы, содержащие все слова запроса), дизъюнкция (OR) и отрицание (-).

Отдельно рассматривается возможность поиска по параметрам документов, которая чаще всего позволяет ограничивать диапазон поиска значениями URL, дат, заголовков. В большей части систем выйти на возможность поиска по параметрам можно из режима расширенного поиска.

В Google, например, обеспечивается поиск по сайту ("site:"), определение ссылок на сайт ("admission site:"), поиск по ценам, например "DVD player $150..250", странам, датам, доменам и т.п. Во многих системах обеспечивается поиск не только по данным в формате HTML, но и в форматах PDF, RTF, DOC (MsWord), PS.

В последнее время получили распространение адаптивные интерфейсы уточнения запросов, чаще всего реализованные путем применения методов кластерного анализа к результатам первичного поиска. Появилось такое понятие, как метод "папок поиска" (Custom Search Folders), объединяющее множество подходов, общее в которых - попытка сгруппировать результаты поиска и

представить группы наиболее связанных документов (кластеры) в удобном для пользователей виде.

Например, в поисковых серверах Vivisimo (http://www.vivisimo.com), Mooter (http://www.mooter.com) или Nigma (http://www.nigma.ru) применяется визуальный подход к представлению результатов поиска путем группирования релевантных документов по категориям. В другом поисковом сервере iBoogie (http://www.iboogie.com/) результаты поиска отображаются в виде, близком к экрану проводника Windows. Слова и словосочетания в так называемых «информационных портретах», применяемых, например, в корпоративных информационно-аналитических системах Галактика Zoom [3] и InfoStream [31], также позволяют адаптивно уточнять первичные запросы.

2.6. Характеристики информационного поиска

Существует много характеристик поиска, из которых две признаны основными - это полнота (*recall*) и точность (*precision*). Много внимания в настоящее время отводится также такой смысловой характеристике, как пертинентность. Эта характеристика информационно-поисковых систем означает соответствие полученных в результате поиска документов информационным потребностям пользователя, а не формальному соответствию документа запросу. Для вычисления показателей качества поиска принято рассматривать таблицу, которую заполняют по результатам поиска в учебной коллекции документов. Этот подход был предложен в рамках созданной Американским Институтом Стандартов (NIST) конференции по оценке систем текстового поиска Text REtrieval Conference (TREC, http://trec.nist.gov/) [125] и поддерживается Российским семинаром по Оценке Методов Информационного Поиска (РОМИП, http://romip.ru/). Таблица результатов поиска имеет следующий вид:

Документы	Выданные	Не выданные
Релевантные	*a*	*c*
Не релевантные	*b*	*d*

С помощью этой таблицы показатели информационного поиска рассчитываются следующим образом:

Коэффициент полноты (*recall*):

$$r = \frac{a}{a+c}.$$

Коэффициент точности (*precision*):

$$p = \frac{a}{a+b}.$$

Коэффициент аккуратности (*accuracy*):

$$acc = \frac{a+d}{a+b+c+d}.$$

Ошибка (error):

$$err = \frac{b+c}{a+b+c+d}.$$

F-мера (F-measure):

$$F = \frac{2}{\frac{1}{p}+\frac{1}{r}}$$

Средняя точность (*average precision*):

$$ArgPrec = \frac{1}{k}\sum_{i=1}^{k} prec\_rel(i),$$

где k - количество документов, релевантных некоторому запросу, i - номер релевантного запросу документа, $prec\_rel(i)$ - точность $i-$го релевантного документа (документы ранжируются по релевантности). Если $i-$й релевантный документ не найден, то $prec\_rel(i)=0$.

Как одна из признанных метрических характеристик информационного поиска рассматривается 11-точечный график полноты/точности TREC (РОМИП), который отражает изменения точности в зависимости от полноты и дает более полную информацию, чем метрическая характеристика в виде одной цифры [142]. По оси абсцисс на графике откладываются значения полноты, по оси ординат - значения точности. Если для запроса известно n релевантных

документов, то полнота может принимать дискретные значения 0, 1/n, 2/n, ... , 1. Для того чтобы получить общий график полноты/точности для множества запросов:

1. Рассматриваются фиксированные значения полноты 0.0, 0.1, 0.2, ... , 1.0 (всего 11 значений).

2. Используется специальная процедура интерполяции точности для данных фиксированных значений полноты.

3. Для множества запросов производится усреднение точности для заданных уровней полноты.

Рассмотрим пример, приведенный в документе «Официальные метрики РОМИП» [54] (рис. 9). Пусть коллекция документов содержит 20 документов, 4 из которых релевантны запросу. Пусть система выдает в качестве результатов запроса все эти документы, ранжированные так, что релевантными являются первый, второй, четвертый и пятнадцатый. Для различных значений точности в этом случае полнота принимает значения 0.25, 0.5, 0.75 и 1.0. В соответствии с правилом интерполяции, для значений полноты от 0 до 0.5 точность равна 1.0 (так как первые два документа задают уровень точности 1.0), для значений полноты 0.6 и 0.7 точность равна 0.75, для значений полноты 0.8, 0.9 и 1.0 точность равна 0.27 (4/15).

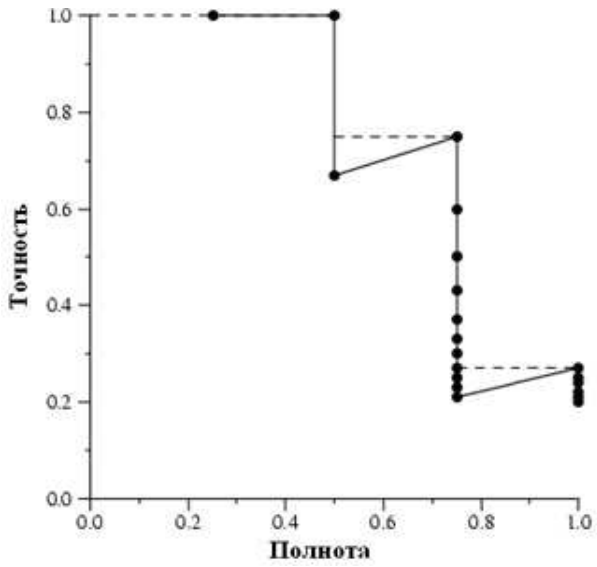


Рис. 9. Зависимость точности от полноты для рассмотренного примера. Пунктирной линией обозначены интерполированные значения

При оценке различных информационно-поисковых систем с помощью 11-точечного графика лучшей считается та система, в которой высокая точность достигается при малой полноте, что свидетельствует о хорошем ранжировании результатов поиска. Кроме того, лучшей признается та система, для которой площадь под соответствующей интерполяционной кривой является наибольшей.

Полнота поиска (recall) тесно связана с оперативностью охвата информации системой. Например, созданная единожды база данных Интернет-ресурсов является "слепком" состояния Интернет в конкретный момент. Если эта база не будет обновляться, имеющиеся в ней ссылки на документы станут «мертвыми». Дополнительно к рассмотренным поисковым характеристикам поисковых систем большое значение имеют такие технологические характеристики, как**:**

- скорость обработки запросов;
- полнота охвата ресурсов;
- доступность, т.е. вероятность получения ответа от системы*;
- нахождение документов, подобных найденным;
- возможность уточнения запросов;
- возможность подключения переводчиков и т.п.

Безусловно, полнота охвата ресурсов Интернет - это один из двух главных аспектов характеристики полноты сетевой информационно-поисковой системы. Второй аспект связан с полнотой информации, которая выдается пользователю по его запросу.

Если под релевантностью понимается формальное соответствие запроса информации, выдаваемой системой, то на практике используется другое, неформальное понятие - пертинентность. Для пользователя пертинентность, соотношение объема полезной для него информации к общему объему полученной информации, имеет решающее значение. При этом следует учитывать, что формальный запрос к системе является предметом творческого

---

* В настоящее время реально существующие в веб информационно-поисковые системы гарантируют доступность уровня «четыре девятки», т.е. 0.9999. Лучшие по этому параметру системы (RBC, Google) обеспечивают «пять девяток».

осмысления информационной потребности и не всегда точно отражает последнюю. Неумение большинством пользователей правильно формулировать запросы и получать приемлемые объемы откликов породило в конце XX столетия мысль относительно веб, как об огромной информационной свалке. Достижение высокой пертинентности - основное поле конкурентной борьбы современных поисковых систем. Именно для максимального удовлетворения информационных потребностей пользователей сетевые информационно-поисковые системы сегодня максимально интеллектуализируются - получили широкое применение технологии и методы семантических и нейронных сетей, Text Mining.

3. КОНЦЕПЦИЯ TEXT MINING

«Остаток дня провел я благонравно,
Приготовлял глаголы, не тужа,
Долбил предлоги и зубрил исправно,
Какого каждый просит падежа.»
Алексей Толстой

Поиск в сетевой среде может стать более эффективным за счет технологий глубинного анализа текстов (Text Mining), нахождения в текстах аномалий и трендов. Разработанные на основе статистического и лингвистического анализа, а также методов искусственного интеллекта, технологии Text Mining предназначены для проведения смыслового анализа. Задача Text Mining - выбирать из текстов наиболее ключевую и значимую информацию для пользователей [75, 32]. Важная компонента технологий Text Mining связана с извлечением из текста характерных элементов или признаков, которые могут использоваться в качестве ключевых слов, метаданных, аннотаций. Еще одна задача Text Mining – отнесение документов к некоторым категориям из заданной схемы их систематизации. Кроме того, Text Mining - это новый вид поиска, который в отличие традиционных подходов не только находит списки документов, формально релевантных запросам, но и помогает в понимании смысла текстов. Таким образом, пользователю будет незачем самому "просеивать" огромное количество неструктурированной информации. Text Mining - это алгоритмическое выявление прежде не известных связей в уже имеющихся данных. Применяя Text Mining, пользователи могут получать новую ценную информацию - знания.

Следует заметить, что технологии глубинного анализа текста исторически предшествовала технология добычи данных (Data Mining), методология и подходы которой широко используются и в методах Text Mining. Для глубинного анализа текстов вполне справедливо определение, данное для Data Mining Г. Пятецким-Шапиро из GTE Labs: "Процесс обнаружения в сырых данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации

знаний, необходимых для принятия решений в различных сферах человеческой деятельности" [122].

Оформившись в середине 90-х годов XX века как направление анализа неструктурированных текстов, технологии Text Mining сразу же взяла на вооружение методы Data Mining, такие как классификация или кластеризация. В Text Mining появились и дополнительные возможности, такие как автоматическое реферирование текстов и выявление феноменов - понятий и фактов. Возможности современных систем Text Mining могут применяться при управлении знаниями для выявления шаблонов в текстах, для автоматического "проталкивания" или распределения информации по интересующим пользователей профилям, создания обзоров.

3.1. Контент-анализ

Один из источников концепции Text Mining - контент-анализ. Понятие контент-анализа, корни которого уходят в психологию и социологию, не имеет однозначного определения:

- Контент-анализ - это методика объективного качественного и систематического изучения содержания средств коммуникации - Д. Джери (J.J. Jerry), Дж. Джери (J. Jerry).
- Контент-анализ - это систематическая числовая обработка, оценка и интерпретация формы и содержания информационного источника - Д. Мангейм (D. Mangeim), Р. Рич (R. Rich).
- Контент-анализ - это качественно-количественный метод изучения документов, который характеризуется объективностью выводов и строгостью процедуры и заключается в квантифицированной обработке текста с дальнейшей интерпретацией результатов (В. Иванов).
- Контент-анализ состоит в нахождении в тексте определенных содержательных понятий (единиц анализа), выявлении частоты их появления и соотношения с содержанием всего документа (Б. Краснов).

Большинство из приведенных определений конструктивны, но из-за различных начальных посылок они порождают различные, а порой и противоречащие друг другу алгоритмы.

Принято разделять методологии контент-анализа на две области: качественную и количественную. Основа количественного контент-анализа - частота появления в документах определенных характеристик содержания (понятий, феноменов). Качественный контент-анализ основан на самом факте присутствия или отсутствия в тексте одной или нескольких характеристик содержания.

3.2. Элементы Text Mining

В соответствии с уже сложившейся методологией, к основным элементам Text Mining относятся: классификация (classification, categorization), кластеризация (clustering), извлечение фактов, понятий (feature extraction), реферирование (summarization), ответ на запросы (question answering), тематическое индексирование (thematic indexing) и поиск по ключевым словам (keyword searching).

При классификации текстов, методы которой детально рассматриваются в четвертой главе, используются статистические корреляции для размещения документов в определенные категории. Задача классификации - это классическая задача распознавания, где по некоторой контрольной выборке система относит новый объект к той или иной категории. Особенность классификации в рамках концепции Text Mining заключается в том, что количество объектов и их атрибутов может быть очень большим, поэтому должны быть предусмотрены механизмы оптимизации этого процесса.

В отличие от классификации, при кластеризации заранее не фиксируются определенные категории. Результатом кластеризации является автоматическое группирование информации, в результате которой создаются классификационные схемы, обеспечивающие эффективный охват больших объемов данных. Кластеризация в Text Mining рассматривается как процесс выделения компактных

подгрупп объектов с близкими свойствами. При кластеризации система должна самостоятельно найти признаки и разделить объекты по группам. Кластеризация, как правило, предшествует классификации, поскольку позволяет определять группы объектов.

Text Mining предусматривает также построение семантических сетей, анализ связей, которые определяются появлением дескрипторов (например, ключевых слов) в текстах.

Кроме того, существует еще несколько задач технологии Text Mining, например, прогнозирование, которое заключается в том, чтобы предсказать по значениям одних признаков текста значения остальных. Еще одна задача - нахождение исключений, то есть поиск документов, которые своими характеристиками выделяются из общей массы [3]. Для этого сначала выясняются средние параметры документов, а затем исследуются те документы, параметры которых наиболее сильно отличаются от средних значений. Обычно поиск исключений зачастую проводится после классификации или кластеризации для того чтобы выяснить, насколько последние были точны.

Несколько отдельно от задачи кластеризации стоит задача поиска связанных признаков (ключевых слов, понятий) отдельных документов. От прогноза эта задача отличается тем, что заранее не известно, по каким именно признакам реализуется взаимосвязь - цель именно в том и состоит, чтобы найти связи признаков. Эта задача сходна с кластеризацией, но не по множеству документов, а по множеству признаков.

3.2.1. Извлечение понятий

Извлечение понятий (Feature Extraction) из текста представляет собой технологию, обеспечивающую получение информации в структурированном виде. В качестве структур могут запрашиваться как относительно простые понятия (ключевые слова, персоны, организации, географические названия), так и более сложные, например, имя персоны, ее должность в конкретной организации и т.п.

Данная технология включает три основных метода:

а) Entity Extraction - извлечение слов или словосочетаний, важных для описания содержания текста. Это могут быть списки терминов предметной области, персон, организаций, географических названий, и др.;

б) Feature Association Extraction - прослеживание связей между извлеченными понятиями;

в) Event and Fact Extraction - извлечение сущностей, распознавание фактов и событий.

Технология извлечения понятий основана на применении специальных семантико-лингвистических методов, которые дают возможность получать приемлемую точность и полноту.

Следует отметить, что подходы к извлечению различных типов понятий из текстов существенно разняться как по контексту их представления, так и по структурным признакам. Так, для выявления принадлежности документа к тематической рубрике могут использоваться специальным образом составленные запросы на информационно-поисковых языках, включающих логические и контекстные операторы, скобки и т.д. Выявление географических названий предполагает использование таблиц, в которых кроме шаблонов написания этих названий используются коды и названия стран, регионов и отдельных населенных пунктов.

В качестве одного из примеров рассмотрим алгоритм выявления названий фирм в текстах документов (рис. 10). На вход системы поступает документ, который анализируется в процессе последовательного считывания (блок «Чтение документа»). Текст документа сравнивается с шаблонами, соответствующими названиям известных фирм, и если такие присутствуют, то они помещаются в специальную таблицу «документ-фирма». Также система извлечения понятий предполагает выявление неизвестных изначально названий фирм на основании как шаблонов, так и результатов структурных исследований текста. При этом, в частности, используется таблица префиксов названий фирм, содержащая такие элементы, как «ООО», «ЗАО», «АО», «Компания» и др.

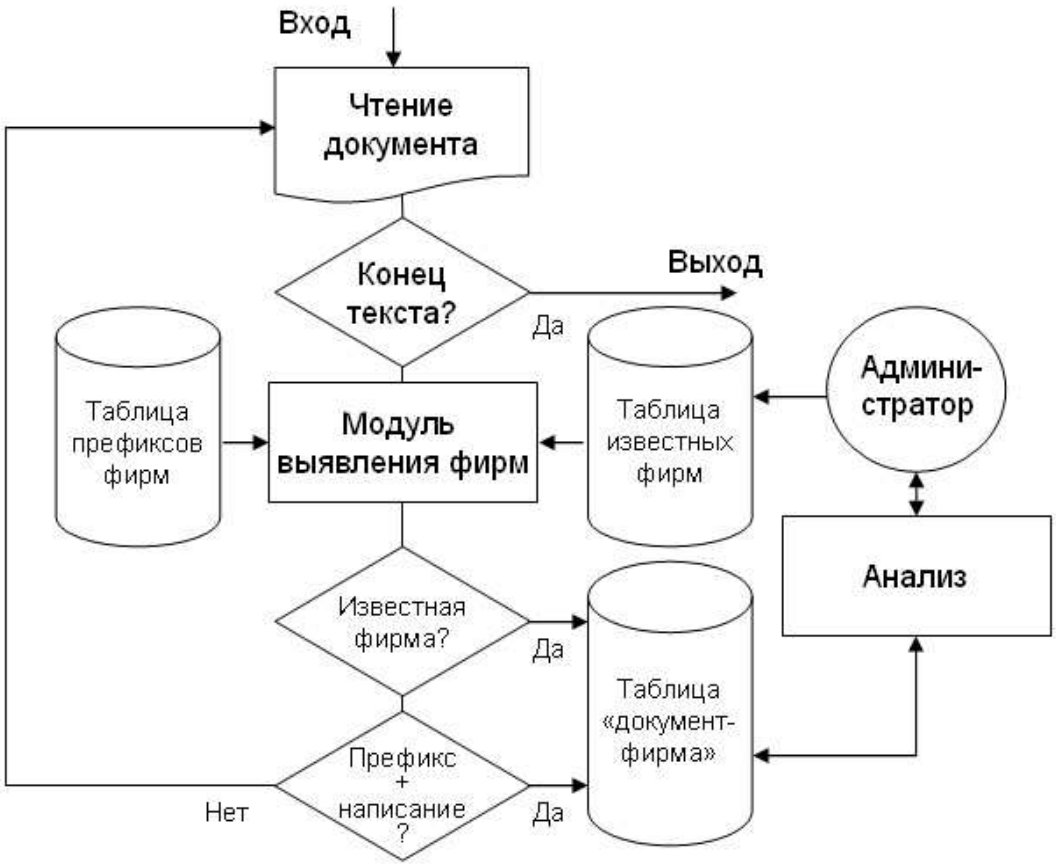


Рис. 10. Алгоритм выявления названий фирм из текстов документов

Выявленные понятия могут служить основой для построения многопрофильных информационных портретов или интерактивных ситуационных карт (сетей, узлами которой являются понятия, а ребрами – информационные связи между ними), соответствующих запросам пользователей. Непосредственно по данным, представленным на ситуационной карте, отражающей наиболее актуальные понятия (термины, тематические рубрики, географические названия, фамилии персон, названия компаний) возможно выявление взаимосвязей, т.е. сами ситуационные карты могут служить исходными данными для построения сетей взаимосвязей понятий.

3.2.2. Определение взаимосвязей понятий

Таблицы взаимосвязей понятий строятся как статистические отчеты, отражающие близость (совместную встречаемость в документах или близость по

сопутствующему контексту) отдельных понятий. Это симметричные матрицы, элементы которых – коэффициенты взаимосвязей понятий, соответствующих ее строкам и столбцам. Эти коэффициенты пропорциональны количеству документов входного информационного потока, которые соответствуют одновременно обоим понятиям, или количеству значимых лексических единиц, употребляемых совместно с данными понятиями. Таким образом, взаимосвязь понятий может быть оценена с помощью двух алгоритмов:

- совместного вхождения – путем расчета совместного вхождения понятий в одни и те же документы;
- контекстной близости - путем расчета корреляций наборов ключевых слов, которые входят в документы, в которых упоминались данные понятия.

Рассмотрим формальное определение таблицы взаимосвязей понятий TVP', построенной с помощью первого алгоритма. Обозначим p_j $(j = 1, ..., M)$ - понятие, D - массив документов, $d^{(i)} \in D$ $(i = 1, ..., N)$ – документ, P_j - подмножество D, соответствующее понятию p_j, $e_j^{(i)}$ – признак соответствия понятия документу:

$$e_j^{(i)} = \begin{cases} 1, & d^{(i)} \in P_j, \\ 0, & d^{(i)} \notin P_j. \end{cases}.$$

Можно определить уровень связи понятий p_j и p_k:

$$v_{j,k} = \sum_{i=1}^{N} e_j^{(i)} e_k^{(i)}.$$

Значения $v_{j,k}$ в совокупности образуют матрицу таблицы взаимосвязей понятий TVP'.

Для случая второго алгоритма, учитывающего контекстную близость, таблицу взаимосвязей понятий TVP'' формально определим следующим образом. Обозначим $W_i = \{w_1^{(i)}, ..., w_n^{(i)}\}$ - множество ключевых слов, входящих в документ $d^{(i)}$.

Введем понятие профайла понятия p_j $(j=1,\ ...,\ M)$ как множества ключевых слов из документов, соответствующих этому понятию:

$$IP(p_j) = \bigcup_{d^{(i)} \in P_j} W_i.$$

Введем также понятия словаря системы $S = \{s_1,\ ...,\ s_K\}$ как множества ключевых слов, входящих в D, и вектора $\vec{t}^{(j)} = (t_1^{(j)},\ ...,\ t_K^{(j)})$ с элементами $t_i^{(j)}$, соответствующими профайлу темы:

$$t_i^{(j)} = \begin{cases} 1, & s_i \in IP(p_j),\ i=1,\ ...,\ K, \\ 0, & s_i \notin IP(p_j),\ i=1,\ ...,\ K. \end{cases}$$

В этом случае уровень связи понятий p_j и p_k можно определить следующим образом:

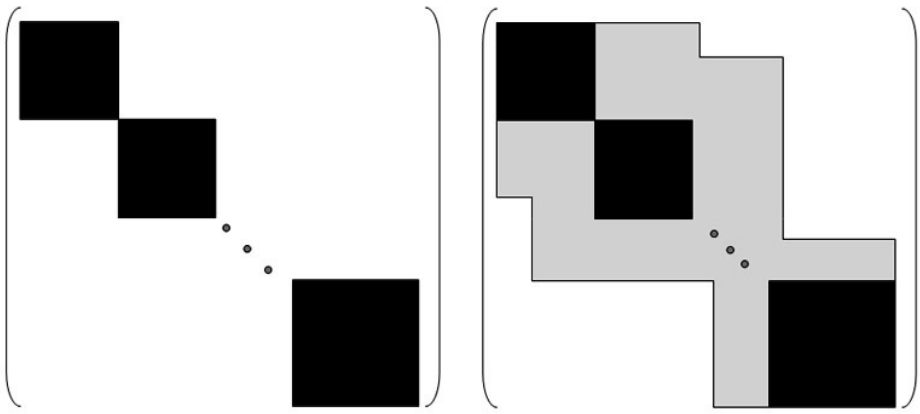
$$\tilde{v}_{j,k} = \vec{t}_j \vec{t}_k = \sum_{i=1}^{K} t_i^{(j)} t_i^{(k)}.$$

Таким образом, таблица взаимосвязей понятий второго типа $TVP"$ будет состоять из значений $\tilde{v}_{j,k}$.

Следует отметить, что таблица взаимосвязей первого типа всегда отражает взаимосвязи понятий точнее, чем таблица взаимосвязей второго типа, однако таблица второго типа учитывает взаимосвязи более полно (рис. 11).

Действительно, из того факта, что $v_{j,k} > 0$ следует, что $\tilde{v}_{j,k} > 0$, так как из первое условие определяет то, что существует хотя бы один такой документ (с индексом i), что $d^{(i)} \in P_j,\ d^{(i)} \in P_k$. Отсюда следует, что пересечения профайлов соответствующих понятий не пусто: $IP(p_j) \bigcap IP(p_k) \neq \varnothing$, а соответственно, $\vec{t}_j \vec{t}_k = \tilde{v}_{j,k} > 0.$

Обратное утверждение в общем случае неверно. Проведем мысленный эксперимент, подтверждающий это замечание. Рассмотрим два понятия «пингвин» и «белый медведь». Эти понятия могут иметь ненулевое контекстное пересечение за счет таких ключевых слов, как «лед», «мороз», «рыба», однако понятие «пингвин» входит в документы, описывающие фауну Антарктики, а «белый медведь» - Арктики.



а) *б)*

Рис. 11. Два варианта таблицы взаимосвязей понятий: а) - TVP', б) - TVP" (нулевые элементы соответствуют белым областям, совпадающие – черным)

Для переупорядочения понятий из таблицы взаимосвязей с целью выявления множеств наиболее взаимозависимых из них (путем выявления диагональных блоков, см. рис. 11) применяются методы кластерного анализа, в частности алгоритм *k*-means (см. п. 5.2), который является одним из самых эффективных для группировки данных из динамически изменяемых массивов.

3.2.3. Автоматическое реферирование

Автоматическое реферирование (Automatic Text Summarization) - это составление коротких изложений материалов, аннотаций или дайджестов, т.е. извлечение наиболее важных сведений из одного или нескольких документов и генерация на их основе лаконичных отчетов [55].

Существует много путей решения этой задачи, которые довольно четко подразделяются на два направления - квазиреферирование и краткое изложение содержания первичных документов. Квазиреферирование основано на экстрагировании фрагментов документов - выделении наиболее информативных фраз и формировании из них квазирефератов.

В рамках квазиреферирования выделяют три основных направления, которые в современных системах применяются совместно:

- статистические методы, основанные на оценке информативности разных элементов текста по частоте появления, которая служит основным критерием информативности слов, предложений или фраз;
- позиционные методы, которые опираются на предположение о том, что информативность элемента текста зависит от его позиции в документе;
- индикаторные методы, основанные на оценке элементов текста, исходя из наличия в них специальных слов и словосочетаний - маркеров важности, которые характеризуют их содержательную значимость.

Определение веса фрагментов (предложений или абзацев) исходного текста выполняется в соответствии с алгоритмами, которые стали уже традиционными. Общий вес текстового блока при этом определяется по формуле:

$$Weight = Location + KeyPhrase + StatTerm.$$

Слагаемое *Location* определяется расположением блока в тексте и зависит от того, где появляется данный фрагмент - в начале, в середине или в конце, а также используется ли он в наиболее важных с содержательной точки зрения разделах текста, например, в выводах. Ключевые фразы (*KeyPhrase*) представляют собой конструкции-маркеры, которые резюмируют содержание, типа "в заключение", "в данной статье", "в результате анализа" и т.п. Весовое значение слагаемого *KeyPhrase* может зависеть также от оценочного термина, например, "отличный". Статистический вес текстового блока (*StatTerm*) вычисляется как нормированная по длине блока сумма весов входящих в него слов и словосочетаний.

После выявления определенного (задаваемого, как правило, коэффициентом необходимого сжатия) количества текстовых блоков с наивысшими весовыми коэффициентами, они объединяются для построения квазиреферата.

Преимущество методов квазиреферирования заключается в простоте их реализации. Однако выделение текстовых блоков, не учитывающее взаимоотношений между ними, часто приводит к формированию бессвязных рефератов. Некоторые предложения могут оказаться пропущены, либо в них могут встречаться слова или фразы, которые невозможно понять без предшествующего пропущенного текста. Попытки решить эту проблему, в основном сводятся к исключению таких предложений из рефератов. Реже

делаются попытки разрешения ссылок с помощью методов лингвистического анализа.

Краткое изложение содержания первичных документов основывается на выделении из текстов наиболее важной информации и порождении новых текстов, содержательно обобщающие первичные документы. В отличие от частотно-лингвистических методов, обеспечивающих квазиреферирование, подход, основанный на базах знаний, опирается на автоматизированный качественный контент-анализ, состоящий, как правило, из трех основных стадий. Первая - сведение исходной текстовой информации к заданному числу фрагментов - единиц значения, которыми являются категории, последовательности и темы. На второй стадии производится поиск регулярных связей между единицами значения, после чего начинается третья стадия - формирование выводов и обобщений. На этой стадии создается структурная аннотация, представляющая содержание текста в виде совокупности концептуально связанных смысловых единиц.

Семантические методы формирования рефератов-изложений предполагают два основных подхода: метод синтаксического разбора предложений и методы, опирающиеся на понимание естественного языка. В первом случае используются деревья разбора текста. Процедуры автоматического реферирования манипулируют непосредственно деревьями, выполняя перегруппировку и сокращение ветвей на основании сответствующих критериев. Такое упрощение обеспечивает построение реферата - структурную "выжимку" исходного текста.

Второй подход основывается на системах искусственного интеллекта, в которых также на этапе анализа выполняется синтаксический разбор текста, но синтаксические деревья не порождаются. В этом случае формируются семантические структуры, которые накапливаются в виде концептуальных подграфов в базе знаний. В частности, известны модели, позволяющие производить реферирование текстов на основе психологических ассоциаций сходства и контраста. В базах знаний избыточная и не имеющая прямого отношения к тексту информация устраняется путем отсечения некоторых подграфов. Затем информация подвергается агрегированию методом слияния

оставшихся графов или их обобщения. Для выполнения этих преобразований выполняются манипуляции логическими предположениями, выделяются определяющие шаблоны в текстовой базе знаний. В результате преобразования формируется концептуальная структура текста - аннотация, т.е. концептуальные "выжимки" из текста.

Многоуровневое структурирование текста с использованием семантических методов позволяет подходить к решению задачи реферирования путем:

- удаления малозначащих смысловых единиц. Преимуществом метода является гарантированное сохранение значащей информации, недостатком - низкая степень сжатия, т.е. сокращения объема реферата по сравнению с первичными документами;
- сокращения смысловых единиц - замена их основной лексической единицей, выражающей основной смысл;
- гибридного способа, заключающегося в уточнении реферата с помощью статистических методов, с использованием семантических классов, особенностей контекста и синонимических связей.

Существуют общедоступные программы квазиреферирования, например, в состав сервисных возможностей системы Microsoft Word входит режим «Автореферат».

3.2.4. Поисковые образы документов

На основе методов автоматического реферирования (в частности, квазиреферирования) возможно формирование поисковых образов документов (ПОД) [1]. По автоматически построенным аннотациям больших текстов (поисковым образам документов) проводится поиск, который характеризуется высокой точностью, достигаемой за счет полноты. То есть вместо поиска в полных текстах в некоторых случаях может оказаться целесообразным поиск в специально созданных аннотациях, которые рассматриваются как поисковые образы документов. Хотя ПОД зачастую для больших документов оказывается образованием, лишь отдаленно напоминающим исходный текст и не всегда

воспринимаемый человеком, но за счет включения наиболее весомых фраз, он может приводить к вполне адекватным результатам при проведении полнотекстового поиска.

3.2.5. Выявление дублирования информации

В сети Интернет важные сообщения многократно дублируются на экспоненциально растущем количестве сайтов, в то время как количество заслуживающих внимания источников растет не такими высокими темпами, скорее всего, линейно.

Выявление дублирующихся сообщений (их принято называть «дубликатами»), а также перепечаток документов с небольшими изменениями («почти дублей») является одной из актуальнейших и сложнейших задач. Понятие содержательных дублей документов достаточно расплывчато, до сих пор остается открытой задача анализа таких явлений, как пересказ одних и тех же событий, описание различных аспектов разными людьми.

В свое время определенные (не оправдавшиеся) надежды возлагались на развитие так называемых семантических методов, которые бы позволили оперировать непосредственно со смыслом сообщений, и таким образом избежать проблем его формализации.

С прагматической точки зрения в применении таких методов следует выделить два главных недостатка. Это существенная зависимость практической реализации метода от языка обрабатываемых дкументов (что фактически делает невозможной работу с многоязычными текстовыми массивами) и его неустойчивость: для некоторых информационных массивов результаты очень хорошие, но для других – очень плохие.

Пессимистический взгляд на применение «семантических» методов в области информационных технологий, в общем-то, вполне понятен. Действительно, семантика занимается отношением лингвистических конструкций к предметам и явлениям реального мира, тогда как компьютерные системы могут манипулировать исключительно формальными элементами. Иными словами, в рамках любой информационной технологии можно устанавливать отношения

только одних лингвистических конструкций с другими лингвистическими конструкциями. Вопрос о том, в какой мере все это может отражать семантические связи, остается открытым.

С другой стороны, игнорировать семантические аспекты информационных технологий, несомненно, было бы ошибкой. Интуиция и опыт подсказывают, что понятие семантической близости документов должно иметь определенный смысл и на уровне машинной обработки текстов.

Серьезное упрощение может быть получено за счет применения содержательных методов, например, путями ранжирования первоисточников, определения и выделения тематических информационных каналов, экспертного формирования словарей значимых слов и т.п.

Преодоление использования явно дублирующейся информации не представляет проблем, однако дублирующиеся по смыслу сообщения выявляются не так легко, здесь на помощь приходят алгоритмы, базирующиеся на вероятностных оценках. На практике явные дубликаты выявляются даже с помощью механизмов контрольных сумм, но этот подход не решает проблем пользователей, для которых чаще всего не имеет значения, с чем они имеют дело: с прямой перепечаткой или с небольшой перефразировкой. Вместе с тем многие недобросовестные издания перепечатывают содержание сообщений, попросту изменяя заглавия (работа «хедлайнеров»). И такой вид дублирования элементарно обходится с помощью контрольных сумм (но уже без учета заголовков). Дальнейший анализ показал, что при перепечатке материалов чаще всего остаются без изменений несколько первых предложений текста или первый абзац. И этот критерий был учтен и успешно внедрен. Вместе с тем качество выявления содержательного дублирования оставалось недостаточно высоким.

Известны подходы, основанные на учете повторений встречаемости цепочек слов, например, метод «шинглов» (чешуек), описанный в работах [82], [103] и [110]. Этот остроумный и эффективный метод поиска «почти дублей» оказался не очень чувствительным для небольших текстов с возможными перефразировками.

Наиболее прямой путь к установлению связи между произвольным документом и семантическим пространством предполагает наличие некоторого

соответствия между устойчивыми сочетаниями слов и единицами смысла. При всей своей внешней банальности, это утверждение отнюдь не тривиально, поскольку речь в нем идет именно о морфизме, но отнюдь не об эквивалентности.

Устойчивое сочетание слов само по себе вовсе не является единицей смысла. Более того, далеко не всегда единица смысла вообще может быть артикулирована с помощью набора слов. Но между наборами слов и единицами смысла всегда или почти всегда могут быть установлены (вообще говоря, неоднозначно) устойчивые отношения.

Метод выявления дубликатов, используемый, например, в системе InfoStream, в частности, заключается в признании документов дубликатами, если у них совпадает более 6 из 12 отобранных по статистическим критериям ключевых слов (термов, образующих так называемые «словарные сигнатуры» документа). Следует отметить, что применение более «мягкого» критерия к множеству отобранных термов позволяет реализовать режим «поиска подобных документов».

Введем обозначения: "$\prec$" – оператора подобия и "$\equiv$" - оператора дублирования. Очевидно, что для алгоритма выявления подобных документов и дубликатов, о котором идет речь, справедливо правило рефлексивности:

$$A \prec A, \quad A \equiv A.$$

где A – произвольный документ.

Оператор подобия не обладает свойством симметричности. Из подобия документа A документу B не следует обратного, т.е.:

$$A \prec B \quad \not\Rightarrow \quad B \prec A.$$

Также не выполняется условие транзитивности:

$$A \prec B, \quad B \prec C \quad \not\Rightarrow \quad A \prec C.$$

Действительно, например, отельный документ может быть подобен тексту из подборки, которая его включает, но сама подборка может не быть подобной этому документу. Или документ может быть подобен двум документам, из которых он скомпилирован, но сами оригиналы могут существенно отличаться.

Для отношения дублирования, наоборот, симметричность и транзитивность выполняются:

$A \equiv B \Rightarrow B \equiv A,$

$A \equiv B,\ B \equiv C \Rightarrow A \equiv C.$

Заметим, что отношение, обладающее свойствами рефлексивности, симметричности и транзитивности является отношением эквивалентности, в нашем случае, отношением содержательного совпадения или дублирования.

Как было замечено, свойство дублирования документов является более жестким критерием подобия, например, совпадение 3, 4 или 5 термов свидетельствуют о некоторой содержательной близости.

На практике каждой паре документов $d^{(i)}$ и $d^{(j)}$ из контрольного документального массива ставился в соответствие вектор с элементами:

$$a_{i,j} = \begin{cases} 1, & d^{(i)} \equiv d^{(j)}, \\ 0, & d^{(i)} \not\equiv d^{(j)}. \end{cases}$$

Условие симметричности в этих обозначениях записывается следующим образом:

$\forall\ i,\ j\ :\ a_{i,j} = a_{j,i},$

а транзитивность определяется выполнением условия:

$\forall\ i,\ j,\ k\ :\ a_{i,j} = 1,\ a_{j,k} = 1 \Rightarrow a_{i,k} = 1.$

Были исследованы критерии подобия (изменяя количество сравниваемых в словесных сигнатурах документов термов), чтобы достичь на контрольном документальном корпусе максимального уменьшения коэффициента асимметричности:

$$\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} |a_{i,j} - a_{j,i}|}{\sum_{i=1}^{N}\sum_{j=1}^{N} a_{i,j}},$$

и увеличения коэффициента транзитивности:

$$\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} a_{i,j}a_{j,k}a_{i,k}}{\sum_{i=1}^{N}\sum_{j=1}^{N} a_{i,j}},$$

где N – количество документов в контрольном массиве.

Очевидно, что так рассчитываемый коэффициент асимметричности ассоциируется с огрублениями при определении дубликатов, а уровень транзитивности – с полнотой.

Вместе с тем следует заметить, что проверка коэффициентов асимметричности и транзитивности может использоваться лишь для формальной проверки приближения отношения к свойствам эквивалентности. Само определение того, что эта эквивалентность – содержательное дублирование должно быть предоставлено аналитиками-экспертами. Приведенный выше алгоритм, кроме своего эмпирического подтверждения, хорош тем, что позволяет варьировать некоторым параметром (количеством сравниваемых термов), значение которого можно подобрать с учетом оптимизации двух названных коэффициентов.

3.2.6. Выявление новых событий

Как правило, задача выявления новых событий из потока сообщений предполагает, что на вход соответствующего программно-технологического комплекса последовательно поступают новые документы. Они могут поступать как непосредственно от средств сканирования, так и будут отобраны по тематическому запросу. При этом зачастую остается открытым прогнозный вопрос, какое событие в данный момент освещено пока мало, но в дальнейшем получит большой резонанс. Этот вопрос связан с общей задачей нахождения исключений или аномалий, т.е. объектов, которые своими характеристиками значительно выделяются из общей массы (хотя в дальнейшем могут породить множество себе подобных). Для решения этой проблемы было предложено несколько путей.

Подход Г. Солтона в определении повых событий заключается в использовании векторно-пространственного представления документов и традиционных методов кластеризации. При этом малый вес приписывается высокочастотным словам из массива документов, что вполне укладывается в модель TF IDF. Документы при этом подходе обрабатываются последовательно в соответствии с таким алгоритмом:

1. Первому рассматриваемому документу ставится в соответствие первый кластер. Каждый кластер представляется вектором термов (ключевых слов), входящих в документы этого кластера. Нормированный каким-то образом вектор термов принято называть центроидом. Иногда центроидом называют документ, самый близкий по некоторому критерию к вектору термов данного кластера, что не меняет сути данного алгоритма.
2. Каждый следующий документ сравнивается с центроидами существующих кластеров (для этого вводится некоторая мера близости).
3. Если документ достаточно близок к некоторому кластеру, то он приписывается этому кластеру, после чего происходит пересчет соответствующего центроида.
4. Если документ не близок к существующим кластерам, то происходит формирование нового кластера, которому приписывается данный документ.
5. Временной диапазон рассматриваемых документов принято называть «окном наблюдения». Кластеры, все документы которых выходят за пределы окна наблюдения, выносятся за рамки рассмотрения.

В результате работы алгоритма каждому новому возникающему кластеру соответствует новое событие, отражаемое в документах данного кластера.

В сответствии с подходом, предлагаемым Р. Папка [120], новые события выявляются из документов, не удовлетворяющих запросам пользователей, построенным с учетом уже известных событий. Алгоритм выявления новых событий заключается в следующем:

1. Формируются запросы по известным темам (при этом используются технологии Text Mining – выявления и выбора понятий из текстов сообщений).

2. Новый поступающий документ сравнивается с существующими запросами.
3. Если документ не соответствует запросам, то он ассоциируется с новым событием.
4. В систему включается новый запрос, соответствующий данному документу.

В реально работающих системах интеграции новостей, как правило, применяются многопараметрические подходы, учитывающие, не только информацию из текста новостей, но и время их публикации, уровень источника, соответствие тематикам пользователей [94]. Один из таких подходов к выявлению новых событий [31] базируется на таких предположениях, относящихся к публикации соответствующих информационных сообщений:

а) минимальное время, прошедшее с момента публикации;

б) минимизация веса термов, входящих в документ, по частотному словарю, сформированному на основании анализа большого массива опубликованных документов (это условие, аналогичное максимизации параметра *IDF* в векторно-пространственной модели);

в) максимизация суммарного веса термов, входящих в документ, по плюс-словарю (содержащему важные для содержания новостей слова типа «теракт», «конфликт», «сенсация» и т.п.);

г) учет ранга «авторитетности» источника (как правило, определяемый экспертами).

Введем следующие обозначения:

n – величина окна наблюдения потока новостей;

D_1 – текущий документ;

D_n – последний документ из окна наблюдения;

D_i – i – й документ;

$PlusDic$ – плюс словарь;

$sim(D_i, D_j)$ – мера близости документа i документу j;

$sim(D_i, PlusDic)$ – мера близости документа i «плюс словарю»;

$Rank_i$ – ранг источника, соответствующего i-му документу.

Мера близости $sim(D_i, D_j)$ может быть определена традиционно для векторно-пространственой модели. При этом может быть дано еще одно определение меры близости документов, использующее аппарат условных вероятностей, а именно, вероятность того, что случайно выбранное слово w входит в документ D_i при условии, что оно входит в документ D_j, умноженную на вероятность вхождения данного слова в документ D_j:

$$sim(D_i, D_j) = \mathrm{P}(w \in D_i \mid w \in D_j) P(w \in D_j).$$

Параметр новизны New_i документа D_i, учитывающий условия а) - г), может быть записан следующим образом:

$$New_i = \frac{Rank_i \cdot sim(D_i, PlusDic)}{\log(i+1) \sum_{j=1}^{N} sim(D_i, D_j)}.$$

Задачи выявления, отслеживания и группировки событий на основе анализа новостей активно обсуждаются, они имеют большое практическое значение имеенно сегодня, когда режим онлайн-доступа к системам интеграции новостей существенно облегчен.

3.3. Реализации систем с элементами Text Mining

В настоящее время существует множество систем глубинного анализа текстов, как встроенных в другие, более комплексные системы, так и автономных. В частности, корпорация IBM (www.ibm.com) создала систему Intelligent Miner for Text, представляющую собой набор утилит, реализующих функции Text Mining:

- Language Identification Tool - утилита определения языка, на котором составлен документ.
- Categorisation Tool - утилита классификации - автоматического отнесения текста к некоторой категории.

- Clusterisation Tool - утилита кластеризации - разбиения большого множества документов на группы по близости стиля, формы, различных частотных характеристик ключевых слов.
- Feature Extraction Tool - утилита определения нового - выявление в документе новых термов, таких как собственные имена, названия, сокращения, на основе анализа заданного заранее словаря.
- Annotation Tool - утилита "выявления содержания" текстов и составления рефератов - аннотаций.

Другая известная система PolyAnalyst компании Мегапьютер Интеллидженс (www.megaputer.com) может применяться для автоматизированного анализа числовых и текстовых баз данных с целью выявления прежде неизвестных, нетривиальных, полезных и доступных пониманию закономерностей. В состав PolyAnalyst входит система TextAnalyst, которая позволяет решать такие задачи Text Mining: построение семантической сети для больших текстов, подготовка резюме текста, поиск по тексту, автоматическая классификация и кластеризация текстов.

Система компании SAS (www.sas.com) содержит компонент SAS Text Miner, который позволяет работать с текстовыми документами в различных форматах из баз данных, файловых систем и веб, а также агрегировать текстовую информацию со структурированными данными.

Средства Text Mining сегодня являются неотъемлемой частью продуктов компании Oracle (www.oracle.com) . Основной задачей, на решение которой нацелены средства Oracle Text, является задача поиска документов по их содержанию. Oracle Text обеспечивает проведение тематического анализа текстов на английском языке. В ходе обработки текст каждого документа подвергается процедурам лингвистического и статистического анализа, в результате чего определяются его ключевые темы и строятся тематическое, а также общее резюме - реферат.

4. МЕТОДЫ КЛАССИФИКАЦИИ ИНФОРМАЦИИ

«Каких зверей, каких там птиц я не видал!
Какие бабочки, букашки,
Козявки, мушки, таракашки!»
Иван Крылов

4.1. Задача классификации

Под классификацией текстов (Text Categorization, TC) понимается распределение текстовых документов по заранее определенным категориям (в противоположность кластеризации, где множество категорий заранее неизвестно).

Методы классификации текстов лежат на стыке двух областей - машинного обучения (machine learning, ML) и информационного поиска (information retrieval, IR) [33, 134]. Соответственно автоматическая классификация может осуществляться:

- на основе заранее заданной схемы классификации и уже имеющегося множества классифицированных документов;
- полностью автоматизировано.

При применении подходов машинного обучения, классификационное правило строится на основе тренировочной коллекции текстов.

Задача классификации текстов заключается в определении принадлежности текста, который рассматривается, одному или нескольким классам. Классификация может определяться общей тематикой текстов, наличием определенных дескрипторов или выполнением определенных условий, иногда довольно сложных.

Для каждого класса эксперты отбирают текстовые массивы (наборы типичных документов), которые используются системой классификации в режиме обучения. После того как обучение закончено, система с помощью специальных алгоритмов сможет распределять входные потоки текстовой информации по классам.

Классификацию можно рассматривать как задачу распознавания образов, при таком подходе для каждого объекта выделяются наборы признаков. В случае текстов признаками являются слова и взаимозависимые наборы слов - термы, которые содержатся в текстах. Для формирования набора признаков для каждого документа используются лингвистические и статистические методы. Признаки группируются в специальную таблицу - информационную матрицу. Каждая строка матрицы соответствует одному из классов, каждый элемент строки – одному из признаков; численное значение этого элемента определяется в процессе обучения системы классификации. Когда обучение завершается, принадлежность нового текста к одному из классов устанавливается путем анализа признаков этого текста с учетом соответствующих весовых значений. Существующие алгоритмы позволяют проводить классификацию с довольно высокой точностью, однако результаты достигаются за счет больших размеров информационной матрицы, которая определяется общим числом дескрипторов - термов.

Автоматическая классификация может применяться в таких процедурах информационного поиска :

- фильтрация (избирательный отбор) информации;
- формирование тематических каталогов;
- поиск по классам;
- реализация обратной связи по релевантности путем классификации результатов поиска и выбора пользователем релевантных классов;
- расширение запросов за счет термов, которые характеризуют тематику класса;
- снятие омонимии (т.е. учет тех случаев, когда одно и то же слово может иметь разный смысл);
- автоматическое реферирование.

4.1.1. Формальное описание задачи классификации

Пусть $D=\{d^{(1)}, ..., d^{(N)}\}$ - множество документов, $C=\{c_1, ..., c_M\}$ -

множество категорий, Φ - целевая функция, которая по паре $\langle d^{(i)}, c_j \rangle$ определяет, относится ли документ $d^{(i)}$ к категории c_j (1 или True) или нет (0 или False). Задача классификации состоит в построении функции $\tilde{\Phi}$, максимально близкой к Φ.

Коллекция заранее классифицированных экспертами документов, т.е. таких, для которых уже точно известно значение целевой функции, разбивается на две части:

1. Учебная коллекция. Классификатор Φ' строится на основе характеристик этих документов.

2. Тестовая коллекция. На ней проверяется качество классификации. Эти документы не должны использоваться в процессе построения классификатора.

Рассматриваемая классификация называется четкой бинарной, то есть подразумевается, что существуют только две категории, которые не пересекаются. К такой классификации сводится много задач, например, классификация по множеству категорий $C = \{c_1, ..., c_M\}$ разбивается на M бинарных классификаций по множествам $\{c_i, \overline{c_i}\}$.

Часто используется ранжирование, при котором множество значений целевой функции - это отрезок [0, 1]. Документ при ранжировании может относиться не только к одной, а сразу к нескольким категориям с разной степенью принадлежности, т.е. категории могут пересекаться между собой.

4.1.2. Ранжирование и четкая классификация

Предположим, что для каждой категории c_i построена функция CSV_i (статус классификации), отображающая множество документов D на отрезок [0; 1], которая задает степень принадлежности документа категории. Рассмотрим задачу, заключающуюся в том, чтобы от функции ранжирования перейти к точной классификации. Наиболее простой способ - для каждой категории c_i выбрать предельное значение (порог) τ_i. Если $CSV_i(d) > \tau_i$, то документ d соответствует категории c_i. Возможен и другой подход - для каждого документа d выбирать k

ближайших категорий, т.е. k категорий, на которых $CSV_i(d)$ принимают наибольшие значения.

Выбор порогового значения возможен, например таким способом. Учебная коллекция разбивается на две части. Для каждой категории c_i на одной части учебной коллекции вычисляется, какая часть документов ей принадлежит. Пороговые значения выбирается так, чтобы на другой части учебной коллекции количество документов, отнесенных c_i, было таким же.

4.1.3. Линейная классификация

Пусть каждой категории C_i соответствует вектор $\vec{c}^{(i)} = (c_1^{(i)}, \ ..., \ c_N^{(i)})$, где N - размерность пространства термов. В качестве правила классификатора документа d используется формула:

$$CSV^{(i)}(d) = \vec{d} \cdot \vec{c}^{(i)}.$$

Нормализация проводится обычно таким образом, чтобы итоговая формула для $CSV^{(i)}(d)$ представляла собой нормированное скалярное произведение - косинус угла между вектором категории C_i и вектором из весовых значений термов $\vec{d} = (d_1, \ ..., \ d_N)$, входящих в документ d:

$$CSV^{(i)}(d) = \frac{\vec{d} \cdot \vec{c}^{(i)}}{|\vec{d}\,||\vec{c}^{(i)}|}.$$

Координаты вектора $\vec{c}^{(i)}$ определяются в ходе обучения, которое проводится по каждой категории независимо от других.

4.2. Метод Rocchio

Некоторые классификаторы используют так называемый профайл (profile, прототип документа) для определения категории. Профайл - это список взвешенных термов, присутствие (или отсутствие) которых позволяет наиболее точно отличать конкретную категорию от других категорий. Метод, предложенный Дж. Роччио (J. Rocchio) [129], относится к линейным

классификаторам, в которых каждый документ представляется в виде вектора весовых значений термов. Профайл категории C_i будем рассматривать как вектор $\vec{c}^{(i)} = (c_1^{(i)}, \ ..., \ c_N^{(i)})$ (N – количество термов в словаре), значения элементов которого $c_k^{(i)}$ при обучении классификатора в рамках метода Rocchio рассчитывается по формуле:

$$c_k^{(i)} = \frac{\alpha}{|POS_i|} \sum_{d^{(j)} \in POS_i} w_k^{(j)} - \frac{\beta}{|NEG_i|} \sum_{d^{(j)} \in NEG_i} w_k^{(j)},$$

где $w_k^{(j)}$ - это вес терма t_k в документе $d^{(j)}$ (рассчитанный, например, по принципу *TF IDF*), POS_i - это положительный пример - множество документов, принадлежащих категории $\vec{c}^{(i)}$, т.е. $POS_i = \{d^{(j)} \mid \Phi(d^{(j)}, c_i) = 1\}$, а NEG_i - отрицательный пример - множество документов, не принадлежащих категории $\vec{c}^{(i)}$: $NEG_i = \{d_j \mid \Phi(d_j, c_i) = 0\}$. В этой формуле, α и β - контрольные параметры, которые характеризуют значимость положительных и отрицательных примеров. Например, если α = 1 и β = 0, C_i будет «центром масс» всех документов, относящихся к соответствующей категории.

Функция $CSV^{(i)}(d)$ в этом случае определяется по различным методикам - либо как величина обратная расстоянию от вектора из весовых значений термов, входящих в документ d, до профайла категории C_i, либо как скалярное произведение этих векторов.

4.3. Метод регрессии

Регрессионный анализ используется, когда признаки категорий могут быть выражены количественно в виде некоторой комбинации векторов весовых значений термов, входящих в документы из учебной коллекции. Полученная комбинация может использоваться для определения категории, к которой будет относиться новый документ.

Метод регрессии является вариантом линейной классификации. При применении регрессионного анализа к классификации текстов рассматривается

множество термов (F) и множество категорий (C). В этом случае учебной коллекции документов ставится в соответствие две матрицы:

- матрица документов D в учебной коллекции, в которой каждая строка – это документ, а столбец – терм;
- матрица ответов $O = \|o_{i,j}\|$, в которой строка i соответствует документу D_i ($i = 1, ..., N$), столбец j – категории ($j = 1, ..., M$), а $o_{i,j}$ - значению $CSV^{(i)}(d^{(i)})$.

Метод регрессии базируется на алгоритме нахождения матрицы правил M, которая минимизирует значение нормы матрицы $\|MD - O\|_F$, что формально записывается следующим образом:

$$M = \arg\min_M \|MD - O\|_F .$$

Напомним, что в линейной алгебре под нормой матрицы понимается функция, которая ставит в соответствие этой матрице некоторую числовую характеристику. В данном случае рекомендуется использовать норму Фробениуса $\|\cdot\|_F$, равную корню квадратному из суммы квадратов всех элементов соответствующей матрицы:

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}{}^2}.$$

Элемент $m_{i,j}$ искомой матрицы M будет отражать степень принадлежности i–го терма j–й категорий.

4.4. ДНФ-классификатор

ДНФ-классификатор состоит из множества правил, условия которых задаются некоторой ДНФ-формулой (ДНФ - дизъюнктивная нормальная форма), представляющей собой дизъюнкцию нескольких выражений, элементы которых соединены некоторым количеством (возможно, нулевым) конъюнкций. В этом случае документ относится к категории, если он удовлетворяет этой формуле, т.е. удовлетворяет хотя бы одному члену дизъюнкции.

На начальной стадии для каждой категории C_i, которая состоит из документов $\{d_1^{(i)}, ..., d_K^{(i)}\}$ определяется следующая формула:

ЕСЛИ $(x = d_1^{(i)})$ ИЛИ $(x = d_2^{(i)})$ ИЛИ . . . ИЛИ $(x = d_K^{(i)})$, ТО C_i.

Классификатор, основанный на таком множестве формул, абсолютно правильно работает на учебной коллекции, но, во-первых, он не может работать на других документах, во-вторых, пользоваться таким классификатором неудобно ввиду большого количества правил. В реально работающих ДНФ-классификаторах происходит переход от документов к множествам термов, которые определяются на основании анализа содержания документов, принадлежащих той или иной категории. Кроме того, проводится ряд упрощений, связанных с объединением или удалением некоторых условий. Приведем небольшой пример:

ЕСЛИ ((кофе И эспрессо) ИЛИ
(кофе И молоко) ИЛИ
(чай И стакан И лимон) ИЛИ
(кофе И чашка И-НЕ зерна))
ТО Напиток
ИНАЧЕ НЕ Напиток

Подобные действия улучшают показатель полноты классификации, но при этом может существенно пострадать точность даже на учебной коллекции.

4.5. Классификация на основе искусственных нейронных сетей

Искуственными нейронными сетями называются вычислительные структуры, моделирующие процессы, которые обычно происходят в мозгу человека. Мозг человека содержит около 10^{11} нейронов. Принято считать, что кадый нейрон в мозгу человека содержит тело, аксон, 10 000 дендритов и синапсы (рис. 12).

В настоящее время искусственные нейронные сети решают задачи классификации, кластеризации, распознавания образов, аппроксимации функций, прогноза и т.п.

Рис. 12. Нейрон

Модель отдельного нейрона можно рассматривать как некий компьютер, действительно, потенциал нейрона (аксона) - это функция от потенциала дендритов. Состояния нейрона - возбужденное или невозбужденное, определяются величиной этого потенциала.

4.5.1. Формальный нейрон

Рассмотрим формальную модель нейрона, приведенную на рис. 13. Входные сигналы через синапсы и дендриты поступают в тело нейрона, при этом их значения умножаются на весовые коэффициенты, соответствующие определенным синапсам (которые могут изменяться при обучении), затем результаты суммируются [16]. На основе полученной суммы (NET), к которой применяется некоторая функция F, называемая активизационной, формируется выходной сигнал нейрона OUT.

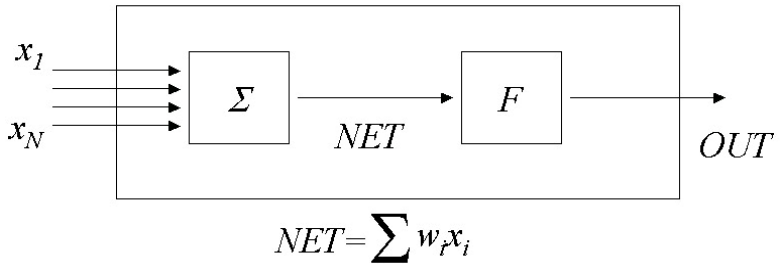


Рис. 13. Формальный нейрон

Сигнал сумматора определяется формулой:

$$NET = \sum_{i=1}^{n} x_i w_i,$$

где n – количество синапсов, индекс сигнала, x_i – входной сигнал i-го дендрита, w_i - весовые показатели, NET – сигнал сумматора.

Сигнал на выходе нейрона имеет следующий вид:

$OUT = F(NET)$.

Чаще всего применяются следующие активизационные функции:

$OUT = K \cdot NET$;

$$OUT = \begin{cases} 1, & NET > T, \\ -1, & NET \le T; \end{cases}$$

$$OUT = \frac{1}{1 + e^{-NET}};$$

$OUT = \text{th}(NET)$.

4.5.2. Искусственная нейронная сеть

Искуственная нейронная сеть - это ориентированный граф, вершины которого – нейроны, распределенные по слоям, а ребра - синапсы. Каждому ребру приписаны свой вес и функция проводимости.

По архитектуре связей можно выделить два класса нейронных сетей:

- сети прямого распространения (FeedForward) – однонаправленные с последовательными связями, где нейроны не имеют обратных связей;
- нейронные сети с обратными связями, где выход нейронов последующего слоя направляется к нейронам предыдущего слоя.

Элемементарную нейронную сеть прямого распространения принято называть перцептроном.

Первая версия перцептрона представляла собой однослойную нейронную сеть (рис. 14), предложенную в 1958 г. нейрофизиологом Ф. Розенблаттом (F. Rosenblatt). В перцептроне каждый нейрон связан через синаптический контакт со всеми рецепторами предыдущего слоя. Перцептрон Розенблатта был способен распознавать простейшие образы. Нейрон в этой модели вычисляет

взвешенную сумму элементов входного сигнала и пропускает результат через жесткую пороговую функцию, выход которой равен +1 или -1.

Ф. Розенблатт (1928 – 1971)

Типовой алгоритм обучения одношагового перцептрона имеет вид:

1. Инициализация синаптических весов w_i $(i = 1, ..., N)$ и сдвига b: синаптические веса принимают случайные значения.
2. Предъявление нейрону входного сигнала x_i $(i = 1, ..., N)$ и желаемого выходного сигнала d.
3. Вычисление выходного сигнала нейрона:

   $$y(t) = sign\left(\sum_{i=1}^{N} w_i(t)x_i(t) - b\right),$$

   где t - шаг итерации, b - сдвиг.
4. Настройка весовых значений:

   $$w_i(t+1) = w_i(t) + r[d(t) - y(t)]x_i(t), \quad i = 1, \ ..., \ N,$$

   где $w_i(t)$ - вес связи i –го элемента входного сигнала нейрона в момент t; r - скорость обучения; $d(t)$ - желаемый выходной сигнал.

   Если сеть принимает правильные решения, то синаптические веса не изменяются.
5. Переход к шагу 2.

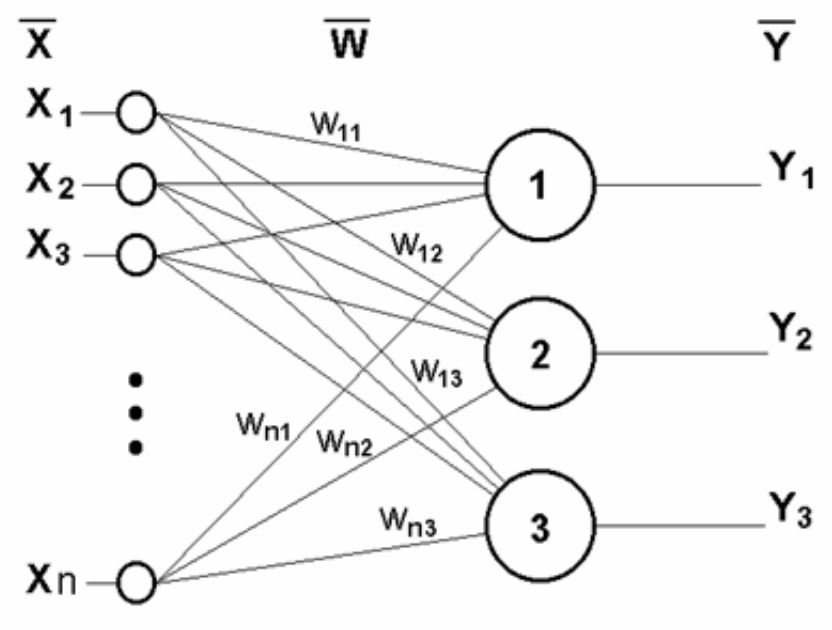


Рис. 14. Однослойный перцептрон:

$\bar{X}$ - входные сигналы; $\bar{W}$ - весовые значения синаптических контактов;

(1), (2), (3) –нейроны; $\bar{Y}$ - выходные сигналы

Перцептрон представляет интерес благодаря его адаптивности, которая используется в задачах распознавания образов. Однако было доказано, что однослойные нейронные сети не способны решать многих задач (например, реализовать логическую операцию исключающего ИЛИ). Для решения этих проблем используются многослойные нейронные сети. Многослойные сети могут образовываться каскадами слоев, в которых выход одного слоя является входом для следующего [45].

Нейронная сеть обучается, чтобы по некоторому множеству входных сигналов выдавать необходимое множество выходных сигналов. Каждое множество сигналов при этом рассматривается как вектор. Обучение осуществляется путем последовательного предъявления входных векторов с одновременной корректировкой весов. В процессе итеративного обучения по некоторым определенным правилам весовые значения становятся такими, что каждый входной вектор порождает необходимый выходной вектор.

4.5.3. Правила обучения перцептрона

Существует много методик обучения нейронных сетей, приведем одно из правил обучения однослойного перцептрона - дельта-правило (или правило Видрова-Хофа). Это правило базируется на простой идее непрерывного изменения синаптических весов для уменьшения разницы («дельты») между значениями желаемого и текущего выходного сигнала нейрона. Алгоритм дельта-правила следующий:

1. Подать на слой перцептрона сигнал $X = \{x_1, ..., x_N\}$ и вычислить $OUT = \{OUT_1, ..., OUT_M\}$.
2. Для всех $j = 1, ..., |X|$ вычислить $\delta_j = T_j - OUT_j$, где необходимое (желаемое) значение T_j задает учитель.
3. Если для всех j выполняется: $\delta_j = 0$, то обучение заканчивается.
4. Если существует j такое, что $\delta_j \neq 0$, то происходит корректировка w_{ij} следующим образом: $w_{ij}(s+1) = w_{ij}(s) + \Delta_{ij}$, где $\Delta_{ij} = \gamma \delta_j x_i$, j - номер нейрона, i - номер синапса, γ - коэффициент скорости обучения.
5. Переход к шагу 1.

Существует множество доступных программных реализаций искусственных нейронных сетей, например Toolbooks в таком пакете, как MatLab.

4.5.4. Нейронная сеть как классификатор

Классификатор может представлять собой нейронную сеть, входы которой соответствуют термам, а выходы - категориям. Для того чтобы классифицировать документ $d^{(j)}$, весовые значения его термов $w_k^{(j)}$ подаются на соответствующие входы сети; активация распространяется по сети, и значения, которые получены на выходах, являются результатами классификации. Типичный метод обучения такой сети - обратное распространение ошибки (back propagation). Если на одном из тренировочных документов получен неправильный ответ на одном из выходов, то ошибка распространяется обратно по сети, и весовые значения ребер

корректируются так, чтобы эту ошибку уменьшить.

4.6. Байесовский классификатор

4.6.1. Байесовская логистическая регрессия

В модели байесовской логистической регрессии рассматривается условная вероятность принадлежности документа D классу C: $P(C \mid D)$.

В рамках данной модели документ – это вектор: $D = (w_1,\ ...,\ w_N)$, где w_i - вес терма i, а N - размер словаря.

Модель байесовской логистической регрессии задается формулой:

$$P(C \mid D) = \varphi(\beta \cdot D) = \varphi(\sum_{i=1}^{N} \beta_i \cdot w_i),$$

где $C \in \{0,\ 1\}$, $\beta = \{\beta_1,\ ...,\ \beta_N\}$ - вектор параметров модели, а φ - логистическая функция, в качестве которой рекомендуется использовать:

$$\varphi(x) = \frac{1}{1 + \exp(-x)}.$$

Основная идея подхода состоит в том, чтобы использовать предшествующее распределение вектора параметров β, в котором каждое конкретное значение β_i с большой вероятностью может принимать значение, близкое к 0. При реальных расчетах принимаются гипотезы о Гауссовском или Лапласовом распределении значений β_i, а также то, что все величины β_i взаимно независимы.

4.6.2. Наивная байесовская модель

Рассматривается условная вероятность принадлежности объекта классу C при том, что он обладает признаками $F_1,\ ...,\ F_n$:

$P(C \mid F_1,\ ...,\ F_n)$.

В соответствии с теоремой Байеса:

$$P(C \mid F_1,\ ...,\ F_n) = \frac{P(C)P(F_1,\ ...,\ F_n \mid C)}{P(F_1,\ ...,\ F_n)}.$$

По определению условной вероятности:

$$P(C \mid F_1,\ ...,\ F_n) = P(C)P(F_1,\ ...,\ F_n \mid C) = P(c)P(F_1 \mid C)P(F_2,\ ...,\ F_n \mid C, F_1) =$$

$$= P(c)P(F_1 \mid C)P(F_2 \mid C)P(F_3,\ ...,\ F_n \mid C, F_1, F_2).$$

В соответствии с «наивным» байесовским подходом предполагается, что события F_i, F_j независимы для любых $i \neq j$:

$$P(F_i \mid C,\ F_j) = P(F_i \mid C).$$

Соответственно:

$$P(C \mid F_1,\ ...,\ F_n) = P(C)P(F_1 \mid C)P(F_2 \mid C) \cdot ... \cdot P(F_n \mid C) = P(C)\prod_{i=1}^{n} P(F_i \mid C).$$

Перейдем к классификации документов. В случае бинарной классификации «наивная» байесовкая вероятность принадлежности документа классу определяется по формуле:

$$P(D \mid C) = \prod_i P(w_i \mid C).$$

В соответствии с теоремой Байеса:

$$P(C \mid D) = \frac{P(C)}{P(D)} P(D \mid C).$$

Допустим, классификация происходит только по двум классам - C и $\bar{C}$. Тогда в соответствии с формулой Байеса имеем:

$$P(C \mid D) = \frac{P(C)}{P(D)} \prod_i P(w_i \mid C);$$

$$P(\bar{C} \mid D) = \frac{P(\bar{C})}{P(D)} \prod_i P(w_i \mid \bar{C}).$$

В качестве критерия принадлежности документа к категории рассматривается следующее отношение вероятностей принадлежности и не принадлежности классу C (аналогично статусу релевантности в вероятностной модели поиска):

$$\frac{P(C \mid D)}{P(\bar{C} \mid D)} = \frac{P(C)}{P(\bar{C})} \prod_i \frac{P(w_i \mid C)}{P(w_i \mid \bar{C})}.$$

На практике используется логарифм отношения вероятностей:

$$\ln\frac{P(C\mid D)}{P(\bar{C}\mid D)}=\ln\frac{P(C)}{P(\bar{C})}+\sum_i\ln\frac{P(w_i\mid C)}{P(w_i\mid\bar{C})}.$$

Если выполняется неравенство $\ln\frac{P(C\mid D)}{P(\bar{C}\mid D)}>0,$ (т.е., попросту, $p(C\mid D)>p(\bar{C}\mid D)$), то считается, что документ D относится к категории C.

4.6.3. Байесовский подход к решению проблемы спама

Метод Байеса широко используется для определения несанкционированных рекламных рассылок по электронной почте (спама). При этом рассматривается учебная база - два массива электронных писем, один из которых составлен из спама, а другой - из обычных писем. Для каждого из корпусов подсчитывается частота использования каждого слова, после чего вычисляется весовая оценка (от 0 до 1), которая характеризует условную вероятность того, что сообщение с этим словом является спамом. Значение веса, близкое к 0.5, не учитываются при интегрированном подсчете, поэтому слова с такими весами игнорируются и изымаются из словарей.

В соответствии с методом, предложенным П. Грэмом (P. Graham) [95], если сообщение содержит n слов с весовыми оценками $w_1,...,w_n$, то оценка условной вероятности того, что письмо является спамом, вычисляется по формуле:

$$Spm=\frac{\prod w_i}{\prod w_i+\prod(1-w_i)}.$$

Предполагается, что S – событие, заключающееся в том, что письмо – спам, A – событие, заключающееся в том, что письмо содержит слово t. Тогда, в соответствии с формулой Байеса, справедливо:

$$P(S\mid A)=\frac{P(A\mid S)P(S)}{P(A\mid S)P(S)+P(A\mid\bar{S})P(\bar{S})}.$$

Если сначала не известно, является ли письмо спамом или нет, исходя из опыта, знания соотношения спама и не-спама в учебной коллекции сообщений, предполагается, что $P(\bar{S})=\lambda P(S),$ откуда следует:

$$P(S \mid A) = \frac{P(A \mid S)}{P(A \mid S) + \lambda P(A \mid \overline{S})}.$$

Далее предполагается, что A_1 и A_2 – это события, состоящие в том, что письмо содержит слова t_1 и t_2. При этом вводится допущение, что эти события независимы («наивный» байесовский подход). Условная вероятность того, что письмо, содержащее оба слова (t_1 и t_2), является спамом, равна:

$$P(S \mid A_1 \,\&\, A_2) = \frac{P(A_1 \mid S)P(A_2 \mid S)}{P(A_1 \mid S)P(A_2 \mid S) + \lambda P(A_1 \mid \overline{S})P(A_2 \mid \overline{S})} =$$

$$= \frac{p(t_1)p(t_2)}{p(t_1)p(t_2) + \lambda(1 - p(t_1))(1 - p(t_2))}.$$

Частным случаем этой формулы на случай произвольного количества слов и $\lambda = 1$ и есть формула Грэма.

Следует отметить, что широкое применение находит именно значение $\lambda = 1$. Хотя это немного упрощает вычисление, но серьезно искажает действительность и снижает качество.

На практике на основе словарей, которые постоянно модифицируются, значение *Spm* рассчитывается для каждого сообщения. Если оно больше некоторого предельного, то сообщение считается спамом.

4.6.4. Определение тональности сообщений

Традиционная экспертная оценка текстовых сообщений оказывается не эффективной для больших и динамичных текстовых массивов. Один из аспектов анализа текстов сообщений из современных информационных потоков - это оценка так называемой тональности или эмоциональной окраски. Под тональностью текста в данном случае понимается позитивная, негативная или нейтральная эмоциональная окраска как всего текстового документа, так и отдельных его частей, имеющих отношения к определенным понятиям, таким как персоны, организации, бренды и т.п.

Описываемый ниже метод основывается на статистических закономерностях, связанных с присутствием определенных термов в текстах, наивном байесовском подходе и методе нейронных сетей (реализации двухслойного перцептрона).

Необходимо отметить, что задача определения тональности сообщений более сложна, чем выявление спама на основе анализа текстов. В то время как выявление спама подразумевает лишь две гипотезы (спам, не спам), то в задаче определения тональности проверяется как минимум три: эмоциональная окраска позитивная, негативная, нейтральная и, зачастую, существует потребность также в проверке комбинации этих гипотез (например, для выявления уровня «экспрессивности» текста).

С другой стороны, в отличие от проблемы выявления спама, где оценка отдельных документов может быть близка к однозначной, в случае определения тональности сообщений разные эксперты порой не приходят к единому мнению.

В случае оценки тональности сообщений пространство гипотез будет содержит: H_{-1} – тональность отрицательная, H_0 - тональность нейтральная и H_1 - тональность положительная. Для упрощения рассмотрим события такого типа: H_1 – тональность положительная, $\overline{H_1}$ – тональность не положительная. Из корпуса документов с положительной тональностью выбираются термы t, характерные для этих документов, со значениями $p(t|H_1)$, превышающими 1/2. Таким же образом выбираются термы и для документов с отрицательной тональностью. Выбранные термы принято называть тонально-окрашенными или просто тональными, несущими в себе оценочную семантику.

Для упрощения модели предположим, что для всех выбранных термов вес будет одинаковым, равным α (может изменяться при обучении модели). Тогда формула для вычисления функции Spm примет вид:

$$Spm(x) = \frac{\alpha^x}{\alpha^x + \lambda(1-\alpha)^x},$$

где x – количество весомых с точки зрения тональности (положительной или отрицательной) термов в информационном сообщении.

Соответственно, для оценки гипотезы об отрицательной тональности сообщения (H_{-1}) может использоваться словарь слов «отрицательной тональности» и та же формула. Вместе с тем, поскольку положительная и отрицательная тональности являются своего рода антагонизмами, окончательное решение о тональности сообщения принимается с учетом разности значений весовых оценок гипотез H_1 и H_{-1}. Пороговое значение этой величины - β определяется в процессе настройки (обучения) системы.

Необходимо сделать еще одно, диктуемое практикой, замечание. Следует учитывать, что отрицательная тональность сообщений почти всегда выражена более явно, чем положительная. Поэтому при расчете веса отрицательной тональности значение x в приведенной выше формуле несколько уменьшается путем умножения его на эмпирически определяемую константу $\gamma \in (0,\ 1)$.

В некоторых случаях определенный интерес для аналитиков представляют документы, у которых достаточно высоки значения весов как положительной, так и отрицательной тональности. Заметим, что разница этих весов может быть минимальной, т.е. документ может характеризоваться как нейтрально окрашенный. Вместе с тем он может получить дополнительную характеристику «экспрессивной» тональности.

Алгоритм определения тональности можно представить в виде нейронной сети. Первый слой этой сети составляют два нейрона - определители весовых значений положительной и отрицательной тональности (положительный и отрицательный нейроны). Можно предположить, что количество синапсов каждого нейрона соответствует количеству значимых для определения тональности сообщений слов из словаря естественного языка.

На вход нейронов поступают слова (рис. 15). При этом $x_i = 1$, если на вход поступило слово из словаря с номером i, в противном случае $x_i = 0$. Весовые значения (вес синапсов), которые соответствуют этим словам, равны $w_1^+, \ldots, w_n^+$ для положительного нейрона и $w_1^-, \ldots, w_n^-$ - для отрицательного. Именно эти весовые значения могут изменяться в процессе обучения первого слоя нейрона. Сумматоры подсчитывают значения NET^+ и NET^- соответственно.

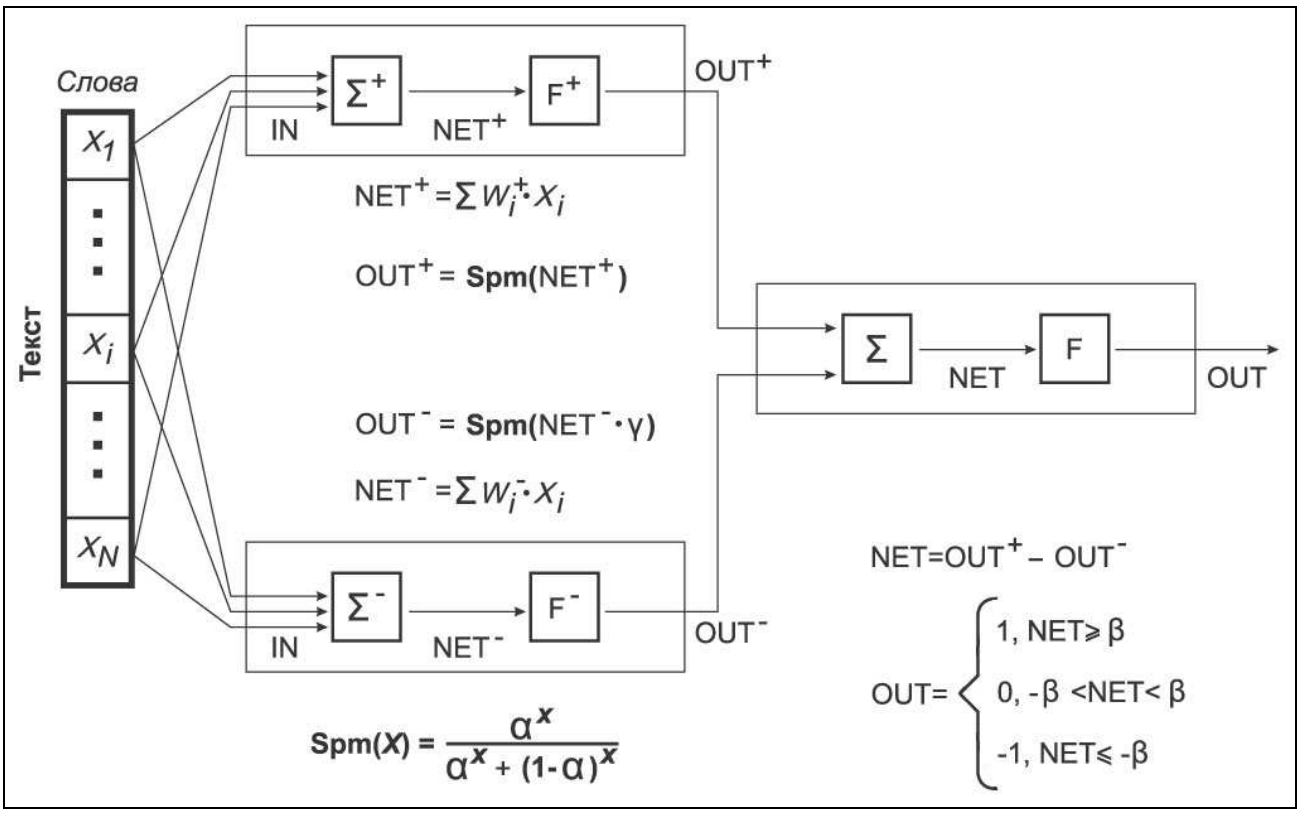


Рис. 15. Двухслойная нейронная сеть для определения тональности текста

Проводимость нейронов рассчитывается по приведенной выше формуле, аргументом в которой выступает значение NET^+ для положительного нейрона и γ NET^- – для отрицательного. Оба нейрона выдают через аксоны значения, OUT^+ и OUT^-, которые являются входными сигналами для нейрона второго уровня, сумматор которого вычисляет разность OUT^+ и OUT^-, а функция проводимости выдает градиентный результат по условию, приведенному на рис. 15.

4.7. Метод опорных векторов

Метод опорных векторов (Support Vector Mashine, SVM), предложенный В.Н. Вапником [146, 84], относится к группе граничных методов классификации. Он определяет принадлежность объектов к классам с помощью границ областей. Будем рассматривать только бинарную классификацию, т.е. классификацию только по двум категориям c и $\overline{c}$, принимая во внимание то, что этот подход

может быть расширен на любое конечное количество категорий. Предположим, что каждый объект классификации является вектором в *N*-мерном пространстве. Каждая координата вектора - это некоторый признак, количественно тем больший, чем больше этот признак выражен в данном объекте.

В.Н. Вапник

Предполагается, что существует учебная коллекция - это множество векторов $\{\vec{x}_1, ..., \vec{x}_n\} \in R^N$ $\{x_1, ..., x_n\} \in R^N$ и чисел $\{y_1, ..., y_n\} \in \{-1, 1\}$. Число y_i равно 1 в случае принадлежности соответствующего вектора x_i категории c, и -1 - в противном случае. Как было показано выше, линейный классификатор - это один из простейших способов решения задачи классификации. В этом случае ищется прямая (гиперплоскость в *N*-мерном пространстве), отделяющая все точки одного класса от точек другого класса. Если удается найти такую прямую (рис. 16), то задача классификации сводится к определению взаимного расположения точки и прямой: если новая точка лежит с одной стороны прямой (гиперплоскости), то она принадлежит классу c, если с другой – классу $\overline{c}$.

Формализуем эту классификацию: необходимо найти вектор $\vec{w}$ такой, что для некоторого значения b и новой точки $\vec{x}_i$ выполняется:

$$y_i = \begin{cases} +1, & \text{если } \vec{w} \cdot \vec{x}_i \geq b, \\ -1, & \text{если } \vec{w} \cdot \vec{x}_i < b, \end{cases}$$

где $\vec{w} \cdot \vec{x}_i$ - скалярное произведение векторов $\vec{w}$ и $\vec{x}_i$:

$$\vec{w} \cdot \vec{x}_i = \sum_{j=1}^{N} w_j x_{i,j} .$$

$\vec{w} \cdot \vec{x}_i = b$ - уравнение гиперплоскости, которая разделяет классы. То есть, если скалярное произведение вектора $\vec{w}$ на $\vec{x}_i$ не меньше значения b, то новая точка принадлежит первому классу, если меньше – ко второму. Вектор $\vec{w}$ перпендикулярен искомой разделяющей гиперплоскости, а значение b зависит от кратчайшего расстояния между этой гиперплоскостью и началом координат. Возникает вопрос, какая из гиперплоскостей разделяет классы лучше всего?

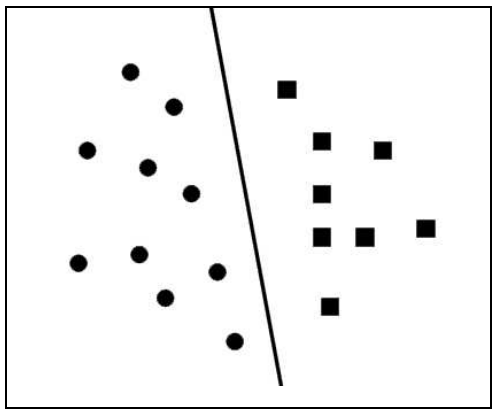


Рис. 16. Разделение классов прямой

Метод SVM базируется на таком постулате: наилучшая разделяющая прямая – это та, которая максимально далеко отстоит от ближайших до нее точек обоих классов. То есть задача метода SVM состоит в том, чтобы найти такие вектор $\vec{w}$ и число b, чтобы для некоторого $\varepsilon > 0$ (половина ширины разделяющей поверхности) выполнялось:

$$\begin{cases} \vec{w} \cdot \vec{x}_i \geq b + \varepsilon \;\Rightarrow\; y_i = +1, \\ \vec{w} \cdot \vec{x}_i \leq b - \varepsilon \;\Rightarrow\; y_i = -1. \end{cases}$$

Умножим после этого обе части неравенства на $1/\varepsilon$ и, не ограничивая общности, выберем ε равным единице. Таким образом, для всех векторов $\vec{x}_i$ из учебной коллекции будет справедливо:

$$\begin{cases} \vec{w} \cdot \vec{x}_i - b \geq +1, \text{ если } y_i = +1, \\ \vec{w} \cdot \vec{x}_i - b \leq -1, \text{ если } y_i = -1. \end{cases}$$

Условие $-1 < \vec{w} \cdot \vec{x}_i - b < 1$ задает полосу, которая разделяет классы (рис. 17).

Границами полосы являются две параллельные гиперплоскости с направляющим вектором $\vec{w}$. Точки, ближайшие к разделяющей гиперплоскости, расположены точно на границах полосы.

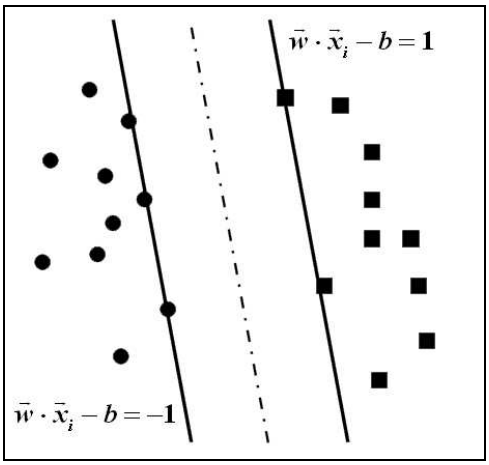


Рис. 17. Разделяющая полоса

Чем шире полоса, тем увереннее можно классифицировать документы, соответственно, в методе SVM считается, что самая широкая полоса является наилучшей.

Сформулируем условия задачи оптимальной разделяющей полосы, определяемой неравенством: $y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1$ (так переписывается система уравнений, исходя из того, что $y_i \in \{-1, 1\}$). Ни одна из точек обучающей выборки не может лежать внутри этой разделяющей полосы. При этих ограничениях $\vec{x}_i$ и y_i - постоянные как элементы учебной коллекции, а $\vec{w}$ и b - переменные.

Легко видеть, что ширина разделяющей полосы равна $2/\|\vec{w}\|$. Поэтому необходимо найти такие значения $\vec{w}$ и b, чтобы выполнялись приведенные линейные ограничения, и при этом как можно меньше была норма вектора $\vec{w}$, то есть необходимо минимизировать:

$$\|\vec{w}\|^2 = \vec{w} \cdot \vec{w}.$$

Это известная задача квадратичной оптимизации при линейных ограничениях.

Если предположить, что на учебных документах возможно были допущены ошибки экспертами при классификации, то необходимо ввести набор

дополнительных переменных $\xi_i \geq 0$, характеризующих величину ошибок на объектах $\{\vec{x}_1, \ ..., \ \vec{x}_n\}$. Это позволяет смягчить ограничения:

$$y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1 - \xi_i.$$

Предполагается, что если $\xi_i = 0$, то на документе $\vec{x}_i$ ошибки нет. Если $\xi_i > 1$, то на документе $\vec{x}_i$ допускается ошибка. Если $0 < \xi_i < 1$, то объект попадает внутрь разделяющей полосы, но относится алгоритмом к своему классу.

Задачу поиска оптимальной разделяющей полосы можно в этом случае переформулировать следующим образом минимизировать сумму:

$$\|\vec{w}\|^2 + C\sum_i \xi_i$$

при ограничениях $y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1 - \xi_i$, где коэффициент C - параметр настройки метода, который позволяет регулировать соотношение между максимизацией ширины разделяющей полосы и минимизацией суммарной ошибки. Приведенная задача является задачей квадратичного программирования, которую можно переписать в следующем виде:

$$\begin{cases} \dfrac{\|w\|^2}{2} + C\sum_i \xi_i \to \min; \\ y_i(\vec{w} \cdot \vec{x}_i - b) + \xi_i \geq 1, \ \ i = 1, \ ..., \ n. \end{cases}$$

По известной теореме Куна-Такера такая задача эквивалентна двойственной задаче поиска седловой точки функции Лагранжа:

$$\begin{cases} \dfrac{1}{2}\vec{w} \cdot \vec{w} + C\sum_i \xi_i - \sum_i \lambda_i(\xi_i + y_i(\vec{w} \cdot \vec{x}_i - b) - 1) \to \min\limits_{w,b} \ \max\limits_{\lambda}; \\ \xi_i \geq 0, \ \ \lambda_i \geq 0, \ \ i = 1, \ ..., \ n. \end{cases}$$

Необходимым условием метода Лагранжа является равенство нулю производных лагранжиана по переменным $\vec{w}$ и b, откуда получаем:

$$\vec{w} = \sum_{i=1} \lambda_i y_i \vec{x}_i,$$

т.е. искомый вектор – это линейная комбинация учебных векторов, для которых $\lambda_i \neq 0$. Если $\lambda_i > 0$, то документ обучающей коллекции называется опорным вектором.

Таким образом, уравнение разделяющей плоскости имеет вид:

$$\sum_{i=1} \lambda_i y_i \vec{x}_i \cdot \vec{x}_i - b = 0 .$$

Приравняв производную лагранжиана по b нулю, получим:

$$\sum_{i=1} \lambda_i y_i = 0.$$

Подставляя последнее выражение и выражение для $\vec{w}$ в лагранжиан получим эквивалентную задачу квадратичного программирования, содержащую только двойственные переменные:

$$\begin{cases} \sum_i \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j (\vec{x}_i \cdot \vec{x}_j) \to \min\limits_{\lambda}; \\ \sum_{i=1} \lambda_i y_i = 0; \\ C \geq \lambda_i \geq 0, \;\; i = 1, \; ..., \; n. \end{cases}$$

При этом существенно, что целевая функция зависит не от конкретных значений $\vec{x}_i$, а от скалярных произведений между ними. Следует заметить, что целевая функция является выпуклой, поэтому любой ее локальный минимум является глобальным.

Метод классификации разделяющей полосой имеет два недостатка:

- при поиске разделяющей полосы важное значение имеют только пограничные точки;
- во многих случаях найти оптимальную разделяющую полосу невозможно.

Для улучшения метода применяется идея расширенного пространства, для чего:

1. Выбирается отображение $\phi(\vec{x})$ векторов $\vec{x}$ в новое, расширенное пространство.
2. Автоматически применяется новая функция скалярного произведения, которая применяется при решении задачи квадратичного программирования, так называемую функцию ядра (kernel function): $K(\vec{x}, \vec{z}) = \phi(\vec{x}) \cdot \phi(\vec{z})$. На практике обычно выбирают не отображение

$\phi(\vec{x})$, а сразу функцию $K(\vec{x},\vec{z})$, которая могла бы быть скалярным произведением при некотором отображении $\phi(\vec{x})$. Функция ядра - главный параметр настраивания машины опорных векторов.

3. Определяется разделяющая гиперплоскость в новом пространстве: с помощью функции $K(\vec{x},\vec{z})$ устанавливается новая матрица коэффициентов для задачи оптимизации. При этом вместо скалярного произведения $\vec{x}_i \cdot \vec{x}_j$ берется значение $K(\vec{x}_i,\vec{x}_j)$, и решается новая задача оптимизации.
4. Найдя $\vec{w}$ и b, получаем поверхность, которая классифицирует, $\vec{w}\cdot\phi(\vec{x})-b$ в новом, расширенном пространстве.

Ядром может быть не всякая функция, однако класс допустимых ядер достаточно широк. Например, в системе классификации новостного контента с применением известного пакета LibSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm) [124] в качестве функции ядра рекомендуется использовать радиальную базисную функцию:

$$K(\vec{x},\vec{z}) = \exp(-\gamma\left\|\vec{x}-\vec{z}\right\|^2),$$

где γ - настраиваемый параметр.

Рассмотрим наглядный пример перехода к расширенному пространству, изображенный на рис. 18. Как видно, круглые и квадратные фигуры не разделяются линейной полосой. Если же «изогнуть» пространство, перейдя к третьему измерению, то эти фигуры можно разделить плоскостью, которая отрезает часть поверхности с квадратными точками. Таким образом, выгнув пространство с помощью отображения $\phi(x)$, можно найти разделяющую гиперплоскость.

Метод SVM обладает определенными преимуществами:

- на тестах с документальными массивами превосходит другие методы;
- при выборах разных ядер позволяет реализовать другие подходы. Например, большой класс нейронных сетей можно представить с помощью SVM со специальными ядрами;
- итоговое правило выбирается не с помощью экспертных эвристик, а

путем оптимизации некоторой целевой функции.

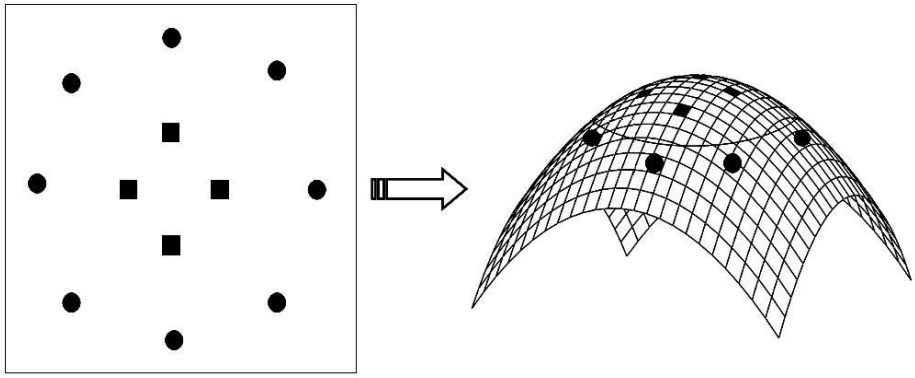


Рис. 18. Пример перехода к расширенному пространству

К недостаткам метода можно отнести:

- иногда слишком малое количество параметров для настройки: после того как фиксируется ядро, единственным изменяемым параметром остается коэффициент ошибки C;
- нет четких критериев выбора ядра;
- медленное обучение системы классификации.

4.8. Оценка качества классификации

Для оценки качества классификации используются характеристики из области информационного поиска. Рассмотрим случай бинарной классификации. Пусть c - категория, а $\overline{c}$ - ее дополнение. Классификатор определяет, принадлежит ли документ к c или нет. Таблица возможных результатов имеет вид:

Категория c		Принадлежность к категории	
		ДА	НЕТ
Результат классификации	ДА	True positive (TP)	False positive (FP)
	НЕТ	False negative (FN)	True negative (TN)

Здесь true positive (TP) - количество документов, правильно отнесенных к

категории, false positive (FP) - количество документов, неправильно отнесенных к категории; false negative (FN) и true negative (TN) определяются аналогично. Принято говорить false negative – это ошибка классификации первого рода, а false positive – ошибка второго рода.

Для анализа качества классификации используются показатели полноты и точности. Полнота π - это часть найденных документов из категории среди всех документов этой категории:

$$\pi = \frac{TP}{TP + FN}.$$

Точность ρ - это доля найденных документов из категории среди всех документов, которые отнесены в эту категорию классификатором (сколько документов, отнесенных классификатором к категории на самом деле ей принадлежит):

$$\rho = \frac{TP}{TP + FP}.$$

Полнота и точность находят широкое применение для оценки качества классификации в ходе обучения. Два разных метода можно корректно сравнивать по полноте и точности, если сравнения проводятся на одной и той же коллекции (одинаковые документы и категории для обоих методов), и коллекция одинаково разделена на учебные составляющие.

Существует и неявный способ оценки. Классификатор сравнивается с некоторым эталонным классификатором при соблюдении вышеперечисленных условий.

5. ЭЛЕМЕНТЫ КЛАСТЕРНОГО АНАЛИЗА

«Возьмемся за руки, друзья,
чтоб не пропасть поодиночке…»
Булат Окуджава

Все рассмотренные выше модели информационного поиска имеют общий недостаток, связанный с большими размерностями (определяемым, как правило, количеством термов). Для обеспечения эффективной работы поисковых систем необходимо группирование как термов, так и тематически подобных документов. Только в этом случае может быть обеспечена обработка современных больших и динамичных информационных массивов в режиме реального времени.

При рассмотрении тематических каталогов, построенных при участии людей (например, Yahoo! или Open Directory) возникает естественный вопрос: могут ли они быть построены автоматически? Один из путей решения этой проблемы – применение кластерного анализа, т.е. методики автоматического группирования данных в классы. При этом документы, которые попадают в один класс, в некотором смысле должны быть ближе друг к другу (например, по тематике), чем к документам из других классов.

С одной стороны, при кластеризации гипертекстовых документов возникают некоторые сложности, связанные с множественностью выбора алгоритмов этого процесса. Разные методологии используют разные алгоритмы подобия документов при наличии большого количества признаков (вместе с тем в случае работы с HTML-документами возникают возможности учета гипертекстовой разметки для выявления текстовых блоков, тегов разметки, имен доменов, URL-адресов, адресных подстрок и т.п.). С другой стороны, как только методами кластерного анализа определяются классы, возникает необходимость их сопровождения, так как веб-пространство постоянно растет. В этом случае на помощь приходит классификация. Механизм классификации обычно обучается на отобранных документах только после того, как заканчивается стадия обучения путем автоматической кластеризации - разбиения множества документов на

классы (кластеры), смысловые параметры которых заранее неизвестны. Количество кластеров может быть произвольным или фиксированным. Если классификация допускает приписывание документам определенных, известных заранее признаков, то кластеризация более сложный процесс, который допускает не только приписывание документам некоторых признаков, но и выявление самых этих признаков как основ формирования классов [25]. Цель методов кластеризации массивов документов состоит в том, чтобы подобие документов, которые попадают в кластер, было максимальным. Поэтому методы кластерного анализа базируются на таких определениях кластера, как множества документов, значение семантической близости между любыми двумя элементами которых (или значение близости между любым документом этого множества и центром кластера) не меньше определенного порога.

Для численного определения значения близости между документами в кластерном анализе используются такие основные правила определения расстояния (метрики), как метрика Минковского:

$$D_p(\vec{x},\vec{y}) = \left(\sum_{k=1}^{N} (x_k - y_k)^p \right)^{1/p},$$

частным случаем при $p = 2$ которого является Евклидова метрика:

$$D_p(\vec{x},\vec{y}) = \sqrt{\sum_{k=1}^{N} (x_k - y_k)^2}.$$

Для группирования документов, представленных в виде векторов весовых значений входящих в них термов, часто используется скалярное произведение весовых векторов:

$$Sim(\vec{x},\vec{y}) = \vec{x} \cdot \vec{y},$$

где $\vec{x}$, $\vec{y}$ - векторы, соответствующие документам, элементами которых являются весовые значения термов, которые, как правило, определяются в результате анализа большого массива документов. Для проведения такого анализа используются разные подходы - весовой, вероятностный, семантический и т. д.

В области информационного поиска кластерный анализ чаще всего применяется для решения двух задач - группирования документов в базах данных (информационных массивах) и группирования результатов поиска.

Для статических документальных массивов методы кластерного анализа в настоящее время получили значительное развитие [111, 134, 32]. Вместе с тем открытым остается вопрос применения этих методов к информационным потокам, которым присущи большие объемы и динамика [31].

Методы кластерного анализа находят широкое применение в процедурах ранжирования откликов информационно-поисковых систем, при построении персонализированных поисковых интерфейсов и папок поиска.

5.1. Латентно-семантический анализ

5.1.1. Матричный латентно-семантический анализ

Метод кластерного анализа LSA/LSI (от англ. Latent Semantic Analysis/Indexing - метод латентно-семантического анализа/индексирования) [106] базируется на сингулярном разложении матриц (SVD) [93]. Пусть массиву документов $D=\{d^{(j)} \mid j=1, \ ..., \ n\}$ ставится в соответствие матрица A, строки которой соответствуют документам, а столбцы – весовым значениям термов (размер словаря термов - m). Сингулярным разложением матрицы A ранга r размерности $m \times n$ называется ее разложение вида $A=USV^T$, где U и V – ортогональные матрицы размерности $m \times r$ и $r \times n$, соответственно, а S – диагональная матрица, диагональные элементы которой неотрицательны ($s_{i,i} \geq 0$). Диагональные элементы матрицы S называют сингулярными значениями матрицы A. Заметим, что матрица S, в отличие от матрицы A, квадратная.

Приведенное выше разбиение матрицы A обладает тем свойством, что если в матрице S оставить только k наибольших сингулярных значений (обозначим такую матрицу как S_k), а в матрицах U и V – только соответствующие этим значениям колонки (соответственно, матрицы U_k, V_k), то матрица $A_k = U_k \cdot S_k \cdot V_k^T$

будет наилучшей по Фробениусу аппроксимацией исходной матрицы A матрицей с рангом, не превышающим k. Напомним, что норма матрицы X размерности $M \cdot N$ по Фробениусу определяется выражением:

$$\|X\|_F = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} x_{ij}^2}.$$

Указанное выше свойство можно перефразировать следующим образом, A_k будет именно той матрицей ранга k, которая минимизирует норму матрицы $\|A - A_k\|_F$, что можно записать в обозначениях, принятых в методах оптимизации:

$$A_k = \arg\min_{X:rank(X)=k} \|A - X\|_F .$$

В соответствии с методом LSA в рассмотрение берутся не все, а лишь k наибольших сингулярных значений матрицы A, и каждому такому значению ставится в соответствие один кластер.

A_k определяет k-мерное факторное пространство, на которое проецируются как документы (с помощью матрицы V), так и термины (с помощью матрицы U). В полученном факторном пространстве документы и термины группируются в массивы (кластеры), имеющие некоторый общий смысл, не заданный в явном виде, т.е. латентный.

Выбор наилучшего значения k для LSA - это проблема отдельных исследований. В идеале, k должно быть достаточно велико для отображения всей реально существующей структуры данных, но в то же время достаточно мало, чтобы не учитывать случайных зависимостей.

В практике информационного поиска особое значение отводится матрицам U_k и V_k^T. Как указывалось ранее, строки матрицы U_k рассматриваются как образы термов в $k-$мерном вещественном пространстве. Аналогично, столбцы матрицы V_k^T рассматриваются как образы документов в том же $k-$мерном пространстве. Иными словами, эти векторы задают искомое представление термов и документов в $k-$мерном пространстве скрытых факторов.

Существуют также методы инкрементного обновления всех значений, используемых в LSA. При пополнении новым документом d (например, новым

результатом поиска по запросу) информационного массива, для которого уже проведено сингулярное разложение, можно не выполнять разложение заново. Достаточно аппроксимировать его, вычисляя образ нового документа на основе ранее вычисленных образов термов и весов факторов. Пусть $d-$ вектор весов термов нового документа (новый столбец матрицы A), тогда его образ можно вычислить по формуле: $d' = S_k^{-1} U_k^T d$.

Если q – вектор запроса, i-й элемент которого равен 1, когда терм с номером i входит в запрос, и 0 - в противном случае, то образ запроса q в пространстве латентных факторов будет иметь вид: $q' = q^T U_k S_k^{-1}$.

В этом случае мера близости запроса q и документа d оценивается величиной скалярного произведения векторов q' и $V_k^T\{d\}$ (здесь $V_k^T\{d\}$ обозначает $d-$й столбец матрицы V_k^T).

При информационном поиске, в результате того, что отбрасываются наименее значимые сингулярные значения, формируется пространство ортогональных факторов, играющих роль обобщенных термов. В результате происходит «сближение» документов из близких по содержанию предметных областей, частично решаются проблемы синонимии и омонимии термов.

Метод LSA широко применяется при ранжировании выдачи информационно-поисковых систем, основанных на цитировании. Это алгоритм HITS (Hyperlink Induced Topic Search) – один из двух самых известных в области информационного поиска. Метод LSA не нуждается в предварительной настройке на специфический набор документов, вместе с тем позволяет качественно выявлять скрытые факторы. К недостаткам метода можно отнести невысокую производительность. Скорость вычисления SVD соответствует порядку $O(N^2 \cdot k)$, где $N = |D| + |T|$, D – множество документов, T – множество термов, k – размерность пространства факторов.

LSA также не предусматривает возможность пересечения кластеров, что противоречит практике. Кроме того, ввиду своей вычислительной трудоемкости метод LSA применяется только для относительно небольших матриц.

5.1.2. Вероятностный латентно-семантический анализ

Вероятностный латентно-семантический анализ (от англ. Probabilitstic Latent Semantic Analysis, PLSA) - это модификация LSA, построенная на использовании вероятностного подхода. Метод PLSA также предназначен для выявления скрытых факторов, присутствующих в информационном массиве и связанных с ними документов и слов.

Как и в предыдущем случае, предполагается, что существует k скрытых факторов $z_1,...,z_k$ (число k задается заранее). Фактору z_i сопоставляется вероятность $P(z_i)$ того, что случайно выбранный из данной коллекции документ наиболее точно характеризуется данным фактором $(\sum_{i=1}^{k} P(z_i)=1)$.

Обозначим через $P(d \mid z_i)$ – вероятность того, что для выбранного фактора z_i из множества фактов Z, именно документ d из всего множества документов D лучше всего характеризуется этим фактором. Тогда $\sum_{d\in D} P(d \mid z_i)=1.$ Аналогично обозначим через $P(t \mid z_i)$ вероятность того, что для выбранного фактора z_i, из всех термов именно терм t из словаря системы T лучше всего характеризуется этим фактором z_i. Тогда $\sum_{t\in T} P(t \mid z_i)=1.$

Вероятность того, что при случайном выборе документа d и терма t, терм t встретится в документе d, можно определить с одной стороны (рис. 19 а, ассиметричная параметризация), как:

$$P(d,t)=P(d)P(t \mid d),$$

$$P(t \mid d)=\sum_{i=1}^{k} P(t \mid z_i)P(z_i \mid d).$$

С другой стороны, эта же вероятность при симметричной параметризации представляется (рис. 19 б) как:

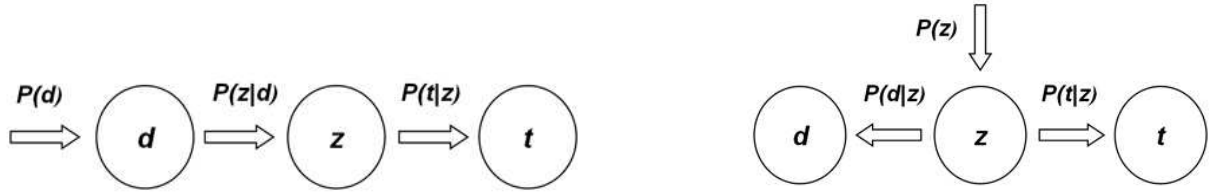
$$P(d,t)=\sum_{i=1}^{k} P(z_i)P(d \mid z_i)P(t \mid z_i).$$

Зафиксировав число скрытых факторов k, с помощью метода PLSA можно оценить следующие величины:

$P(z_i)$ – вероятность того, что случайно выбранный из коллекции документ в соответствует фактору z_i;

$P(d_j \mid z_i)$ – вероятность того, что документ d_j попадет в группу документов, соответствующих фактору z_i;

$P(t_j \mid z_i)$ – вероятность того, что терм t_j попадет в группу слов, связанных с фактором z_i.



а) *б*)

Рис. 19. Графическое представление модели: а) - ассиметричная; б) - симметричная параметризация

Для оценки приведенных выше вероятностей на контрольном массиве документов определяется наблюдаемая частота вхождения терма t в документ d, традиционно обозначаемая как $tf(d,t)$.

Упомянутые выше вероятности определяются исходя из условия максимизации функции максимального правдоподобия:

$$L = \sum_{d \in D} \sum_{t \in T} tf(d,t) \log P(d,t),$$

где внешняя сумма берется по всем документам, а внутренняя по всем термам словаря.

В PLSA используется алгоритм EM (Expectation Maximization – оценочной максимизации), в котором на каждом шагу выполняются два шага – 1) оценивание, при котором вычисляются и оцениваются послеопытные

вероятности латентных переменных, и 2) максимизация, в результате которой параметры изменяются [101].

На первом шаге оценивается:

$$P(z \mid d,t) = \frac{P(z)P(d \mid z)P(t \mid z)}{\sum_{z' \in Z} P(z')P(d \mid z')P(t \mid z')},$$

после чего выполняется шаг максимизации L на основе вычисления:

$$P(t \mid z) \propto \sum_{d \in D} tf(d,t)P(z \mid d,t),$$

$$P(d \mid z) \propto \sum_{t \in T} tf(d,t)P(z \mid d,t),$$

$$P(z) \propto \sum_{d \in D} \sum_{t \in T} tf(d,t)P(z \mid d,t).$$

Данный алгоритм обеспечивает сходимость функции L к некоторому локальному максимуму. Эксперименты показывают, что сходимость достигается после нескольких десятков итераций.

Покажем, как представить PLSA в виде матричной записи. Определим матрицы: 1) $\hat{U}$, элементами которой $\hat{u}_{i,k}$ будут условные вероятности $P(d^{(i)} \mid z_k)$, 2) $\hat{V}$, элементами которой $\hat{v}_{j,k}$ будут условные вероятности $P(t_j \mid z_k)$, 3) $\hat{S}$ - диагональную матрицу ранга k, на диагонали которой будут размещены значения вероятностей $P(z_i)$. Объединенная вероятностная модель $P(z)$ аппроксимируется выражением $\hat{U}\hat{S}\hat{V}^T$. Сравнивая это разложение с SVD, можно заметить, что $\hat{U}$ и $\hat{V}$ в PLSA также независимы ввиду предположения независимости термов и документов. Хотя приведенное разложение не является сингулярным, вместе с тем, k наибольших компонент $\hat{S}$ определяют правила кластеризации PLSA. Основное отличие PLSA от LSA заключается в выборе целевой функции L.

5.2. Метод *k*-means

Итеративный алгоритм кластерного анализа *k*-means (*k*-средних) группировки документов $\{\vec{d}^{(1)}, ..., \vec{d}^{(N)}\}$ (как и в векторно-пространственной модели информационного поиска документы представляют собой векторы, координаты которых - весовые значения термов) по фиксированному количеству

кластеров заключается в следующем: случайным образом выбирается k документов, которые определяются как центроиды (наиболее типичные представители) кластеров. Т.е. каждый кластер C_j ($j = 1, ..., k$) представляется вектором $\vec{C}_j$, соответствующим центроиду. Затем k кластеров наполняются – для каждого из $N - k$ документов, которые остались, некоторым образом определяется близость к центроиду соответствующего кластера. Близость может определяться различными способами, например, как нормированное скалярное произведение:

$$Sim(\vec{d}, \vec{C}_j) = \frac{\vec{d} \cdot \vec{C}_j}{|\vec{d}||\vec{C}_j|}.$$

После этого документ приписывается к тому кластеру, значение $Sim(\vec{d}, \vec{C}_j)$ для которого оказывается наибольшим. Далее для каждого из новых кластеров заново определяется центроид $\vec{C}_j$ ($j = 1, ..., k$) - вектор, координаты которого определяются, например, как среднее арифметическое соответствующих весовых документов, входящих в данный кластер.

После этого снова осуществляется процесс наполнения кластеров, затем вычисление новых центроидов и т.д., пока процесс формирования кластеров не стабилизируется (или, если уменьшение суммы расстояния от каждого элемента до центра его кластера не станет меньше некоторого заданного порогового значения).

Алгоритм *k*-means максимизирует функцию качества кластеризации Q:

$$Q(C_1, ..., C_k) = \sum_{j=1}^{k} \sum_{\vec{d} \in C_j} Sim(\vec{d}, \vec{C}_j).$$

В отличие от метода LSI, *k*-means может использоваться для группирования динамических информационных потоков благодаря своей вычислительной простоте - $O(kn)$, где n - количество объектов группирования (документов). Недостатком метода является то, что каждый документ может попасть всего лишь в один кластер.

5.3. Иерархическое группирование-объединение

Иерархическое группирование-объединение (Hierarchical Agglomerative Clustering, HAC) начинается с того, что каждому объекту в соответствие ставится отдельный кластер, а затем происходит объединение кластеров, которые наиболее близки друг к другу, в соответствии с выбранным критерием. Алгоритм завершается, когда все объекты объединяются в единый кластер. История объединений образует бинарное дерево иерархии кластеров.

Разновидности алгоритма HAC различаются выбором критериев близости между кластерами. Например, близость между двумя кластерами может вычисляться как максимальная близость между объектами из этих кластеров.

Могут использоваться и другие меры близости, например, близость «центров масс», средняя близость между всеми парами объектов в объединенных кластерах и т.п. Мера близости между двумя кластерами C_i и C_j в последнем случае вычисляется по формуле:

$$Sim(C_i,\ C_j) = \frac{1}{\left|C_i \cup C_j\right|\left(\left|C_i \cup C_j\right| - 1\right)} \sum_{x,y \in C_i \cup C_j, x \neq y} Sim(x,\ y).$$

В этом выражении $|C_i \cup C_j|$ - количество объектов в множестве $C_i \cup C_j$, а x и y - объекты, принадлежащие $C_i \cup C_j$.

Сложность алгоритма HAC составляет $O(n^2 s)$, где n - количество объектов, а s - сложность вычисления близости между кластерами.

5.4. Метод суффиксных деревьев

Изначально метод суффиксных деревьев (Suffix Tree Clustering) был разработан для быстрого поиска подстрок в строках.

Суффикс W строки S - это такая строка, в которой конкатенация (сцепление строк) VW совпадает с S для некоторой (возможно, пустой) строки V. Суффикс называется собственным, если $|V| \neq 0$. Например, для строки «substring»

подстрока «sub» является собственным префиксом, «ring» — собственным суффиксом. Срока V называется при этом префиксом.

Суффиксное дерево – это дерево, содержащее все суффиксы строки. Оно состоит из вершин, ветвей и суффиксных указателей (меток). Метка узла в дереве определяется как конкатенация подстрок, маркирующих ребра пути от корня дерева до этого узла.

Существуют алгоритмы (например, алгоритм Укконена (E. Ukkonen) [143]), реализующие построение суффиксных деревьев за $O(n)$ шагов, где n - длина строки. Ветви дерева обозначаются отдельными буквами или частями суффиксов строки. Суффикс, соответствующий определенной вершине, можно получить путем объединения букв, которые находятся на ветвях, начиная от корневой вершины и заканчивая данной.

Рассмотрим простейший вариант построения суффиксного дерева, формируемого «налету», по мере поступления новых символов в «хвост» строки. Пусть строка S в окончательном виде состоит из последовательности символов $t_1,..., t_n$. Вводится понятие префикса, т.е. последовательности символов $t_1, ..., t_i$ – начало строки S. Суффиксное дерево строится не только для всей строки, но и для всех ее префиксов. Рассмотрим алгоритм детальнее по шагам:

0. Строится суффиксное дерево для t_1.

1. Суффиксное дерево расширяется путем добавления ветвей t_1t_2.

. . .

n - 1. Суффиксное дерево для $t_1, ..., t_{n-1}$ расширяется до дерева для $t_1,..., t_n$.

n. Суффиксное дерево для $t_1, ..., t_n$ расширяется до дерева для $t_1,..., t_n$\$ (\$ - конец текста).

Последовательность шагов приведенного алгоритма для строки abca\$ приведена на рис. 20.

В настоящее время идеология суффиксных деревьев применяется для кластеризации результатов работы информационно-поисковых систем в интерактивном режиме. Именно такой подход используется, например, на поисковых серверах Clusty (http://www.clusty.com) или Nigma

(http://www.nigma.ru). Остановимся подробнее на принципах построения и применения суффиксных деревьев в случае, когда в качестве текстовой строки используется весь текст документа, а в качестве символов – слова.

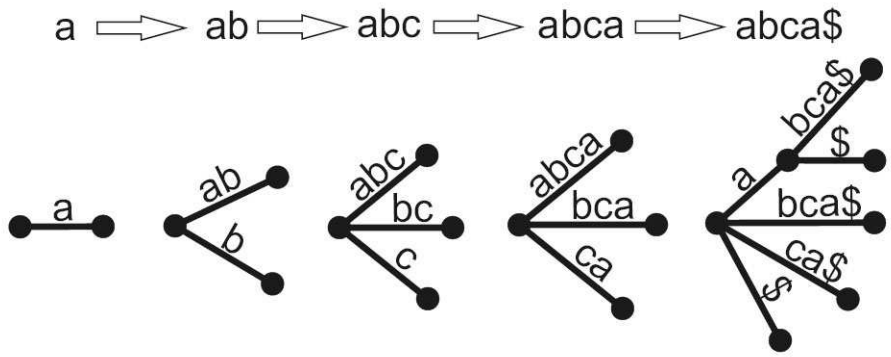


Рис. 20. Процесс построения суффиксного дерева

Вначале при построении дерева документы, получаемые от поисковой системы, подвергаются отчистке от пунктуации, слова из документов приводятся в канонические формы (лемматизация) и т.д. После этого для найденных документов строится суффиксное дерево, но в этом случае ветвям приписываются термы (слова или словосочетания), а не буквы как при традиционном подходе. В результате вершинам дерева соответствуют фразы, которые можно получить, объединив все термы, находящиеся на ветвях, ведущих от корня к данной вершине дерева. В вершинах дерева, имеющих потомков, расположены ссылки на документы, в которых встречается фраза, соответствующая вершине. Ее можно получить, объединив все слова, находящиеся на ребрах на пути от корня дерева к данной вершине. Множества документов, на которые указывают эти ссылки, образуют базовые кластеры. Затем происходит укрупнение базовых кластеров и получение окончательного набора кластеров. На рис. 21 приведено суффиксное дерево для предложения «I know you know I know». 6 узлов в этом примере маркированы как прямоугольники с цифрами от 1 до 6.

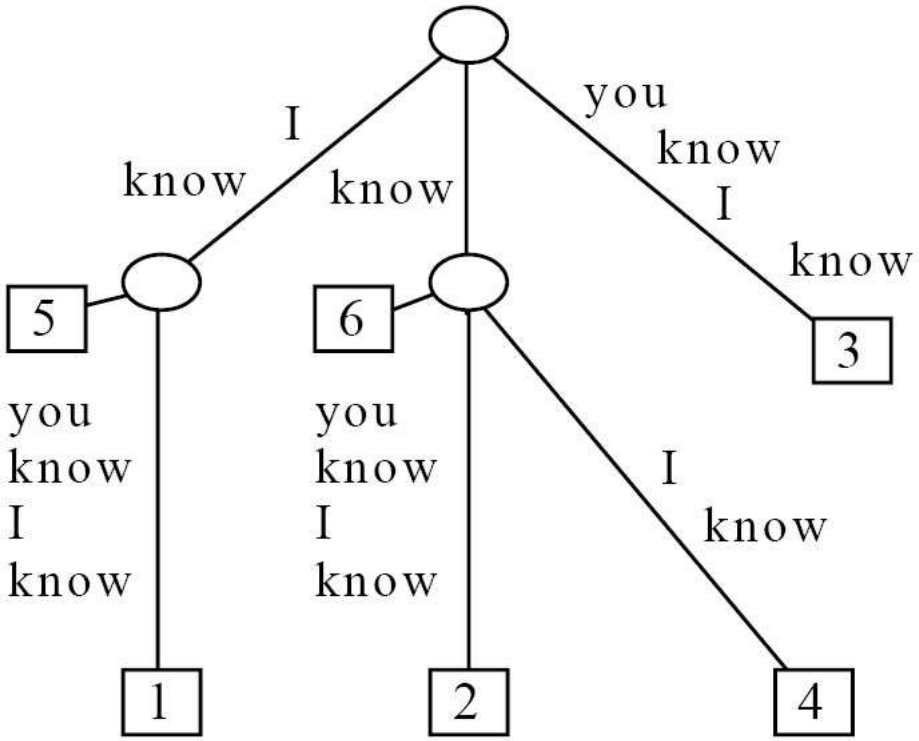


Рис. 21 . Пример суффиксного дерева для одного предложения

Суффиксное дерево можно генерировать также из нескольких предложений. На рис. 22 приведен пример общего суффиксного дерева для трех предложений «cat ate cheese», «mouse ate cheese too» и «cat ate mouse too». Внутренние узлы на данном графе представлены как кружки, которые отмечены буквами от a до f. Одиннадцать листьев в этом примере маркированы прямоугольниками, первая цифра в которых – номер предложения, из которого взят суффикс, а вторая цифра – позиция в предложении.

Кластеры определяются на основе близости между множествами документов, соответствующих узлам суффиксного дерева, следующим образом. Пусть B_m и B_n – базовые кластеры, $|B_m|$, $|B_n|$ - их размеры. $|B_m \cap B_n|$ – количество общих документов для этих кластеров. Близость $sim(B_m, B_n)$ между B_m и B_n задается условием:

$$sim(B_m, B_n) = \begin{cases} 1, & \text{если } \dfrac{|B_m \cap B_n|}{|B_m|} > \alpha, \ \dfrac{|B_m \cap B_n|}{|B_n|} > \alpha; \\ 0, & \text{иначе,} \end{cases}$$

где α – некоторый порог, принимающий значение от 0 до 1, например 0,6.

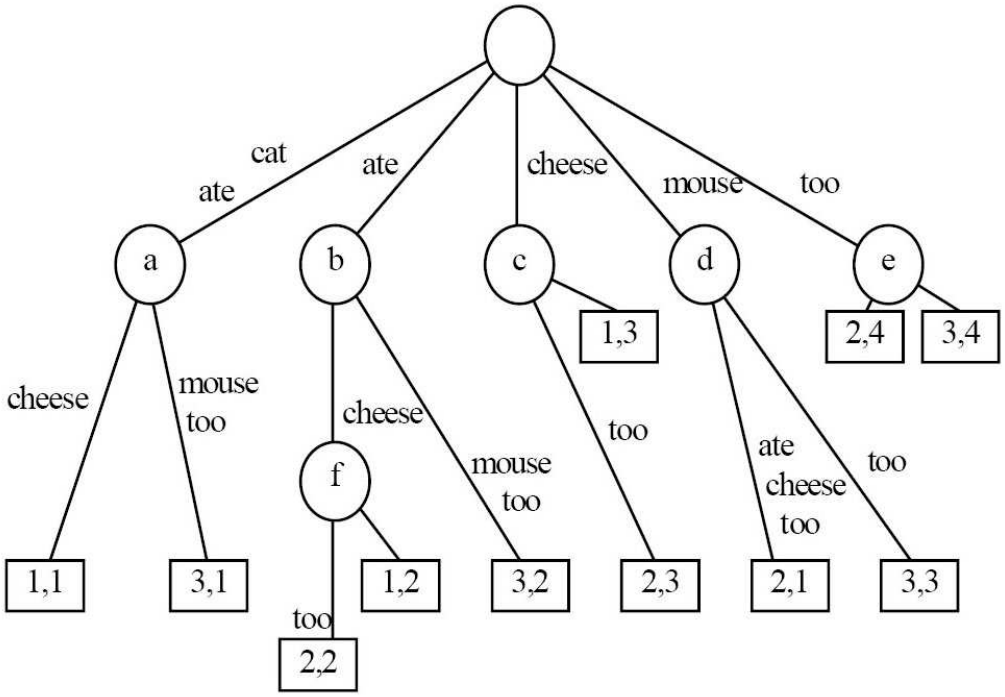


Рис. 22. Пример общего суффиксного дерева для нескольких предложений

На рис. 23. показаны результаты формирования кластеров для случая, приведенного на рис. 22 с 6-ю узлами (от *a* до *f*). В случае *a* порог $\alpha = 0.7$, в случае *b* $\alpha = 0.6$, в случае *c* порог также равен 0.6, однако слово ate исключено с помощью «стоп-словаря».

В отличие от большинства моделей кластеризации, рассматривающих текст как “Bag of Words”, в методе суффиксных деревьев учитывается порядок слов, что существенно влияет на учет лингвистических особенностей группируемых текстов. В отличие от многих других методов, кластеры, сформированные с помощью метода суффиксных деревьев могут пересекаться, что вполне соответствует реальности. К достоинствам этого метода можно отнести также его наглядность представления результатов и высокую скорость работы.

5.5. Гибридные методы

На практике достаточно часто применяются так называемые гибридные методы, которые зачастую объединяют несколько теоретических подходов. В качестве реализации гибридного метода рассмотрим алгоритм выявления основных сюжетов из потока новостей, применяемый в системе InfoStream [31].

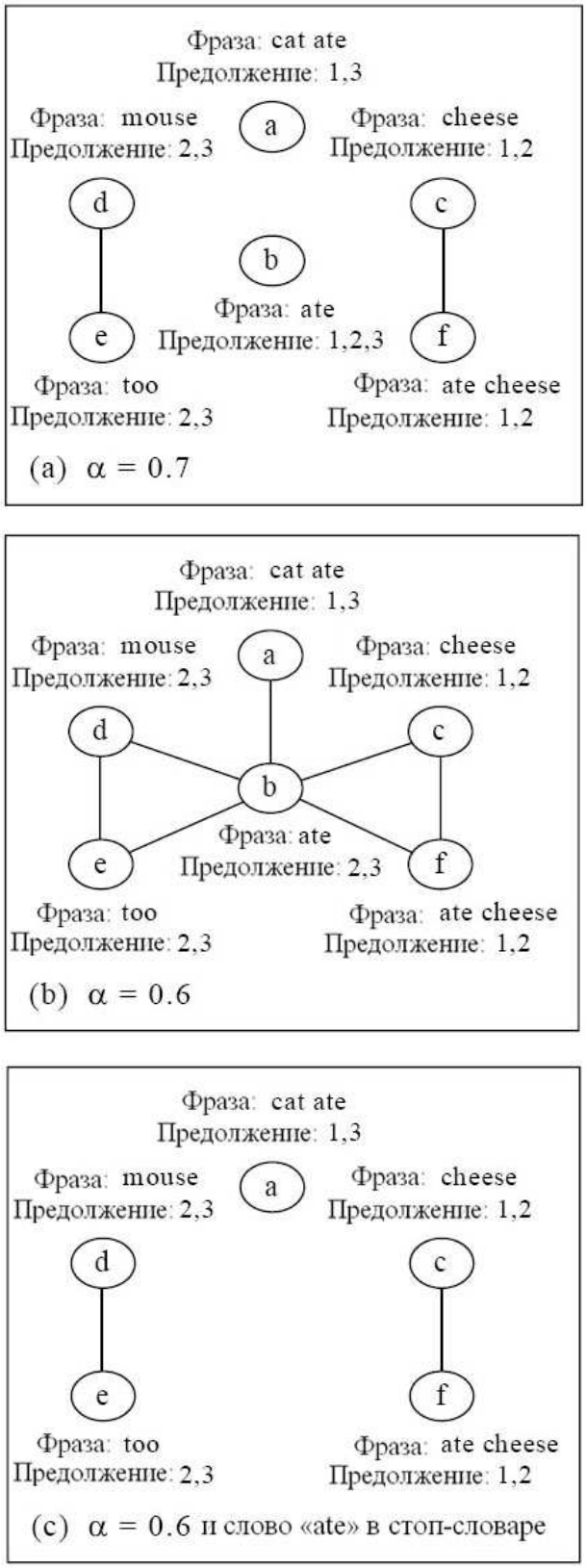


Рис. 23. Графы кластеров суффиксного дерева

В соответствии с этим алгоритмом последний документ, который поступает на вход системы (документ с номером 1 при обратной нумерации), порождает

первый кластер и сравнивается со всеми предыдущими в соответствии с некоторой метрикой. Если эта мера близости для какого-нибудь документа оказывается больше заданной, то текущий документ приписывается первому кластеру. Сравнение длится до тех пор, пока не исчерпывается список актуальных документов потока. После такой обработки документа с номером 1, происходит обработка следующего документа, который не вошел в первый кластер, с которым последовательно сравниваются все актуальные документы потока и т.д. В результате формируется некоторое неизвестное заранее количество кластеров, которые ранжируются по своим весам. Для выбранных кластеров, как и в методе *k-means* заново пересчитываются центроиды - документы, которые наиболее близки в смысле векторно-пространственной модели и предположительно лучше всего отражают тематику кластеров.

Укрупнение рубрик - актуальная задача кластерного анализа и она может быть решена путем их группирования по признакам подобия. Рассмотрим множество T всех термов t_i $(i = 1, \ ..., \ N)$ в некоторой системе тематических информационных портретов (профайлов) P_j $(j = 1, \ ..., \ M)$ и его проекцию на это множество - матрицу P, строки которой соответствуют профайлам, а столбцы – термам.

Произведение матриц $E = P^T \cdot P$ будет таблицей взаимосвязей тематик, построенной в результате анализа состава термов из соответствующих профайлов.

В некоторых случаях можно эффективно выделить некоторое число групп взаимозависимых тематических рубрик, используя, например, методы кластерного анализа *k*-means или LSI, заменяя их в последующем одной, укрупненной рубрикой.

При заранее определенных тематических профайлах P для любого документа может быть вычислен его вес в пространстве этих профайлов (задача линейной классификации). На практике тематические профайлы чаще всего формируются путем лингвостатистического анализа массивов документов, полученных в результате поиска по тематическим запросам. Эти запросы в

большинстве промышленных информационно-поисковых систем составляются на языках, которые являются расширением булевой алгебры.

Окончательная же рубрикация документов допускает более «экономный» весовой подход на основе массивов термов, входящих в соответствующие профайлы. Таким образом, в результате учитываются «логические» преимущества первого подхода и эксплуатационные - второго.

Определим матрицу M отображения потока документов на пространство тематических профайлов, строки которой соответствуют документам, а столбцы - тематикам. Введем понятие ядра этой операции как произведения матриц $A = M^T \cdot M$. Матрица A по смыслу представляет собой матрицу взаимосвязей тематических профайлов.

Еще одна матрица, полученная в результате умножения $B = M \cdot M^T$, выражает взаимосвязь документов. Для современных информационных потоков размерность матриц B намного превышает размерность матриц A. Соответственно, обнаруживая явные группы взаимозависимых тем в матрице A, можно определять группы взаимозависимых документов в матрице B, группирование элементов которой ввиду ее размерности и динамики роста – достаточно сложная задача.

5.6. Ранжирование результатов поиска

Ранжирование - процесс, при котором поисковая система выстраивает результаты поиска в определенном порядке по принципу наибольшего соответствия конкретному запросу. Представление результатов поиска конечно зависит от алгоритма ранжирования. Ранжирование результатов поиска по уровню релевантности возможно не для всех моделей поиска (например, невозможно для булевой модели).

Перспективный подход к ранжированию - использование многопрофильных шкал, сформированных на основе метаданных, сетевых свойств, данных о пользователях.

Например, реализация сюжетных цепочек в тематических информационных массивах и их взвешивание рассматриваются как один из алгоритмов ранжирования. Ранжирование текстовых и гипертекстовых документов существенно различается. Текстовые документы могут ранжироваться по уровням релевантности и другим параметрам, экстрагируемым из текстов. Ранжирование гипертекстовых документов возможно также по свойствам, обуславливаемым сетевой структурой, гиперссылками. Например, для определения авторитетности веб-страницы как источника информации или посредника используется анализ графа, образованнго веб-документами и соответствующими гиперссылками. Два самых известных алгоритма ранжирования веб-страниц, основанных на связях, HITS (hyperlink induced topic search) и PageRank, были разработаны в 1996 году в IBM Дж. Клейнбергом (J. M. Kleinberg) [105] и в Стенфордском Университете С. Брином (S. Brin) и Л. Пейджем (L. Page) [80].

Оба алгоритма предназначены для решения "проблемы избыточности", свойственной широким запросам, увеличения точности результатов поиска на основе методов анализа сложных сетей.

5.6.1. Алгоритм HITS

Алгоритм HITS (Hyperlink Induced Topic Search), предложенный Дж. Клейнбергом, является реализацией латентно-семантического индексирования (см. п. 5.1) к ранжированию выдачи информационно-поисковых систем.

Алгоритм HITS обеспечивает выбор из информационного массива лучших «авторов» (первоисточников, на которые введут ссылки) и «посредников» (документов, от которых идут ссылки цитирования). Понятно, что страница является хорошим посредником, если она содержит ссылки на ценные первоисточники, и наоборот, страница является хорошим автором, если она упоминается хорошими посредниками.

Дж. Клейнберг

Для каждого документа $d_j \in D$ рекурсивно вычисляется его значимость как автора $a(d_j)$ и посредника $h(d_j)$ по формулам:

$$a(d_j) = \sum_{i=1,\ i \neq j}^{|D|} h(d_i),\ \ h(d_j) = \sum_{i=1,\ i \neq j}^{|D|} a(d_i).$$

Покажем, что алгоритм HITS подобен LSA. Введем понятие матрицы инциденций A, элемент которой a_{ij} равен единице, когда документ d_i содержит ссылку на документ d_j, и нулю в противном случае. Воспользуемся сингулярным разложением: $A = USV^T$, где S - квадратная диагональная матрица с неотрицательными диагональными элементами $s_{i,i}$. Рассмотрим матрицу $A^T A$, для которой справедливо: $A^T A = VSU^T USV^T = VS^2V^T$, где S^2 – диагональная матрица с элементами $s_{i,i}^2$. Соответственно, для матрицы AA^T будет справедливо $AA^T = US^2U^T$. Очевидно, что как и при LSA, собственные векторы, которые соответствуют наибольшим сингулярным значениям AA^T (или $A^T A$), будут соответствовать статистически наиболее важным авторам (или посредникам).

Алгоритм вычисления рангов HITS приводит к росту рангов документов при увеличении количества и степени связанности документов соответствующего сообщества. В этом случае в результаты поиска системы, использующей алгоритм HITS, могут попасть в большом количестве документы по темам, отличным от информационной потребности пользователя, но тесно связанных между собой, т.е. часть выдаваемых результатов может отклониться от доминирующей

тематики. В этом случае происходит, так называемый, сдвиг тематики (topic drift) за счет наличия «тесно связанных сообществ» документов (Tightly-Knit Community, TKC).

Для решения этой проблемы как некоторое расширение стандартного алгоритма HITS был предложен алгоритм PHITS. В рамках этого алгоритма предполагается: D – множество цитирующих документов, C – множество ссылок, Z – множество классов (факторов). Предполагается также, что событие $d \in D$ происходит с вероятностью $P(d)$.

Условные вероятности $P(c|z)$ и $P(z|d)$ используются для описания зависимостей между наличием ссылки $c \in C$, латентным фактором $z \in Z$ и документом - $d \in D$.

Оценивается функция правдоподобия:

$$L(D,C) = \prod_{c \in C, d \in D} P(d,c) = \prod_{c \in C, d \in D} P(d)P(c|d),$$

где

$$P(c|d) = \sum_{z \in Z} P(c|z)P(z|d).$$

Цель алгоритма PHITS состоит в том, чтобы подобрать $P(z)$, $P(c|z)$, $P(d|z)$, чтобы максимизировать $L(D,C)$.

После этого:

$P(c|z)$ – ранги авторов;

$P(d|z)$ – ранги посредников.

Для вычисления рангов необходимо задать количество факторов в множестве Z, и тогда $P(c|z)$ будет характеризовать качество страницы как автора в контексте тематики z. К недостаткам метода надо отнести то, что итеративный процесс чаще всего останавливается не на абсолютном, а на локальном максимуме функции правдоподобия L. Вместе с там в ситуациях, когда в множестве найденных веб-страниц нет явного доминирования тематики запроса, PHITS превосходит алгоритм HITS.

5.6.2. Алгоритм PageRank

Алгоритм PageRank близок по идеологии к литературному индексу цитирования и рассчитывается для произвольного документа с учетом количества ссылок из других документов на данный документ. При этом PageRank как и HITS, в отличие от литературного индекса цитирования, не считает все ссылки равнозначными.

Принцип рассчета ранга веб-страниц PageRank основывается на модели «случайного блуждания» пользователя по следующему алгоритму: он открывает случайную веб-страницу, с которой переходит по случайно выбранной гиперссылке. Затем он перемещается на другую веб-страницу и снова активизирует случайную гиперссылку и т.д., постоянно переходя от страницы к странице, никогда не возвращаясь. Иногда ему такое блуждание надоедает, и он снова переходит на случайную веб-страницу - не по ссылке, а набрав вручную некоторый URL. В этом случае вероятность того, что блуждающий в WWW пользователь перейдет на некоторую определенную веб-страницу - это ее ранг. Очевидно, PageRank веб-страницы тем выше, чем больше других страниц ссылается на нее, и чем эти страницы популярнее.

Лэрри Пейдж и Сергей Брин, авторы PageRank

Пусть есть n страниц $\{d_1, ..., d_n\}$, которые ссылаются на данный документ (веб-страницу A) , а $C(A)$ – общее число ссылок с веб-страницы A на другие документы. Определяется некоторое фиксированное значение δ как вероятность того, что пользователь, пересматривая какую-нибудь веб-страницу из множества D, перейдет на страницу A по ссылке, а не набирая ее URL в явном виде. В рамках модели вероятность продолжения этим пользователем веб-серфинга по сети из N веб-страниц без использования гиперссылок, путем ручного ввода адреса (URL) со случайной страницы составит $1-\delta$ (альтернатива перехода по гиперссылкам). Индекс PageRank $PR(A)$ для страницы A рассматривается как вероятность того, что пользователь окажется в некоторый случайный момент времени на этой странице:

$$PR(A) = (1-\delta)/N + \delta\sum_{i=1}^{n}\frac{PR(d_i)}{C(d_i)}.$$

По этой формуле индекс страницы легко подсчитывается простым итерационным алгоритмом. На практике применяется до 30 шагов итерации для достижения устойчивых результатов.

Несмотря на различия HITS и PageRank, в этих алгоритмах общее то, что авторитетность (вес) узла зависит от веса других узлов, а уровень "посредника" зависит от того, насколько авторитетны узлы, на которые он ссылается.

Расчет авторитетности отдельных документов сегодня широко используется в таких приложениях, как определение порядка сканирования документов в сети роботом ИПС, ранжирование результатов поиска, формирование тематических обзоров и т.п.

В настоящее время приобрели широкое распространение технологии искусственного повышения рангов отдельных веб-документов или их групп (веб-сайтов) путем установления гиперссылок, не имеющих отношения к их содержанию. Эти технологии, называемые методами поисковой оптимизации (SEO, Search Engine Optimization), основываются на приспособлении к существующим алгоритмам ранжирования веб-документов наиболее популярными поисковыми системами.

В свою очередь, такие технологии приводят к необходимости постоянного совершенствования алгоритмов ранжирования в поисковых системах, ориентации на содержательную составляющую веб-документов при определении их рангов.

5.6.3. Алгоритм Salsa

Алгоритм ранжирования Salsa (Stochastic Approach for Link-Structure Analysis - Стохастический Алгоритм Анализа Структуры Связей) [108] был предложен Ш. Мораном (Sh. Moran) и Р. Лемпелем (R. Lempel) как некоторый симбиоз алгоритмов PageRank и HITS, позволяющий сократить последствия образования ТКС – «тесно связанных сообществ» документов.

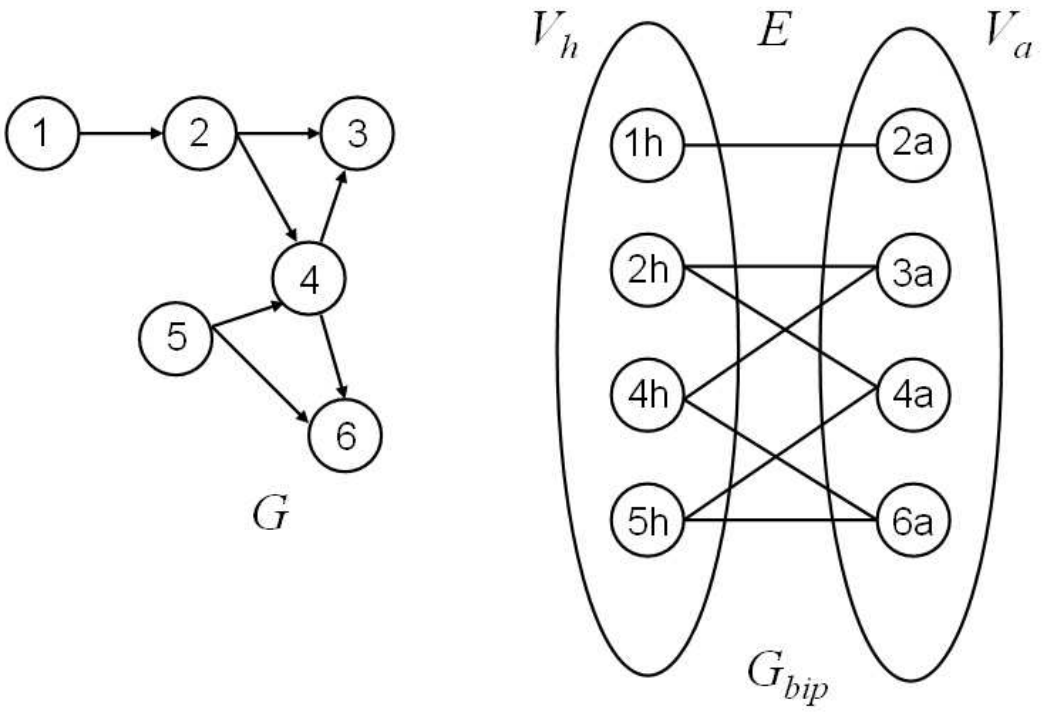
Как и в методе PageRank в случае Salsa предполагается модель случайного блуждания пользователя по веб-графу, однако предполагается наличие двухстороннего «серфинга». В соответствии с алгоритмом Salsa:

1. Из произвольного узла v, пользователь случайным образом возвращается к узлу u, который ссылается на v. Выбор v делается случайно, при условии, что узлы v и u принадлежат веб-графу.

2. Из u пользователь наугад переходит к узлу w, если существует связь $(u,\ w)$.

Веб-граф G (рис. 24 *а*) может быть преобразован в двудольный ненаправленный граф G_{bip}, (рис. 24 *б*) и определен как совокупность $G_{bip} = (V_h,\ V_a,\ E)$, где h обозначает посредников, V_h - совокупность узлов-посредников (тех, из которых исходят ссылки), a - авторов, V_a - совокупность узлов-авторов (тех, на которые ведут ссылки). Необходимо отметить, что одни и те же узлы могут быть одновременно и авторами и посредниками.

Каждая неизолированная страница $s \in G$ представлена в G_{bip} одним или двумя узлами s_h и s_a. В этом двудольном графе Salsa реализует два разных случайных перехода. При каждом переходе возможно «посещение» узлов только из одной из двух долей графа G_{bip}.



а) *б)*

Рис. 24: Salsa: конструкция двудольного графа

Каждый путь длины два в G_{bip} представляет собой обход одной гиперсвязи (при прохождении от доли посредников к доле авторов в G_{bip}), и отход вдоль гиперсвязи (при прохождении в обратном направлении). Это движение в обратном направлении напоминает танец сальса, который ассоциируется с названием данного алгоритма.

Так как посредники и авторы, относящиеся к теме t должны быть явно выражены в G_{bip} (доступны из многих узлов благодаря прямым ссылкам или коротким путям), предполагается что авторы из V_a и посредники из V_h, относящиеся к теме t, будут наиболее часто посещаемыми при случайных «блужданиях» пользователей.

В алгоритме Salsa исследуются две различных цепи Маркова, которые ассоциируются с этими случайными блужданиями: цепь на стороне авторов G_{bip} (цепь авторов), и цепь на стороне посредников G_{bip}.

Такой подход позволяет ввести две стохастические матрицы перехода цепей Маркова, которые определяются следующим образом: строится матрица инциденций W ориентированного графа G. Обозначим как W_r матрицу, полученную делением каждого ненулевого элемента W на сумму значений соответствующей строки, а через W_c - матрицу, полученную делением каждого ненулевого элемента W на сумму элементов в соответствующем столбце. Тогда, матрица H, соответствующая посредникам будет состоять из ненулевых строк и столбцов $W_r W_c^T$, а матрица авторов A, соответственно, будет состоять из ненулевых строк и столбцов $W_c^T W_r$. В рамках алгоритма Salsa игнорируются строки и столбцы матриц A и H, которые состоят полностью из нулей, так как по определению, все узлы G_{bip} имеют не менее одной связи. В результате матрицы A и H используются для вычисления рангов тем же путем, что и в алгоритме HITS.

В [108] показано что, сходящиеся в процессе итерационного процесса вероятность перехода к узлу v как к автору, имеет очень простую форму:

$$\pi_v = c_1 \cdot InDegree(v),$$

а вероятность возврата к узлу u как к посреднику:

$$\pi_u = c_2 \cdot OutDegree(u),$$

где c_1 и c_2 - некоторые константы, а $InDegree$ и $OutDegree$ - это количество исходящих и входящих ссылок, соответственно.

Р. Лемпель и С. Моран продемонстрировали, что алгоритм Salsa менее чувствителен к эффекту тесно связанных сообществ, чем HITS, но при условиях, что вручную в документах удаляются ссылки, не относящиеся к исследуемой теме. Это требование на практике ведет к большим издержкам, в результате чего авторам пока не известно случаев использования этого алгоритма ранжирования в реально работающих системах.

5.6.4. Ранжирование «по Хиршу»

В 2005 г. Й. Хиршем (J. Hirsch) был предложен новый метод оценивания индекса цитирования (h-индекс), который претендует на большую объективность в сравнении с традиционным индексом цитирования [100].

Метод состоит в подсчете количества h публикаций одного автора, на которые есть не менее h ссылок. Автор p имеет индекс Хирша h, если h из его N_p статей цитируются как минимум h раз каждая, в то время как $N_p - h$ статей из тех, что остались, цитируются меньше чем по h раз (рис. 25).

Параметр Хирша для сайта-источника может определяться, например, равным максимальному количеству дней (h), на протяжении которых было зафиксировано не менее h внешних ссылок на данный сайт (таким образом параметр на практике может подсчитываться для новостных сайтов [49]).

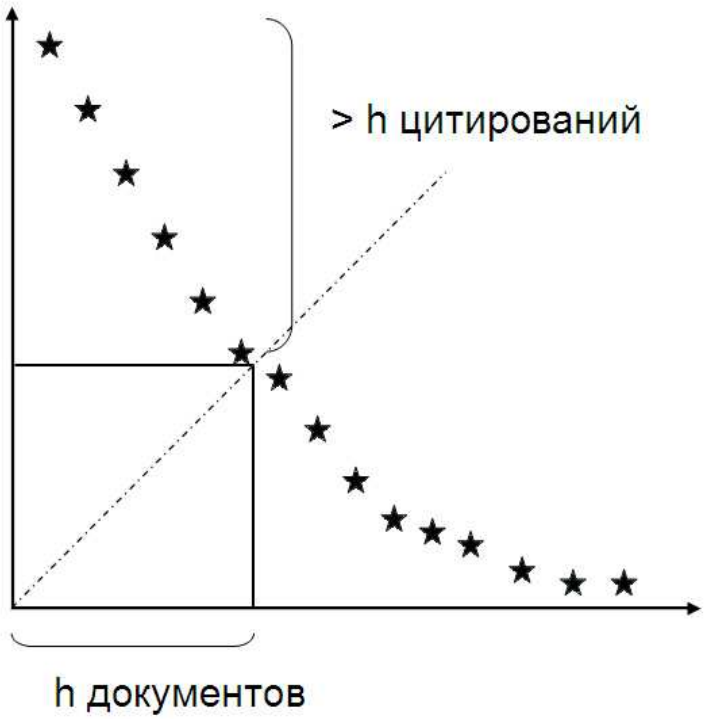


Рис. 25. Графическое представление алгоритма h-индекса:ось 0X – номер документа в ранжированном списке; ось 0Y – количество цитирований

Рейтинг Хирша характеризует как регулярность ссылок на источник, так и количество этих ссылок. Этот показатель учитывает стабильность авторитетности источника на протяжении продолжительного периода времени.

6. ЭМПИРИЧЕСКИЕ РАСПРЕДЕЛЕНИЯ И МАТЕМАТИЧЕСКИЙ ФОРМАЛИЗМ

« Я хотел сказать, что въедливость
и точность в математическом смысле –
все что у нас есть.»
Нил Стивенсон

Так как большинство моделей информационного поиска учитывают статистическую природу текстов, остановимся на вопросах статистического распределения текстовых данных. В естественных науках хорошо известны и широко распространены такие статистические распределения, как Гауссово, показательное, биноминальное. Однако, в практике информационного поиска самое большое распространение имеет степенное распределение.

6.1. Эмпирические закономерности

Ниже будут обсуждаться параметры некоторых распределений, присущих многим информационным процессам, с учетом которых можно строить модели одновременно в рамках теории информационного поиска и концепции сложных сетей.

6.1.1. Распределение Парето

Анализируя общественные процессы, В. Парето (V. Pareto) рассмотрел социальную среду как пирамиду, на вершине которой находятся люди, представляющие элиту. Парето в 1906 году установил, что около 80 процентов земли в Италии принадлежит лишь 20 процентам ее жителей. Он пришел к заключению, что параметры полученного им распределения приблизительно одинаковы и принципиально не различаются в разных странах и в разное время. Парето также установил, что точно такая же закономерность наблюдается и в распределении доходов между людьми, которое описывается уравнением $N = A/X^{p}$, где X – величина дохода, N - количество людей с доходом, равным

или превышающим X, A и p - параметры распределения. В математической статистике это распределение получило имя Парето, при этом предполагаются естественные ограничения на параметры: $X \geq 1,\ p > 1$. Распределению Парето присуще свойство устойчивости, т.е. сумма двух случайных переменных, которые имеют распределение Парето, также будет распределена по Парето. Замеченное распределение, называемое "законом Парето" или "принципом 80/20", применимо в очень многих областях. Например, при информационном поиске достаточно определить 20% важнейших ключевых слов, чтобы найти 80% необходимых документов, а затем расширить поиск или воспользоваться опцией "найти похожие" для полного решения задачи. Еще один пример: 80% посещений веб-сайта приходится лишь на 20% его веб-страниц.

Вильфредо Парето (1848 - 1923)

При построении систем массового обслуживания, в том числе и информационно-поисковых систем, необходимо учитывать тот факт, что наиболее сложным функциональным возможностям системы, на реализацию которых уходит 80 и больше процентов трудозатрат, будут пользоваться не более чем 20 процентов пользователей данной системы.

Перейдем к более строгой формулировке закона Парето. Предположим, что последовательность $x_1, x_2, ...x_n, ...$ соответствует размерам доходов отдельных людей. После ранжирования этой последовательности по убыванию получается

новая последовательность $x_{(1)}$, $x_{(2)}$, ..., $x_{(r)}$, ... (элементы $x_{(r)}$ расположены в порядке убывания).

Предположим, что N - общее число людей, у которых доход составляет не менее $x_{(r)}$, т.е. $N = r$. Тогда правило Парето можно переписать в таком виде:

$$r = \frac{A}{x_{(r)}^{\,p}}.$$

Отсюда:

$$x_{(r)} = \left(\frac{A}{r}\right)^{\frac{1}{p}}$$

Рассматривается сумма первых n ($n = 1,\ 2,\ ...,\ N$) значений величины $x_{(r)}$, то есть общая величина дохода наиболее богатых людей - $m(n)$ составляет:

$$m(n) = \sum_{r=1}^{n} x_{(r)} = \sum_{r=1}^{n} \left(\frac{A}{r}\right)^{1/p} = \sum_{r=1}^{n} \frac{C}{r^{\gamma}},$$

где $\gamma = 1 - 1/p;\ \ C = A^{1/p}$.

Переходя от дискретных величин к непрерывным (предполагая, что $n >> 1$), имеем:

$$m(n) \approx \int_{1}^{n} \frac{C}{r^{\gamma}} dr \approx \frac{C}{1-\gamma} n^{1-\gamma}.$$

В безразмерных переменных $\mu = m(n)/m(N)$ - и $\nu = n/N$ последнее равенство имеет вид (см. рис. 26):

$$\mu = \nu^{1-\gamma}.$$

Величина μ - в нашем примере - относительное количество дохода, получаемого первыми по рангу n людьми, доля которых (относительно всех людей) равна ν.

Для последних двух случаев, представленных на рис. 26, $\nu \approx 0.2$ - 20% людей имеют $\mu \approx 0.8$ - 80% доходов (близкие к этим значениям явления наблюдаются в реальной жизни).

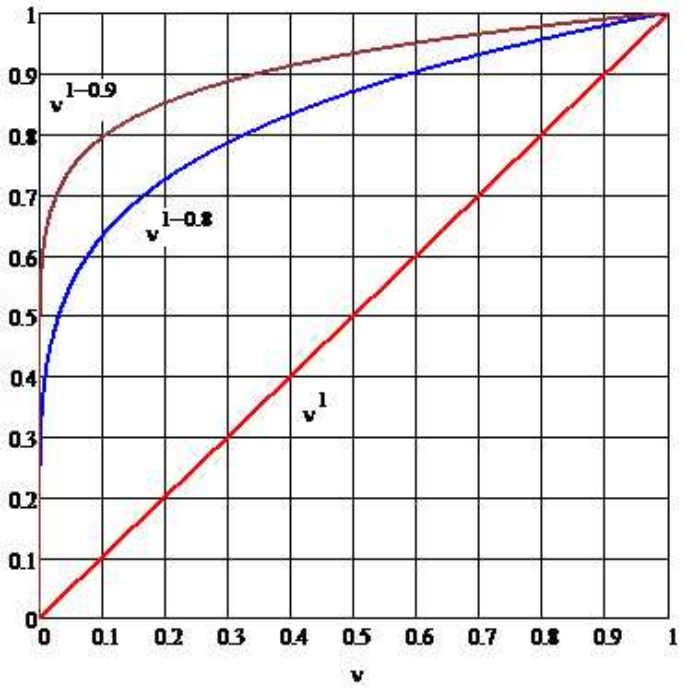


Рис. 26. Распределение Парето для различных значений параметров: зависимость $\mu = \nu^{1-\gamma}$ *для трех случаев -* $\gamma = 0$ *(доходы всех одинаковы) и* $\gamma = 0.8,\ \ \gamma = 0.9$

6.1.2. Законы Ципфа

Дж. Ципф (G. Zipf) изучал использование статистических свойств языка в текстовых документах и выявил несколько эмпирических законов, которые представил как эмпирическое доказательство своего «принципа наименьшего количества усилий». Он экспериментально показал, что распределение слов естественного языка подчиняется закону, который часто называют первым законом Ципфа, относящимся к распределению частоты слов в тексте. Этот закон можно сформулировать таким образом. Если для какого-нибудь довольно большого текста составить список всех слов, которые встретились в нем, а потом ранжировать эти слова в порядке убывания частоты их появления в тексте, то для любого слова произведение его ранга и частоты появления будет величиной постоянной: $f \cdot r = c$, где f - частота встречаемости слова в тексте; r - ранг слова в списке; c - эмпирическая постоянная величина (коэффициент Ципфа). Для

славянских языков, в частности, коэффициент Ципфа составляет приблизительно 0,06-0,07.

Джордж Ципф (1902 -1950)

Приведенная зависимость отражает тот факт, что существует небольшой словарь, который составляет большую часть слов текста. Это главным образом служебные слова. Например, приведенный в [111] анализ романа «Том Сойер», позволил выделить 11.000 английских слов. При этом было обнаружено двенадцать слов (the, and, и др.), каждое из которых охватывает более 1 % лексем в романе. Закон Ципфа был многократно проверен на многих массивах. Ципф объяснял приведенное выше гиперболическое распределение «принципом наименьшего количества усилий» предполагая что при создании текста меньше усилий уходит на повторение некоторых слов, чем на использование новых, т.е. на обращение к «оперативной памяти, а не к долговременной».

Ципф сформулировал еще одну закономерность, так называемый второй закон Ципфа, состоящий в том, что частота и количество слов, которые входят в текст с данной частотой, также связанны подобным соотношением, а именно:

$$N(f) = \frac{B}{f^{\beta}},$$

где $N(f)$ - количество различных слов, каждое из которых используется в тексте f раз, B - константа нормирования.

Существует простая количественная модель определения зависимости частоты от ранга. Предположим, что генерируется случайный текст обезьяной на пишущей машинке. С вероятностью p генерируется пробел, а с вероятностью $(1-p)$ - другие символы, каждый из которых имеет равную вероятность. Показано, что полученный таким образом текст будет давать результаты, близкие по форме к распределению Ципфа. Эта модель была усовершенствована в соответствии с фактическими эмпирическими данными, когда вероятности генерации отдельных символов были заданы на основе анализа большого текстового массива [73]. Полученное соответствие не доказывает закона Ципфа, но вполне его объясняет с помощью простой модели.

Более сложную модель генерации случайного текста, удовлетворяющего второму закону Ципфа, предложил Г.А. Саймон (H.A. Simon) [135]. Условия этой модели достаточно просты: если текст достиг размера в n слов, тогда то, каким будет $(n+1)$-е слово текста определяется двумя допущениями:

1. Пусть $N(f,n)$ - количество разных слов, каждое из которых использовалось f раз среди первых n слов текста. Тогда вероятность того, что $(n+1)$-ым окажется слово, которое до того использовалось f раз пропорционально $f \cdot N(f,n)$ - общему количеству появления всех слов, каждое из которых до этого использовалось f раз.
2. С вероятностью δ $(n+1)$-ым словом будет новое слово.

Распределение Ципфа часто искажается на практике ввиду недостаточных объемов текстовых корпусов, что приводит к проблеме оценки параметров статистических моделей. Вместе с тем соотношение между рангом и частотой была взята Солтоном в 1975 г. [131] как отправная точка для выбора терминов для индексирования. Далее им рассматривалась идея сортировки слов в соответствии с их частотой в текстовом массиве. Как второй шаг высокочастотные слова могут быть устранены, потому что они не являются хорошими различительными признаками для отдельных документов из текстового массива. На третьем шаге термы с низкой частотой, определяемой некоторым порогом (например слова, которые встречаются только единожды или

дважды) удаляются, потому что они встречаются так нечасто, что редко используются в запросах пользователей. Используя этот подход, можно значительно уменьшить размер индекса поисковой системы. Более принципиальный подход к подбору индексных термов – учет их весовых значений. В весовых моделях среднечастотные термы оказываются самыми весомыми, так как они являются наиболее существенными при отборе того или иного документа (наиболее частотные слова встречаются одновременно в большом количестве документов, а низкочастотные могут не входить в документы, интересующие пользователя).

Еще один эмпирический закон, сформулированный Ципфом состоит в том, что количество значений слова коррелирует с квадратным корнем его частоты. Подразумевалось, что нечасто используемые слова более однозначны, а это подтверждает то, что высокочастотные слова не подходят для внесения в индексы информационно-поисковых систем.

Ципф также определил, что длина слова обратно пропорциональна его частоте, что может быть легко проверено путем простого анализа списка служебных слов. Последний закон действительно служит примером принципа экономии усилий: более короткие слова требуют меньше усилий при воспроизведении, и таким образом, используются более часто. Этот «закон» можно подтвердить, рассматривая приведенную выше модель генерации слов обезьяной. Легко видеть, что вероятность генерации слова уменьшается с длиной, вероятность слова из n непробельных символов равна:

$$(1-p)^n \cdot p,$$

где p - вероятность генерации пробела.

Хотя закон Ципфа дает интересные общие характеристики слов в текстовых массивах, в общем случае замечены некоторые ограничения его применимости при получении статистических характеристик документальных массивов, состоящих из множества независимых документов разных авторов.

Законам Ципфа удовлетворяют не только слова из одного текста, но многие объекты современного информационного пространства.

6.1.3. Закономерность Бредфорда

Закономерность С. Бредфорда (S. Bredford), известного документалиста, одного из авторов универсальной десятичной классификации – УДК, состоит в следующем: если научные журналы расположить в порядке убывания числа помещенных в них статей по конкретному предмету, то полученный список можно разбить на три зоны таким образом, чтобы количество статей в каждой зоне по заданному предмету была одинаковой. Эти три зоны представляют: ядро - профильные журналы, непосредственно посвященные рассмотренной тематике, журналы, частично посвященные заданной области и журналы, тематика которых довольно далека от рассмотренного предмета. С. Бредфорд в 1934 г. установил следующее соотношение для количества журналов в разных зонах [79]:

$$\frac{N_3}{N_2} = \frac{N_2}{N_1} = const,$$

где количество журналов в первой зоне - N_1, во второй - N_2, в третьей - N_3.

Бредфорд вначале рассматривал найденную закономерность только как специфический случай распределения Ципфа для системы периодических изданий по науке и технике. Однако в дальнейшем оказалось, что эта же закономерность справедлива и для периодических изданий из многих других предметных областей [2, 31], а также для наборов веб-сайтов, относящихся к некоторой выбранной тематике.

6.1.4. Закон Хипса

В компьютерной лингвистике эмпирический закон Г.С. Хипса (H.S. Heaps) связывает объем документа с объемом словаря уникальных слов, которые входят в этот документ [98]. Казалось бы, словарь уникальных слов должен насыщаться, а его объем стабилизироваться при увеличении объемов текста. Оказывается, это не так! Для всех известных сегодня текстов в соответствии с законом Хипса, эти значения связаны соотношением (рис. 27):

$$v(n) = \alpha n^{\beta},$$

где $v-$ это объем словаря уникальных слов, составленный из текста, который состоит из n уникальных слов, α и $\beta-$ определенные эмпирически параметры. Для европейских языков α принимает значение от 10 до 100, а β - от 0.4 до 0.6.

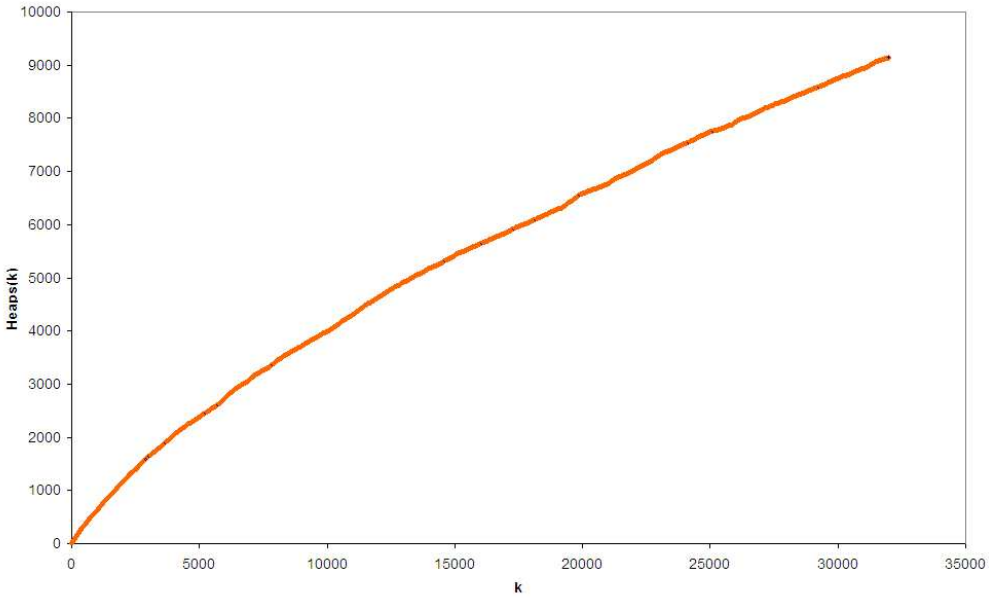


Рис. 27. Типичный график, подтверждающий закон Хипса: по оси абсцисс – количество слов в тексте, по оси ординат – объем словаря – количество уникальных слов

Закон Хипса справедлив не только для уникальных слов, но и для многих других информационных объектов, что вполне естественно, так как уже доказано [96] , что он является следствием закона Ципфа.

6.2. Степенные распределения случайных величин

Наиболее частыми (как обычно считается), универсальными законами распределения случайных величин, встречаемыми в различных естественнонаучных исследованиях, является нормальный закон – распределение Гаусса и так называемое логнормальное распределение (рис. 28):

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}},$$

$$f(x)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x)^2}{2\sigma^2}},\quad x>0$$

Частая встречаемость нормального закона объясняется тем, что когда случайная величина является суммой независимых случайных величин, то ее распределение приближается к нормальному. Именно это утверждение является содержанием так называемой центральной предельной теоремы теории вероятностей. Заметим, что часто в конкретных исследованиях гауссово распределение случайной величины принимается в силу привычки или удобства.

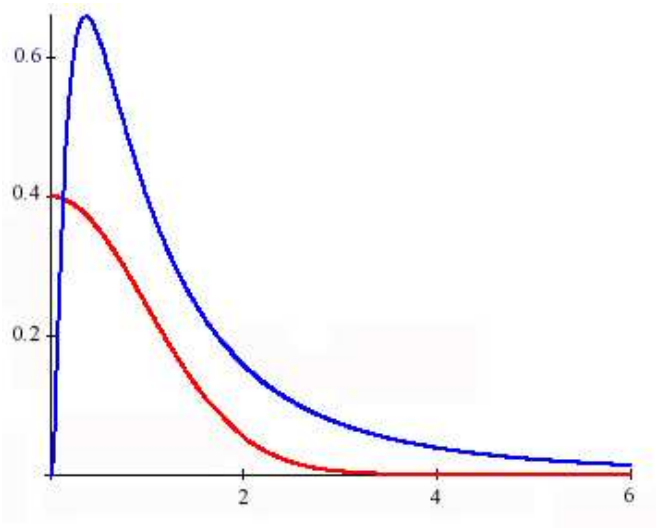


Рис. 28. Графики нормального и логнормального распределения. Среднее значение для нормального распределения выбрано равным нулю

Б. Мандельброт был одним из первых, кто обратил пристальное внимание на то, что не менее универсальным, часто встречаемым законом распределения случайной величины является степенное (часто говорят гиперболическое) распределение с плотностью вероятности:

$$f(x)=\frac{B}{x^{\beta}},$$

или

$$P(X\geq x)=\frac{A}{x^{\alpha}},\qquad 0<x<\infty,\qquad \alpha=\beta-1,$$

где $P(X\geq x)$ - вероятность того, что $X\geq x$, а A и α - некоторые положительные константы, параметры распределения.

Следует отметить, что приведенное выше распределение рассматривалось Б. Мандельбротом (B. Mandelbrot) как уточнение закона Ципфа и его часто называют распределением Ципфа-Мандельброта. При этом оказалось, что α - близкая к единице величина, которая может изменяться в зависимости от свойств текста и языка. Соответственно,

$$P(X \geq x) = \int_x^{\infty} \frac{B}{x^{\beta}} dx, \qquad \frac{\partial P(x)}{\partial x} = -f(x).$$

Справедливости ради надо отметить, что степенные функции распределения рассматривались еще Коши. Как наглядный пример распределения Коши можно привести модель стрельбы из вращающегося в горизонтальной плоскости пулемета (рис. 29).

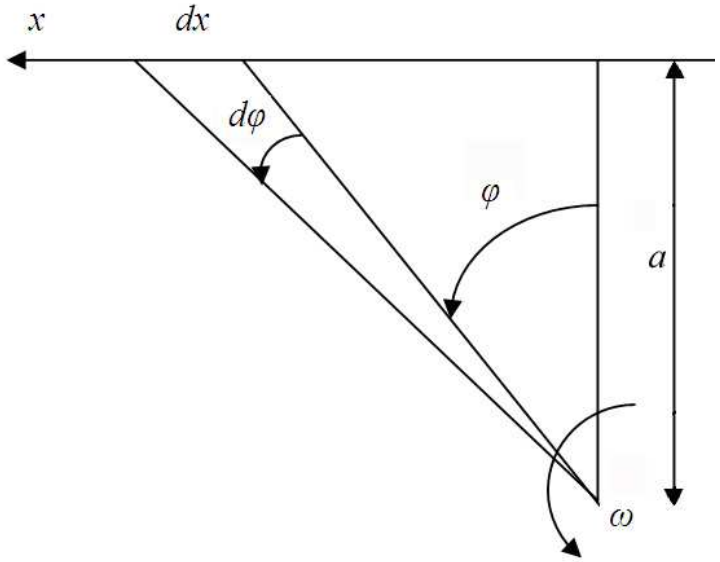


Рис. 29. Модель, приводящая к распределению Коши

Если, производя одиночные выстрелы, нажимать на курок равновероятно при любом его положении, то функция распределения выстрелов по углу - φ будет величиной постоянной: $F(\varphi) = const$. С другой стороны, вероятность попадания в бесконечно малый участок dx бесконечной плоской мишени равна $f(x)dx = F(\varphi)d\varphi$. Откуда, с учетом $x = a \cdot tg(\varphi)$, после элементарных преобразований находим распределение Коши:

$$f(x) = \frac{1}{\pi} \frac{a}{a^2 + x^2}, \qquad -\infty < x < \infty.$$

Так как для этой функции интеграл

$$\int_{-\infty}^{\infty} x^{\alpha} f(x)dx$$

не определён для $\alpha \geq 1$, то ни математическое ожидание, ни дисперсия, ни моменты старших порядков этого распределения не определены. В этом случае говорят, что математическое ожидание не определено, а дисперсия бесконечна.

Напомним, гиперболическое распределение A/x названо в честь В. Парето, а дискретный закон распределения с ранжированной переменной был назван в честь Д. Ципфа, который сформулировал его для описания частоты употребления слов.

Множество примеров применения ранговых статистик и гиперболических распределений приведено, например, в [39].

6.3. Однородные функции и скейлинг

В теории сложных сетей широко применяются методы, приемы и понятия из теоретической физики, которые возникли при рассмотрении конкретных физических задач, например, при построении теории фазовых переходов II рода. Естественно, они несут в своей терминологии отпечаток всех этих физических проблем. Поэтому весьма уместным дать краткое описание некоторых из них, что позволит с пониманием оперировать необходимыми понятиями при чтении следующих глав. К числу таких понятий относятся, например, однородные функции. По определению, однородная функция это такая функция переменной x, что:

$$f(\lambda x) = g(\lambda) f(x),$$

например,

$$f(x) = \alpha x^3.$$

В этом случае:

$$f(\lambda x) = \alpha \lambda^3 x^3 = \lambda^3 f(x), \quad g(\lambda) = \lambda^3.$$

Однородные функции обладают тем свойством, что если известна функция $g(\lambda)$, то, зная значение однородной функции в некоторой точке x_0, можно найти ее значение в любой другой - x. Для этого необходимо ввести масштабный множитель λ, такой что $x = \lambda x_0$, тогда:

$$f(x) = f(\lambda x_0) = g(\lambda) f(x_0).$$

Заметим, что $g(\lambda)$ должна иметь вид:

$$g(\lambda) = \lambda^p,$$

где константа p называется степенью однородности. Кроме того:

$$g(\lambda\mu) = g(\lambda) g(\mu).$$

Аналогично однородной функции одной переменной вводится и однородная функция многих переменных:

$$f(\lambda x_1, \lambda x_2, ...) = g(\lambda) f(x_1, x_2, ...).$$

Например,

$$f(\lambda x, \lambda y) = \lambda^p f(x, y).$$

Эти соотношения можно обобщить, введя так называемую обобщенную однородную функцию, для которой (в случае двух переменных)

$$f(\lambda^a x, \lambda^b y) = \lambda^p f(x, y).$$

Отметим, что в обобщенной однородной функции каждая переменная умножается на свою масштабирующую константу $\lambda^a, \lambda^b, ...$

Самое важное для различных приложений свойство однородной функции заключается в том, что ее можно свести к функции меньшего числа переменных. Например, для однородной функции двух переменных из соотношения

$$f(\lambda x, \lambda y) = \lambda^p f(x, y),$$

выбирая $\lambda = 1/y$ находим:

$$f(\lambda x, \lambda y) = f\left(\frac{x}{y}, 1\right) = y^{-p} f(x, y)$$

или

$$f(x, y) = y^{p} f\left(\frac{x}{y}, 1\right).$$

Вводя обозначение:

$$F\left(\frac{x}{y}\right) = f\left(\frac{x}{y}, 1\right)$$

имеем:

$$f(x, y) = y^{p} F\left(\frac{x}{y}\right).$$

Таким образом, функция двух переменных $f(x, y)$, если она является однородной функцией, может быть сведена к функции одной переменной $F(z)$.

Функция $F(z)$ называется скейлинговой функцией. Это название связано с тем, что термин «скейл» - масштаб подразумевает, что переменная x, в нашем примере, в скейлинговой функции $F(z)$ описывается в «масштабе» переменной y – $z = x / y$.

Для обобщенной однородной функции скейлинговая функция включает в себя более сложную, нежели x / y комбинацию переменных. В самом деле, выбирая для $f(\lambda^{a} x, \lambda^{b} y) = \lambda^{p} f(x, y)$ значение $\lambda = 1 / y^{1/b}$, после элементарных преобразований, аналогичных приведенным выше, получаем:

$$f(x, y) = y^{\alpha} F(z), \qquad z = \frac{x}{y^{a/b}}, \qquad \alpha = \frac{p}{b}.$$

Приведем вначале пример из физики. Задача описания спектральной плотности излучения абсолютно черного тела $u(\omega, T)$ (ω - частота, T - температура) потребовала введения принципиально нового направления - квантовой физики. До ее создания, некоторые общие положения относительно

функции $u(\omega,T)$ удалось сформулировать в терминах понятий однородных функций. В частности, это так называемая теорема В. Вина, согласно которой:

$$u(\omega,T)=\omega^3\varphi(\omega/T).$$

Сама скейлинговая функция $\varphi(z)$ этой теоремой не определялась, однако уже то, что $u(\omega,T)$ выражается через однородную функцию, позволило получить нетривиальные выводы. Например, то, что значение частоты $\omega_{\max}$, при которой $u(\omega,T)$ достигает максимального значения прямо пропорционально температуре ($\omega_{\max}\sim T$, так называемая формула смещения Вина).

Еще один пример скейлингового соотношения можно привести из теории так называемых «малых миров»:

$$\ell(N,p)\sim p^{-\tau}\psi(Np^{\tau}),$$

где $\ell(N,p)$ средний путь от одного узла сети до другого при условии, что каждый узел связан с некоторым (небольшим) числом своих соседей и, кроме того, есть случайные связи (их число порядка pN) между случайными узлами. В этом случае для скейлинговой функции установлено, что:

$$\psi(z<<1)\sim x,\ \ \psi(z>>1)\sim \ln x.$$

Ниже мы увидим примеры скейлинга в теории перколяции (протекания).

6.4. Параметр порядка и фазовые переходы

Для того чтобы говорить о фазовых переходах, необходимо определить, что такое фазы. Понятие фаз встречается во множестве явлений, поэтому вместо того, чтобы давать общее определение (чем оно более общее, тем оно, как и положено, более абстрактное и ненаглядное), приведем несколько примеров.

Вначале пример их физики. Для обычной, наиболее часто встречаемой в нашей жизни жидкости – воды известно три фазы: жидкая, твердая (лед) и газообразная (пар). Каждая из них характеризуется своими значениями

параметров. Существенно то, что при изменении внешних условий одна фаза (лед) переходит в другую (жидкость). Еще один любимый объект теоретиков – ферромагнетик (железо, никель и множество других чистых металлов и сплавов). При низких температурах (для никеля ниже $T_c = 360^0 C$) никель является ферромагнетиком, при снятии внешнего магнитного поля он остается намагниченным, т.е. может использоваться как постоянный магнит. При температуре выше T_c это свойство теряется, при выключении внешнего магнитного поля он переходит в парамагнитное состояние и не является постоянным магнитом. Здесь, как и в предыдущем примере, четко видно существование двух фаз – парамагнитной и ферромагнитной. При изменении температуры происходит переход – фазовый переход – из одной фазы в другую.

В [52] приведен геометрический пример из теории перколяции. Случайно вырезая из сетки связи, в конце концов, когда концентрация оставшихся связей - p станет меньше некоторого значения p_c, по решетке уже нельзя будет пройти «из конца в конец». Таким образом, сетка из состояния протекания - фаза «протекания», перейдет в состояние фазы «непротекания».

Из этих примеров ясно, что для каждой из рассмотренных систем существует так называемый параметр порядка, определяющий в какой из фаз находится система. В ферромагнетизме параметр порядка – намагниченность в нулевом внешнем поле, в теории перколяции – связность сетки.

Фазовые переходы бывают разного рода. Фазовый переходы I рода - это такой переход, когда в системе может одновременно существовать несколько фаз. Например, при температуре $0^0 C$ лед плавает в воде. Если система находится в термодинамическом равновесии (нет привода и отвода тепла), то лед не тает и не нарастает. Для фазовых переходов II рода существование одновременно нескольких состояний невозможно. Кусочек никеля либо находится в парамагнитном состоянии η, либо в ферромагнитном. Сетка со случайно вырезанными связями либо связна, либо нет.

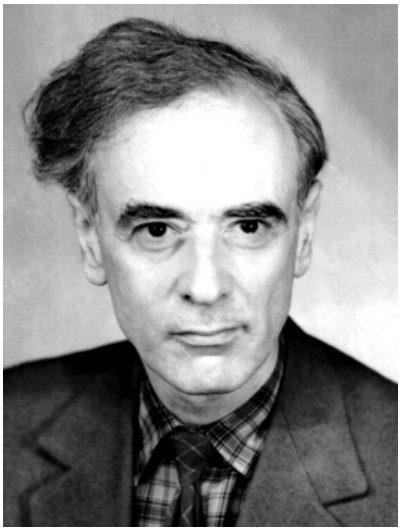
Решающим в создании теории фазовых переходов II рода, начало которой положил Л.Д. Ландау, было введение параметра порядка (будем обозначать его

η) как отличительного признака фазы системы. В одной из фаз, например, парамагнитной, $\eta = 0$, а в другой, ферромагнитной, $\eta \neq 0$. Для магнитных явлений параметр порядка η - это намагниченность системы.

Для описания фазовых переходов вводится некоторая функция параметров, определяющих состояние системы - $G(\eta, T, ...)$. В физических системах это энергия Гиббса. В каждом явлении (перколяция, сеть «малых миров» и т.д.) эта функция определятся «самостоятельно». Главное свойство этой функции, первое предположение Л.Д. Ландау – в состоянии равновесия эта функция принимает минимальное значение:

$$\frac{\partial G}{\partial \eta} = 0, \qquad \frac{\partial^2 G}{\partial \eta^2} > 0.$$

В физических системах говорят о термодинамическом равновесии, в теории сложных цепей можно говорить об устойчивости. Заметим, что условие минимальности определяется варьированием параметра порядка.



Лев Давидович Ландау (1908-1968)

Второе предположение Л.Д. Ландау – при фазовом превращении $\eta = 0$. Согласно этому предположению, функцию $G(\eta, T, ...)$ вблизи точки фазового перехода можно разложить в ряд по степеням параметра порядка η:

$$G(\eta, T) = G_0(T) + A(T)\eta^2 + B\eta^4 + ...,$$

где $\eta = 0$ в одной фазе (парамагнитной, если речь идет о магнетизме и несвязной, если о сетке) и $\eta \neq 0$ в другой (ферромагнитной или связной).

Из условия $\partial G / \partial \eta = 0$ находим:

$$2A\eta + B\eta^3 = 0,$$

что дает нам два решения $\eta = 0$ и $\eta = -A/2B \neq 0$.

Для $T > T_c$ должно иметь место решение $\eta = 0$, а для $T < T_c$ решение $\eta \neq 0$. Этому можно удовлетворить, если для случая $T > T_c$ и $\eta = 0$ выбрать $A > 0$. В этом случае второго корня не существует. А для случая $T < T_c$ должно иметь место второе решение, т.е. должно выполняться $A < 0$. Таким образом:

$A > 0$ при $T > T_c$,

$A < 0$ при $T < T_c$,

$A(T_c) = 0$ – второе предположение Ландау.

Простейший вид A, удовлетворяющий этим требованиям, есть

$$A = \alpha(T - T_c).$$

Тогда

$$\eta^2 = -A/2B \sim \alpha(T - T_c),$$

откуда

$\eta \sim \sqrt{T_c - T}$, а функция $G(\eta, T)$ принимает вид

$$G(\eta, T) = G_0(T) + \alpha(T - T_c)\eta^2 + B\eta^4 + \ldots.$$

На рис. 30 изображена зависимость $G(\eta, T)$ для $T > T_c$ и $T < T_c$.

В теории фазовых переходов II рода интересно не только поведение функции $G(\eta, T)$, но и ее производных по температуре (или концентрации разорванных связей, если речь идет о сложных сетях). В физических задачах эти производные носят названия энтропии S и теплоемкости C

$$S = -\frac{\partial G}{\partial T}, \quad C = -T\frac{\partial S}{\partial T}.$$

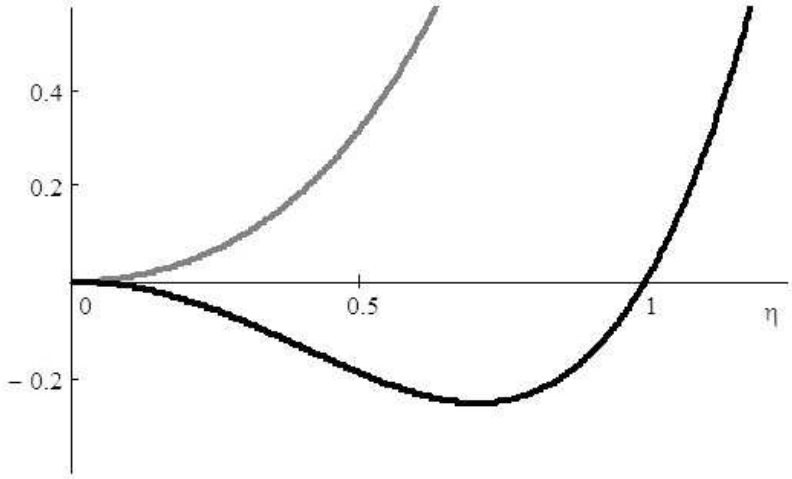


Рис. 30. Графики функции параметров $G(\eta,T)$ *для* $T > T_c$ *и* $T < T_c$

7. ЭНТРОПИЯ И КОЛИЧЕСТВО ИНФОРМАЦИИ

«Исчезни все, мне чуждое! исчезни город каменный!
Исчезни все, гнетущее! исчезни вся вселенная!
Все краткое, все хрупкое, все мелкое! все тленное!»
Игорь Северянин

Одной из основ современной теории информационного поиска является классическая теория информации, оформившаяся в 40-х годах XX века благодаря работам К. Шеннона (C.E. Shannon) [62].

Понятие энтропии первоначально возникло в физике, в таком ее разделе как термодинамика, а позже в статистической физике. В термодинамике изучают макроскопические состояния систем, которые задаются макроскопическими параметрами, такими как, например, энергия, объем, давление и т.п. В статистической физике, кроме понятия макроскопических состояний, вводят понятие микроскопического состояния, определяемого так называемыми микроскопическими параметрами, например, значениями в данный момент времени всех импульсов и координат всех частиц, из которых состоит система.

Естественно, одному макросостоянию может соответствовать множество микросостояний. В статистической физике все микросостояния считаются равновероятными, речь идет о микроканоническом ансамбле (подробности приведены в серии книг «Теоретическая физика» Л.Д. Ландау и Е.М. Лифшица). Поэтому, чем больше микросостояний соответствуют данному макросостоянию, тем большая вероятность этого макросостояния. В качестве классического примера рассматривается закрытый ящик, мысленно разделенный на две равные части. По всему объему ящика равномерно распределены частицы, каждая из которых равновероятно может находиться как в левой, так и в правой части. Выбирается количество частиц, равное 100.

Первое из рассматриваемых макросостояний следующее: все частицы расположены в левой части, этому макросостоянию соответствует только одно микросостояние ($N=1$). Второе макросостояние следующее – в левой части находится только одна частица - такому макросостоянию соответствует уже сто

микросостояний ($N=100$). В качестве третьего макросостояния выбирается то, которое чаще всего наблюдается на практике - частицы равномерно распределены по объему – количество частиц в левой части составляет половину от всего числа частиц. Такое макросостояние должно быть наиболее вероятным, соответственно, ему должно соответствовать наибольшее число микросостояний. Действительно, как показывает простое вычисление, количество сочетаний из 100 элементов по 50 составляет $N \approx 10^{29}$. При таком гигантском отличие числа микросостояний для разных макросостояний ясно, что вероятностью встретить первое макросостояние, по сравнению с вероятностью встретить третье ($\sim 10^{-29}$) можно пренебречь. В реальных системах, как правило, количество частиц значительно большее – порядка 10^{24} (и больше), соответственно количество микросостояний становится очень большими.

Чтобы не работать с большими числами рассматривают логарифм от количества микросостояний, соответствующих данному макросостоянию, который и называют энтропией:

$$S = k \ln N ,$$

где k - некоторая константа, которую в физике выбирают раной постоянной Больцмана.

В случае, когда все микросостояния равновероятны $p_i = p = 1/N = const$ выражение для энтропии может быть записано как:

$$S = k \ln N = k \sum_{i=1}^{N} p \ln N = -k \sum_{i=1}^{N} p \ln p .$$

Вообще говоря, вероятности микросостояний могут быть разными, в этом случае выражение для энтропии надо записывать так:

$$S = -k \sum_{i=1}^{N} p_i \ln p_i .$$

В теории информации константу k принято выбирать равной $k = 1/\ln 2$ (информация измеряется в битах!), таким образом:

$$S = -\sum_{i=1}^{N} p_i \log_2 p_i ,$$

это и есть как раз энтропия, которая была предложена К. Шенноном.

Классическая теория информации была ориентирована прежде всего на исследование процессов передачи данных по каналам связи. Благодаря использованию таких понятий, как информационная энтропия, количество информации, взаимная информация и т.д. теория информации приобрела универсальный характер, и ее методы стали широко использоваться во многих областях науки и технологий. Многие эффективные методы решения задач глубинного анализа текстов базируются на понятии взаимной информации (mutual information), которая широко используется, в частности, в области статистической обработки естественных языков [111], позволяя определять близость между словами или какими-либо другими языковыми явлениями. В этой области взаимная информация описывает количество информации о принадлежности документа к определенной категории c, которое, например, связано с наличием некоторого терма t. В этом случае взаимная информация определяется по формуле:

$$I(t,c) = \log \frac{P(t,c)}{P(t)P(c)},$$

где $P(t,c)$ - эмпирически оцененная вероятность одновременной встречаемости терма t и принадлежности документа к категории c; $P(t)$ - вероятность появления терма t, $P(c)$ - вероятность принадлежности документа к категории c.

Таким образом, взаимная информация между термом и категорией описывает степень ассоциации терма t и категории c.

На информационной теории базируются очень многие информационно-поисковые и аналитические системы. Так, в частности, компания Autonomy создала аналитический сервер IDOL (Intelligent Data Operating Layer), идеология которого базируется на использовании байесовских вероятностей и теории Шеннона, которая рассматривается как математическая основа построения коммуникационных систем, позволяющая определять и интерпретировать численные значения количества информации. По мнению создателей сервера IDOL, естественные языки обладают высокой степенью избыточности, несущественного содержания. С помощью анализа энтропии, а точнее, используя

методологию взаимной информации, сервер IDOL обеспечивает извлечение «сущности» из избыточных текстов. По мнению идеологов системы IDOL, чем реже контекст встречается в процессе коммуникации, тем он важнее, тем больше информации он передает. Благодаря такому подходу обеспечивается нахождение наиболее информативных понятий в документах.

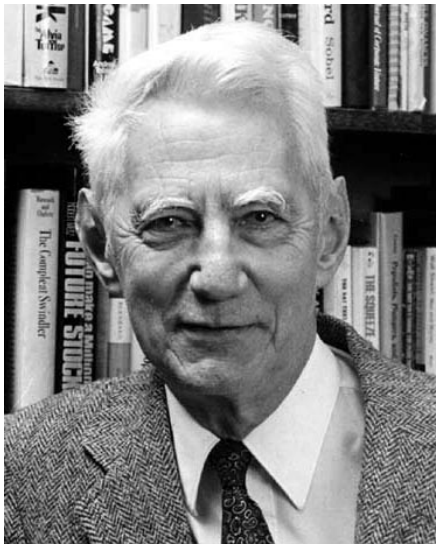
Остановимся подробнее на основных моментах классической теории информации и ее применимости к теории и практике информационного поиска.

7.1. Энтропия Шеннона

Клодом Шенноном была предложена энтропия как мера неопределенности ансамбля $U = \{u_1, \ ..., \ u_N\}$, определяемая следующим функционалом:

$$H(U) = -K\sum_{i=1}^{N} p_i \log_2 p_i ,$$

где p_i - вероятность состояния u_i, K - неотрицательная константа.



Клод Шеннон (1916 -2001)

В случае, если все состояния источника информации равновероятны, формула для энтропии принимает вид:

$$H(U) = -\sum_{i=1}^{N} p_i \log_2 p_i = -\sum_{i=1}^{N} \frac{1}{N} \log_2 \frac{1}{N} = -\log_2 \frac{1}{N} = \log_2 N,$$

который совпадает с мерой Хартли, таким образом подтверждая тот факт, что она является частным случаем энтропии Шеннона.

Для пояснения понятия информационной энтропии можно рассмотреть процесс получения сообщения длиной N символов (букв или пробела).

Итак, пусть передается сообщение, состоящее из n различных символов - $u_1,\ u_2,\ ...,\ u_n$. Данное сообщение можно представить в виде таблицы:

u_{10}	u_5	u_{21}	…	u_3
1	2	3	…	N

где первая строка - это символы сообщения, а вторая – соответствующие этим символам номера мест в сообщении.

Пусть для любого i символ u_i $(i = 1,\ 2,\ ...,\ n)$ генерируется с вероятностью p_i, причем это значение не зависит от предыдущих символов. Тогда при достаточно большом N количество символов u_i будет с высокой точностью соответствовать значению Np_i. Таким образом, вероятность p получить сообщение, в котором содержится Np_1 символов u_1, Np_2 символов u_2 и т.д. (без учета их местоположения в сообщении), равна:

$$p = p_1^{Np_1} p_2^{Np_2} ... p_n^{Np_n}.$$

Двоичный логарифм от этой вероятности можно записать следующим образом:

$$\log_2 p = N \sum_{i=1}^{n} p_i \log_2 p_i .$$

Сомножитель у N, взятый с обратным знаком, и есть энтропия Шеннона:

$$S = -\sum_{i=1}^{n} p_i \log_2 p_i ,$$

Таким образом вероятность появления сообщения длиной N символов с указанными выше свойствами, равна:

$p = 2^{-NS}$.

Так как все подобные сообщения равновероятны (с вероятностью p), то их число K равно:

$$K = \frac{1}{p} = 2^{NS}.$$

Таким образом, информационная энтропия (или энтропия Шеннона) определяет количество сообщений, в которых символы встречаются с «правильной» по статистическим соображениям частотой (u_1 с p_1, u_2 с p_2, и т.д.).

Следует заметить, что введенная Шенноном энтропия - это та же энтропия из физики, хотя и используется она для других целей, о которых речь пойдет позже. По физической терминологии, макросостояние задается набором $\{p_1,\ p_2,\ ...,\ p_n\}$. Каждому макросостоянию соответствует $K(\{p_1, p_2, ...p_n\}) = 2^{NS(\{p_1, p_2, ... p_n\})}$ микросостояний.

Для пояснения, приведем два примера.

Макросостоянию $\{1,\ 0,\ ...,\ 0\}$ (с вероятность 1 встречается символ u_1) соответствует только одно сообщение, т.е. только одно микросостояние - $(u_1,\ u_1,\ ...,\ u_1)$. Энтропия такого макросостояния (с учетом известного предела $0\log 0 = 0$) равна:

$$S = -\sum_{i=1}^{n} p_i \log_2 p_i = -1 \cdot \log_2 1 - 0 \cdot \log_2 0 - ... = 0.$$

Такой результат вполне соответствует интуиции, неопределенности нет, сообщение, которое мы можем получить, полностью определено (предсказуемо) – энтропия минимальна.

А вот для, например, макросостоянию, в котором каждый символ встречается с одной и той же вероятностью $p_i = 1/n$ - $\{1/n,\ 1/n,\ ...,\ 1/n\}$ соответствует намного большая энтропия:

$$S = -\sum_{i=1}^{n} p_i \log_2 p_i = \log_2 n .$$

Количество разных сообщений длиной N, в которых каждый символ встречается N/n раз естественно намного больше единицы, оно равно

$$K\left(\{1/n, 1/n, \ldots 1/n\}\right) = 2^{NS\left(\{1/n, 1/n, \ldots 1/n\}\right)} = 2^{N \log_2 n} = n^N .$$

7.2. Свойства энтропии

Энтропия, введенная Шенноном как мера неопределенности состояния дискретного источника информации, обладает следующими свойствами:

1) Энтропия является вещественной неотрицательной величиной в интервале [0, 1].
2) Энтропия - величина ограниченная.
3) Энтропия равна нулю лишь тогда, когда вероятность одного из состояний равна единице, т.е. состояние источника точно определено.

Рассмотрим график зависимости энтропии источника с двумя состояниями, характеризующимися вероятностями p и $1-p$, соответственно (рис. 31).

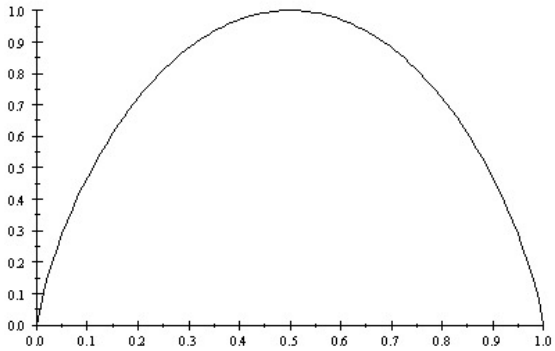


Рис. 31. Энтропия системы с двумя состояниями
(ось абсцисс - p, ось ординат – энтропия H)

Энтропия в этом случае равна:

$$H(U) = -\left[p \log p + (1-p) \log(1-p) \right].$$

Для построения графика энтропии системы с тремя состояниями (рис. 32), характеризующимися вероятностями $p, q, 1-p-q$, используется формула:

$$H(U) = -\left[p \log p + q \log q + (1-p-q) \log(1-p-q) \right].$$

На практике очень важным является и такое свойство энтропии, как полное игнорирование содержательной стороны состояний источника, а лишь учет вероятности этих состояний. При этом рассматривается лишь степень неопределенности. Так, например, если рассмотреть полное множество исходов лечения больного, состоящее из двух событий – благоприятного и неблагоприятного, то результат благоприятного исхода лечения больного с вероятностью 0.9 и с вероятностью 0.1 имеют одинаковое значение энтропии [17].

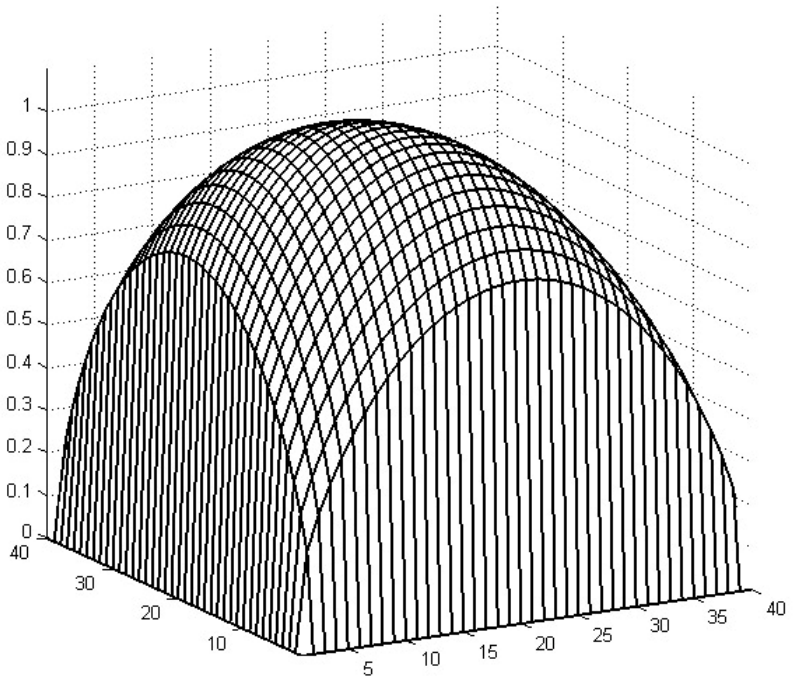


Рис. 32. Энтропия системы с тремя состояниями
(ось OX - p, ось OY – q, ось OZ - энтропия)

7.3. Условная энтропия

Определим энтропию объединения двух статистически связанных ансамблей U и V, которое характеризуется матрицей вероятностей $p(U,V) = \left\| p(u_i, v_j) \right\|$, $(i = 1, \ldots, N; \quad j = 1, \ldots, M)$.

Из теории вероятностей известно: $p(u_i, v_j) = p(u_i)p(v_j / u_i) = p(v_j)p(u_i / v_j)$, соответственно, энтропия объединения событий выражается формулой:

$$H(U,V) = -\sum_{i=1}^{N}\sum_{j=1}^{M} p(u_i v_j)\log p(u_i v_j) = -\sum_{i=1}^{N}\sum_{j=1}^{M} p(u_i)\,p(v_j / u_i)\log[p(u_i)\,p(v_j / u_i)] =$$

$$= -\sum_{i=1}^{N} p(u_i)\log p(u_i) - \sum_{i=1}^{N} p(u_i)\cdot\sum_{j=1}^{M} p(v_j / u_i)\log p(v_j / u_i).$$

Назовем $H_{u_i}(V)$ частной условной энтропией ансамбля V по отношению к состоянию $u_i \in U$:

$$H_{u_i}(V) = -\sum_{j=1}^{M} p(v_j / u_i)\log p(v_j / u_i).$$

Соответственно, степень неопределенности, приходящаяся на одно состояние ансамбля V при известных состояниях ансамбля U или условная энтропия V по отношению к U определяется как:

$$H_U(V) = \sum_{j=1}^{N} p(u_i)H_{u_i}(V) = -\sum_{i=1}^{N} p(u_i)\sum_{j=1}^{M} p(v_j / u_i)\log p(v_j / u_i).$$

Условная энтропия обладает такими основными свойствами:

1) Энтропия объединения двух ансамблей V и U равна безусловной энтропии одного ансамбля плюс условная энтропия другого относительно первого:

   $H(UV) = H(U) + H_U(V),$

   $H(UV) = H(V) + H_V(U).$

2) Наличие сведений о результатах реализации состояния одного ансамбля никак не может увеличить неопределенность выбора состояния из другого ансамбля:

   $H_U(V) \le H(V),$

   $H_V(U) \le H(U).$

3) В случае отсутствия статистической связи в реализациях состояний из ансамблей U и V:

   $H_U(V) = H(V),$

   $H_V(U) = H(U).$

7.4. Энтропия непрерывного источника информации

Непрерывным источником информации будем называть источники, множество состояний которых составляет континуум.

Вероятности значений непрерывного источника информации не могут непосредственно использоваться для оценки неопределенности, так как в этом случае вероятность любого конкретного значения равна нулю. Поэтому в данном случае естественно перейти к рассмотрению плотности распределения вероятностей.

Разобьем диапазон случайной величины U с плотностью вероятностей $p(u)$ на n интервалов шириной Δu. Тогда для вероятности нахождения случайной величины в заданном интервале справедливо:

$$p(u_i \le u \le u_i + \Delta u) = \int_{u_i}^{u_i+\Delta u} p(u)du \approx p(u_i)\Delta u.$$

Если перейти от случайной величины U к дискретной случайной величине $\tilde{U}$ с n состояниями и вероятностями состояний, равными по значению $p(u_i \le u \le u_i + \Delta u)$, то можно записать формулу для энтропии $\tilde{U}$:

$$H(\tilde{U}) = -\sum_{i=1}^{n} p(u_i)\Delta u \log\left[p(u_i)\Delta u\right] = -\sum_{i=1}^{n} p(u_i)\Delta u \log p(u_i) - \sum_{i=1}^{n} p(u_i)\Delta u \log \Delta u.$$

Так как $\sum_{i=1}^{n} p(u_i)\Delta u \approx 1,$ то справедливо:

$$H(\tilde{U}) \approx -\sum_{i=1}^{n} p(u_i)\Delta u \log p(u_i) - \log \Delta u.$$

Предполагая, что энтропия непрерывного источника информации совпадает с пределом при $\Delta u \to 0$ энтропии определенного выше дискретного источника информации, имеем:

$$H(U) = \lim_{\Delta u \to 0} H(\tilde{U}) = -\int_{-\infty}^{\infty} p(u) \log p(u) du - \lim_{\Delta u \to 0} \log \Delta u.$$

Именно из-за второго члена приведенного выражения величина $H(U)$ является бесконечной (что соответствует бесконечной неопределенности при

выборе из континуума состояний). Вместе с тем первый член выражения похож на сумму, соответствующую дискретному источнику информации [9]. Поэтому в качестве меры неопределенности непрерывного источника информации договорились использовать именно первый член приведенного выше выражения:

$$h(U) = -\int_{-\infty}^{\infty} p(u)\log p(u)du.$$

Такая мера неопределенности получила название дифференциальной энтропии.

Приведем некоторые свойства дифференциальной энтропии.

1) Дифференциальная энтропия является относительной мерой неопределенности, зависящей от масштаба величины U. Если изменить масштаб случайной величины U в k раз, введя $u_k = ku$, соответственно, $p(u_k) = p(u)/k$, то дифференциальная энтропия будет зависеть от логарифма k:

$$h(U) = -\int_{-\infty}^{\infty} p(u_k)\log p(u_k)du_k = -\int_{-\infty}^{\infty} \frac{p(u)}{k}\log\left[\frac{p(u)}{k}\right]kdu = h(u) + \log k.$$

Таким образом, $h(U)$ не может служить абсолютной мерой неопределенности непрерывного сообщения. Кроме того, дифференциальная энтропия может принимать положительные, отрицательные и нулевые значения.

2) Для дифференциальной энтропии объединения статистически зависимых источников справедливы те же соотношения, что и для дискретных источников:

$h(UV) = h(U) + h_U(V) = h(V) + h_v(U),$

где $h_U(V)$, $h_v(U)$ - условные дифференциальные энтропии.

3) Максимальной дифференциальной энтропией при заданной дисперсии σ обладает нормальное распределение вероятностей:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}},$$

$$h_{\max}(U) = \log_2 \sigma\sqrt{2\pi e}.$$

Как видим, дифференциальная энтропия в этом случае не зависит от математического ожидания m.

7.5. Количество информации

Рассмотрим дискретный источник Z как множество возможных сообщений $Z = \{z_1, \ldots, z_N\}$. Допустим, сообщения передаются по каналу связи и принимаются как некоторое новое, искаженное множество $W = \{w_1, \ldots, w_N\}$.

Средняя неопределенность относительно любого состояния источника, остающаяся у адресата после получения сообщения w_j характеризуется условной энтропией:

$$H_{w_j}(Z) = -\sum_{i=1}^{N} p(z_i / w_j) \log p(z_i / w_j).$$

Тогда средняя неопределенность по всему ансамблю принимаемых сообщений равна сумме по всем j:

$$H_W(Z) = -\sum_{i=1}^{N} p(w_j) H_{w_j}(Z).$$

Определим количество информации, содержащееся в каждом принятом элементе сообщения относительно любого переданного сообщения, следующим образом:

$$I(ZW) = H(Z) - H_W(Z).$$

Очевидно, справедливо соотношение:

$$I(ZW) = \sum_{i=1}^{N} \sum_{j=1}^{N} p(z_i w_j) \log \frac{p(z_i w_j)}{p(z_i)\, p(w_j)}.$$

Рассмотрим некоторые свойства количества информации:

1) Количество информации величина неотрицательная. Действительно,
   $H(Z) \geq H_W(Z) \;\Rightarrow\; I(ZW) = H(Z) - H_W(Z) \geq 0.$
2) При отсутствии статистической связи между Z и W:
   $H(Z) = H_W(Z) \;\Rightarrow\; I(ZW) = 0.$
3) $I(ZW) = I(WZ)$. Действительно:

$$I(ZW) = H(Z) - H_W(Z) = H(ZW),$$

$$I(WZ) = H(W) - H_Z(W) = H(WZ).$$

При этом $H(ZW) = H(WZ)$.

4) При взаимно однозначном соответствии между Z и W:

$$I(ZW) = H(Z).$$

Это максимальное количество информации о состоянии дискретного источника.

7.6. Взаимная информация

Взаимная информация определяется аналогично понятию количества информации, содержащейся в каждом принятом элементе сообщения, относительно любого переданного сообщения. Вместе с тем, приведенные выше формулы для количества информации характеризуют информационные свойства одного дискретного источника или ансамбля. Однако в теории информационного поиска особенный интерес представляет выявление количества информации в ансамбле категорий V, количество которых равно N, относительно другого – словаря текста U, содержащим M ключевых слов (термов). Можно рассматривать и другие интерпретации этой задачи, например, относящиеся к взаимной информации, содержащейся в параллельно генерируемых текстах.

Для определения такой информационной характеристики рассмотрим условную энтропию $H_U(V)$, определяющую среднее количество информации, выдаваемое сообщением ансамбля V при условии, что сообщение ансамбля U уже известно. Эта условная энтропия задается формулой:

$$H_U(V) = \sum_{j=1}^{M} P(u_j) H_{u_j}(V).$$

Подставляя в эту формулу значение частной условной энтропии $H_{u_i}(V)$, получаем:

$$H_U(V) = -\sum_{j=1}^{M} \sum_{k=1}^{N} P(v_k, u_j) \cdot \log(P(v_k, u_j) / P(u_j)).$$

Взаимная информация между V и U определяется как:

$$I(V,U) = H(V) - H_U(V).$$

Взаимная информация измеряется в тех же единицах что и энтропия, (например, в битах). Величина $I(V,U)$ показывает, сколько в среднем бит информации о реализации ансамбля V дает наблюдение о реализации ансамбля U.

Выразим взаимную информацию через вероятности:

$$I(V,U) = H(V) - H_U(V) =$$

$$= -\sum_{k=1}^{N} P(v_k) \cdot \log P(v_k) + \sum_{k=1}^{N}\sum_{j=1}^{M} P(v_k,u_j) \cdot \log(P(v_k,u_j)/P(u_j)) =$$

$$= -\sum_{k=1}^{N}\sum_{j=1}^{M} P(v_k,u_j) \cdot \log P(v_k) + \sum_{k=1}^{N}\sum_{j=1}^{M} P(v_k,u_j) \cdot \log(P(v_k,u_j)/P(u_j)) =$$

$$= \sum_{k=1}^{N}\sum_{j=1}^{M} P(v_k,u_j) \cdot \log \frac{P(v_k,u_j)}{P(v_k) \cdot P(u_j)}.$$

Взаимная информация обладает следующими свойствами:

1) $I(V,U) \geq 0$, причем равенство имеет место только в том случае, когда V и U взаимно независимы.
2) $I(V,U) = I(U,V)$, т.е U содержит столько же информации относительно V, сколько V содержит относительно U. Поэтому можно так же записать $I(V,U) = H(U) - H_V(U)$.
3) $I(V,U) \leq H(V)$, $I(V,U) \leq H(U)$, причем равенство имеет место, когда по реализации U можно точно восстановить реализацию V или наоборот.
4) $I(V,V) = H(V)$, что позволяет интерпретировать энтропию источника как информацию ансамбля V о самом себе.

8. ОСНОВЫ ТЕОРИИ СЛОЖНЫХ СЕТЕЙ

«…но всегда легче создавать новые сложности,
чем распутывать старые.»
Нил Стивенсон

В последнее время все большую популярность получает область дискретной математики, называемая теорией сложных сетей (complex networks) [116], изучающая характеристики сетей, учитывая не только их топологию, но и статистические феномены, распределение весов отдельных узлов и ребер, эффекты протекания и проводимости в таких сетях тока, жидкости, информации и т.д. Оказалось, что свойства многих реальных сетей существенно отличаются от свойств классических случайных графов.

Несмотря на то, что в рассмотрение теории сложных сетей попадают различные сети – электрические, транспортные, информационные, наибольший вклад в развитие этой теории внесли исследования социальных сетей. Термин «социальная сеть» обозначает сосредоточение социальных объектов, которые можно рассматривать как сеть (или граф), узлы которой - объекты, а связи - социальные отношения. Этот термин был введен в 1954 году социологом из «Манчестерской школы» Дж. Барнсом (J. Barnes) в работе «Классы и сборы в норвежском островном приходе». Во второй половине XX столетия понятие «социальная сеть» стало популярным у западных исследователей, при этом как узлы социальных сетей стали рассматривать не только представителей социума, но и другие объекты, которым присущий социальные связи. В теории социальных сетей получило развитие такое направление, как анализ социальных сетей (Social Network Analysis, SNA). Сегодня термин «социальная сеть» обозначает понятие, оказавшееся шире своего социального аспекта, оно включает, например, многие информационные сети, в том числе и WWW.

В рамках теории сложных сетей рассматривают не только статистические, но динамические сети, для понимания структуры которых необходимо учитывать принципы их эволюции [87].

8.1. Параметры сложных сетей

В теории сложных сетей выделяют три основных направления: исследование статистических свойств, которые характеризуют поведение сетей; создание модели сетей; предсказание поведения сетей при изменении структурных свойств. В прикладных исследованиях обычно применяют такие типичные для сетевого анализа характеристики, как размер сети, сетевая плотность, степень центральности и т.п.

При анализе сложных сетей как и в теории графов исследуются параметры отдельных узлов; параметры сети в целом; сетевые подструктуры.

8.1.1. Параметры узлов сети

Для отдельных узлов выделяют следующие параметры:

- входная степень узла - количество ребер графа, которые входят в узел;
- выходная степень узла - количество ребер графа, которые выходят из узла;
- расстояние от данного узла до каждого из других;
- среднее расстояние от данного узла до других;
- эксцентричность (eccentricity) - наибольшее из геодезических расстояний (минимальных расстояние между узлами) от данного узла к другим;
- посредничество (betwetnness), показывающее, сколько кратчайших путей проходит через данный узел;
- центральность - общее количество связей данного узла по отношению к другим.

8.1.2. Общие параметры сети

Для расчета индексов сети в целом используют такие параметры, как: число узлов, число ребер, геодезическое расстояние между узлами, среднее расстояние от одного узла к другим, плотность - отношение количества ребер в сети к возможному максимальному количеству ребер при данном количестве узлов,

количество симметричных, транзитивных и циклических триад, диаметр сети - наибольшее геодезическое расстояние в сети и т.д..

Существует несколько актуальных задач исследования сложных сетей, среди которых можно выделить следующие основные:

- определение клик в сети. Клики - это подгруппы или кластеры, в которых узлы связаны между собой сильнее, чем с членами других клик;
- выделение компонент (частей сети), которые связаны внутри и не связаны между собой;
- нахождение блоков и перемычек. Узел называется перемычкой, если при его изъятии сеть распадается на несвязанные части;
- выделение группировок - групп эквивалентных узлов (которые имеют максимально похожие профили связей).

8.1.3. Распределение степеней узлов

Важной характеристикой сети является функция распределения степеней узлов $P(k)$, которая определяется как вероятность того, что узел i имеет степень $k_i = k$. Сети, характеризующиеся разными $P(k)$, демонстрируют весьма разное поведение. $P(k)$ в некоторых случаях может быть распределениями Пуассона ($P(k) = e^{-m} m^k / k!$, где m – математическое ожидание), экспоненциальным ($P(k) = e^{-k/m}$) или степенным ($P(k) \sim 1/k^{\gamma}$, $k \neq 0$, $\gamma > 0$).

Сети со степенным распределением степеней узлов называются безмасштабными (scale-free). Именно безмасштабные распределения часто наблюдаются в реально существующих сложных сетях. При степенном распределении возможно существование узлов с очень высокой степенью, что практически не наблюдается в сетях с пуассоновым распределением.

8.1.4. Путь между узлами

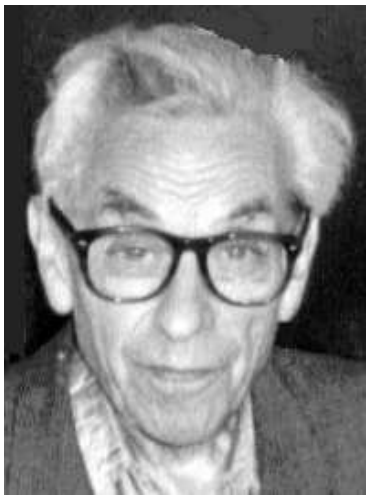
Расстояние между узлами определяется как количество шагов, которые необходимо сделать, чтобы по существующим ребрам добраться от одного узла

до другого. Естественно, узлы могут быть соединены прямо или опосредованно. Путем между узлами d_{ij} назовем кратчайшее расстояние между ними. Для всей сети можно ввести понятие среднего пути, как среднее по всем парам узлов кратчайшего расстояния между ними:

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

где n - количество узлов, d_{ij} – кратчайшее расстояние между узлами i и j.

Венгерскими математиками П. Эрдёшем (P. Erdős) и А. Реньи (A. Rényi) было показано, что среднее расстояние между двумя вершинами в случайном графе растет как логарифм от числа вершин [88, 89].



Пауль Эрдёш (1913-1996)

С именем П. Эрдёша связаны не только исследования сложных сетей, но и популярное число Эрдёша, которое используется как один из критериев определения уровня математиков в соответствующем социуме, базирующийся на так называемой сети соавторства. Известно, что Эрдёш написал около полутора тысяч статей, а также, что количество его соавторов превышало 500. Столь большое число соавторов и породило такое понятие, как число Эрдёша, которое определяется следующим образом: у самого Эрдёша это число равно нулю; у соавторов Эрдёша это число равно единице; соавторы людей с числом Эрдёша, равным единице, имеют число Эрдёша два; и так далее.

Таким образом, число Эрдёша это длина пути от некоторого автора до самого Эрдёша по совместным работам. Известен факт, что 90% математиков обладают числом Эрдёша не выше 8, что соответствует теории "малых миров", речь о которой пойдет ниже.

Некоторые сети могут оказаться несвязными, т.е. найдутся узлы, расстояние между которыми окажется бесконечным. Соответственно, средний путь может оказаться также равным бесконечности. Для учета таких случаев вводится понятие среднего инверсного пути между узлами, рассчитываемое по формуле:

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

Сети также характеризуются таким параметром как диаметр или максимальный кратчайший путь, равный максимальному значению из всех d_{ij}.

8.1.5. Коэффициент кластерности

Д. Уаттс (D. Watts) и С. Строгатц (S. Strogatz) в 1998 году определили такой параметр сетей, как коэффициент кластерности [147], который соответствует уровню связности узлов в сети. Этот коэффициент характеризует тенденцию к образованию групп взаимосвязанных узлов, так называемых клик (clique). Кроме того, для конкретного узла коэффициент кластеризации показывает, сколько ближайших соседей данного узла являются также ближайшими соседями друг для друга.

Коэффициент кластерности для отдельного узла сети определяется следующим образом. Пусть из узла выходит k ребер, которые соединяют его с k другими узлами, ближайшими соседями. Если предположить, что все ближайшие соседи соединены непосредственно друг с другом, то количество ребер между ними составляло бы $\frac{1}{2}k(k-1)$. То есть это число, которое соответствует максимально возможному количеству ребер, которыми могли бы соединяться ближайшие соседи выбранного узла. Отношение реального количества ребер, которые соединяют ближайших соседей данного узла к максимально возможному

(такому, при котором все ближайшие соседи данного узла были бы соединены непосредственно друг с другом) называется коэффициентом кластерности узла i – $C(i)$. Естественно, эта величина не превышает единицы.


Д. Уаттс и С. Строгатц

Коэффициент кластерности может определяться как для каждого узла, так и для всей сети. Соответственно, уровень кластерности всей сети определяется как нормированная по количеству узлов сумма соответствующих коэффициентов отдельных узлов. Рассмотренный ниже феномен «малых миров» непосредственно связан с уровнем кластерности сети.

8.1.6. Посредничество

Посредничество (betweenness) – это параметр, показывающий, сколько кратчайших путей проходит через узел. Эта характеристика отражает роль данного узла в установлении связей в сети. Узлы с наибольшим посредничеством играют главную роль в установлении связей между другими узлами в сети. Посредничество b_m узла m определяется по формуле:

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

где $B(i, j)$ - общее количество кратчайших путей между узлами i и j, $B(i, m, j)$ - количество кратчайших путей между узлами i и j, проходящих через узел m.

8.1.7. Эластичность сети

Свойство эластичности сетей относится к распределению расстояний между узлами при изъятии отдельных узлов. Эластичность сети зависит от ее связности, т.е. существовании путей между парами узлов. Если узел будет изъят из сети, типичная длина этих путей увеличится. Если этот процесс продолжать достаточно долго, сеть перестанет быть связной. Р. Альберт (Réka Albert) из университета штата Пенсильвания, США при исследовании атак на интернет-серверы изучала эффект изъятия узла сети, представляющей собой подмножество WWW из 326000 страниц [65].

Среднее расстояние между двумя узлами, как функция от количества изъятых узлов, почти не изменилось при случайном удалении узлов (высокая эластичность). Вместе с тем целенаправленное удаление узлов с наибольшим количеством связей приводит к разрушению сети. Таким образом, Интернет является высоко эластичной сетью по отношению к случайному отказу узла в сети, но высокочувствительной к намеренной атаке на узлы с высокими степенями связей с другими узлами.

8.1.8. Структура сообщества

О "структуре сообщества" можно говорить тогда, когда существуют группы узлов, которые имеют высокую плотность ребер между собой, при том, что плотность ребер между отдельными группами - низкая. Традиционный метод для выявления структуры сообществ - кластерный анализ. Существуют десятки приемлемых для этого методов, которые базируются на разных мерах расстояний между узлами, взвешенных путевых индексах между узлами и т.п. В частности, для больших социальных сетей наличие структуры сообществ оказалось неотъемлемым свойством.

8.2. Модель слабых связей

Некоторые свойства реальных сетей не укладываются в рамки традиционных моделей. К таким свойствам относятся и так называемые «слабые» связи. Аналогом слабых социальных связей являются, например, отношения с далекими знакомыми и коллегами. В некоторых случаях эти связи оказываются более эффективными, чем связи «сильные». Так, группой исследователей из Великобритании, США и Венгрии, был получен концептуальный вывод в области мобильной связи, заключающийся в том, что «слабые» социальные связи между индивидуумами оказываются наиболее важными для существования социальной сети [77].

Для исследования были проанализированы звонки 4.6 млн. абонентов мобильной связи, что составляет около 20% населения одной европейской страны. Это был первый случай в мировой практике, когда удалось получить и проанализировать такую большую выборку данных, относящихся к межличностной коммуникации.

В социальной сети с 4.6 млн. узлов было выявлено 7 млн. социальных связей, т.е. взаимных звонков от одного абонента другому и обратно, если обратные звонки были сделаны на протяжении 18 недель. Частота и продолжительность разговоров использовались для того, чтобы определить силу каждой социальной связи.

Было выявлено, что именно слабые социальные связи (один-два обратных звонка на протяжении 18 недель) связывают воедино большую социальную сеть. Если эти связи проигнорировать, то сеть распадется на отдельные фрагменты. Если же не учитывать сильных связей, то связность сети нарушится (рис. 33). Оказалось, что именно слабые связи являются тем феноменом, который связывает сеть в единое целое. Надо полагать, что данный вывод справедлив и для веб-пространства, хотя исследований в этой области до сих пор не проводилось.

1 2 3

Рис. 33. Структура сети:

1) полная карта сети социальных коммуникаций; 2) социальная сеть, из которой изъятые слабые связи; 3) сеть, из которой изъятые сильные связи: структура сохраняет связность

8.3. Модель малых миров

Несмотря на огромные размеры некоторых сложных сетей, во многих из них (и в WWW, в частности) существует сравнительно короткий путь между двумя любыми узлами – геодезическое расстояние. В 1967 г. психолог С. Милгран в результате проделанных масштабных экспериментов вычислил, что существует цепочка знакомств, в среднем длиной шесть, практически между двумя любыми гражданами США [113].

Д. Уаттс и С. Строгатц обнаружили феномен, характерный для многих реальных сетей, названный эффектом малых миров (Small Worlds) [146]. При исследовании этого феномена ими была предложена процедура построения наглядной модели сети, которой присущ этот феномен. Три состояния этой сети представлены на рис. 34: регулярная сеть - каждый узел которой соединен с четырьмя соседними, та же сеть, у которой некоторые «ближние» связи случайным образом заменены «далекими» (именно в этом случае возникает феномен «малых миров») и случайная сеть, в которой количество подобных замен превысило некоторый порог.

На рис. 35 приведены графики изменения средней длины пути и коэффициента кластеризации искусственной сети Д. Уаттса и С. Строгатца от вероятности установления «далеких связей» (в полулогарифмической шкале).

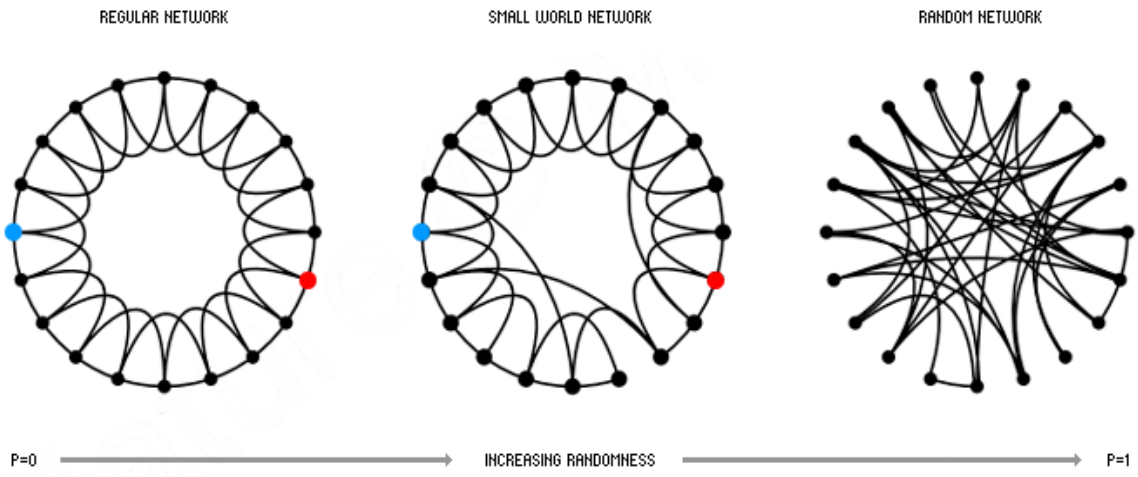


Рис. 34. Модель Уаттса-Строгатца

В реальности оказалось, что именно те сети, узлы которых имеют одновременно некоторое количество локальных и случайных «далеких» связей, демонстрируют одновременно эффект малого мира и высокий уровень кластеризации.

WWW является сетью, для которой также подтвержден феномен малых миров. Анализ топологии веб, проведенный Ши Жоу (S. Zhou) и Р. Дж. Мондрагоном (R.J. Mondragon) из Лондонского университета, показал, что узлы с большой степенью исходящих гиперссылок имеют больше связей между собой, чем с узлами с малой степенью, тогда как последние имеют больше связей с узлами с большой степенью, чем между собой. Этот феномен был назван "клубом богатых" (rich-club phenomenon). Исследование показало, что 27% всех соединений имеют место между всего 5% наибольших узлов, 60% приходится на соединение других 95% узлов с 5% наибольших и только 13% - это соединение между узлами, которые не входят в лидирующие 5%.

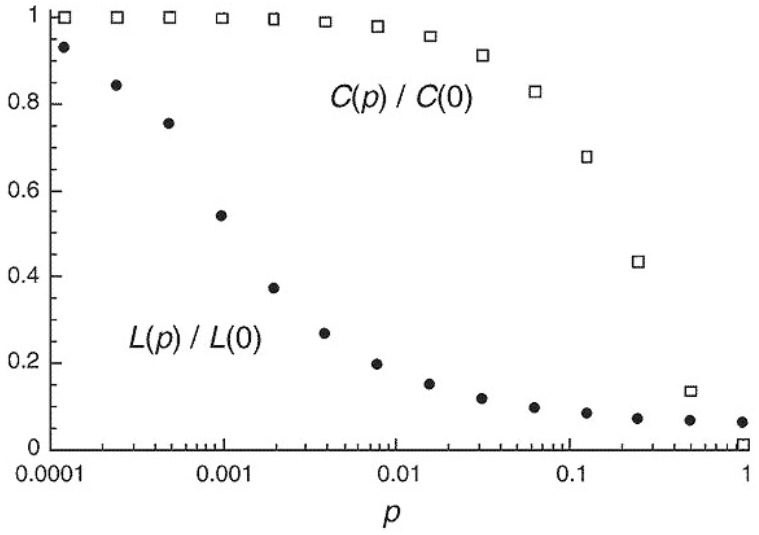


Рис. 35. Динамика изменения длины пути и коэффициента кластерности в модели Уаттса-Строгатца в полулогарифмической шкале (ось 0X – вероятность замены ближних связей далекими)

Эти исследования дают основания полагать, что зависимость WWW от больших узлов значительно существеннее, чем предполагалось ранее, т.е. она еще более чувствительна к злонамеренным атакам. С концепцией «малых миров» связан также практический подход, называемый «сетевой мобилизацией», которая реализуется над структурой «малых миров». В частности, скорость распространения информации благодаря эффекту «малых миров» в реальных сетях возрастает на порядки по сравнению со случайными сетями, ведь большинство пар узлов реальных сетей соединены короткими путями.

Кроме того, сегодня довольно успешно изучаются масштабируемые, статические, иерархические "малые миры" и другие сети, исследуются их фундаментальные свойства, такие, как стойкость к деформациям и перколяция. Недавно было показано, что наибольшую информационную проводимость имеет особый класс сетей, называемых "запутанными" (entangled networks). Они характеризуются максимальной однородностью, минимальным расстоянием между любыми двумя узлами и очень узким спектром основных статистических параметров. Считается, что запутанные сети могут найти широкое применение в области информационных технологий, в частности, в новых поколениях веб, позволяя существенным образом снизить объемы сетевого трафика.

8.4. WWW как сложная сеть

8.4.1. Топология WWW

Следует отметить, что как вся информационная сеть WWW, так и ее отдельные фрагменты и даже сайты несут значительную социальную нагрузку, которая позволяет сравнивать их на содержательному уровне с социальными сетями, образованными отношениями людей или цитированием в науке. Веб, будучи, наверное, самой динамической частью информационного пространства, характеризуется большим количеством скрытых в нем неявных экспертных оценок, реализованных в виде гиперссылок. Поэтому WWW можно с полным правом считать социальной сетью, исследование которой можно проводить, базируясь на существующем подходе анализа таких сетей - SNA. Много сетевых служб, которые позволяют людям устанавливать связи в Сети, автоматически формируют социальные сети. Кроме того, сегодня бурно развился специальный сервис по целенаправленному построению социальных сетей в веб-пространстве.

В 1999 г. А. Брёдер (A. Broder) из IBM и его соавторы из компаний AltaVista, IBM и Compaq математически описали «карту» ресурсов и гиперсвязей веб-пространства [83], получившей благодаря своей форме название «галстука-бабочки» (Bow Tie, рис. 36). С помощью баз данных и поискового механизма AltaVista было проанализировано свыше 200 млн. веб-страниц и несколько миллиардов ссылок, размещенных на этих страницах.

Андрей Брёдер

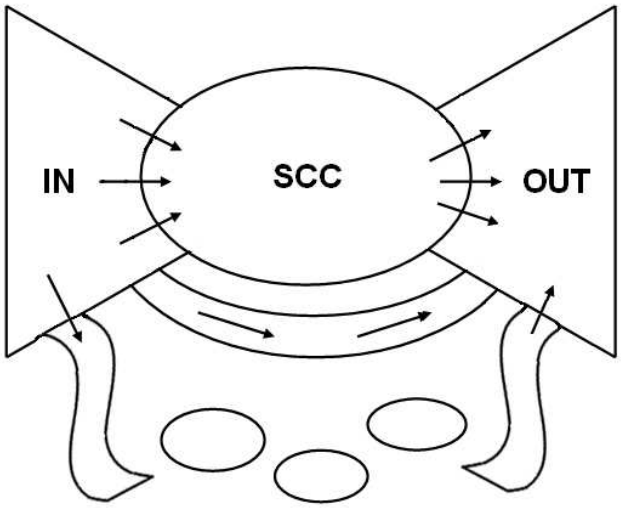


Рис. 36. Модель веб-пространства Bow Tie

В рамках общей задачи определения структуры связей между отдельными веб-страницами было выявлено:

- центральное ядро (28% веб-страниц) - область сильной связности (Strongly Connected Component, SCC), которая образована веб-страницами, связанными между собой так тесно, что следуя по гиперссылкам, из любой из них в конечном счете можно попасть на любую другую не выходя из этой области;
- 22% веб-страниц - это "отправные веб-страницы" (IN). Они содержат гиперссылки, которые в конечном итоге ведут к ядру, но из ядра к ним попасть нельзя;
- столько же - 22% - "конечных веб-страниц" (OUT), к которым можно прийти по ссылкам из ядра, но нельзя возвратиться назад;
- 22% веб-страниц - отростки - полностью изолированные от центрального ядра: это или "мысы", связанные гиперссылками со страницами любой другой категории, или "перешейки", соединяющие веб-страницы, которые не входят в ядро.

Выявлено, что четыре основных множества - более 90% веб-страниц, топологически относятся к одной компоненте связности. Существуют и "острова", которые вообще не пересекаются с остальными ресурсами Интернет.

Единственный способ обнаружить ресурсы этой группы - знать адрес. Никакие поисковые машины не смогут найти эти острова, если они в прошлом каким-то образом не соединялись с другими частями Интернет.

Было обнаружено, что пропорции этих четырех категорий в течение нескольких месяцев оставались неизменными, несмотря на значительное увеличение общего объема веб-ресурсов. Топология и характеристики модели оказались примерно одинаковыми для различных подмножеств веб-пространства, подтверждая тем самым наблюдение о том, что свойства структуры всего веб-пространства Bow Tie также верны и для его отдельных подмножеств. Таким образом, алгоритмы, использующие информацию о структуре веб-пространства, предположительно будут работать и на отдельных его подмножествах [78].

Были исследованы такие параметры модели Bow Tie, как среднее количество сайтов, через которые связываются любые два сайта гиперссылками, а также распределение входящих и исходящих ссылок. Оказалось, что распределение степеней узлов (входящих и исходящих гиперссылок) веб-пространства (исследовались сайты домена edu в количестве 325729) подчиняется степенному закону, т.е. вероятность того, что соответствующая степень вершины равна i, пропорциональна $1/i^k$ (для входящих ссылок $k \approx 2.1$, а для исходящих $k \approx 2.45$). Кроме того, оказалось, что сеть WWW является «малым миром» со средней длиной кратчайшего пути, равной 11 и относительно большим значением коэффициента кластерности, приблизительно равным 0.15 (для классического случайного графа это значение составило бы 0.0002) .

С большой вероятностью случайно выбранные веб-страницы вообще никак не связаны, если же путь все-таки существует, среднее количество кликов, необходимых для переходов между такими страницами, составляет 16. Если же этот путь рассматривать как двусторонний, то среднее число промежуточных кликов сокращается до семи. Благодаря полученным результатам, уже сегодня может быть создан инструментарий, способный превратить веб-пространство в систему двустороннего движения. "Сейчас трафик по существу односторонний. Если бы браузер был наделен средствами серфинга в обратном направлении, это

открыло бы доступ к гораздо большему числу ресурсов", - заявил по этому поводу представитель IBM Нам Ламор (N. Lamour) [32].

Это свойство структуры веб-пространства сегодня уже довольно широко используется при решении многих задач, например, для оптимизации эффективности механизмов сканирования, при построении новых веб-сервисов, для решения задач анализа и прогноза.

В то же время существует совокупность веб-ресурсов, не видимая «глобальными» поисковыми системами, называемая «скрытым веб». К таким ресурсам, в частности, относятся некоторые динамически формируемые веб-страницы и документы из баз данных. В этой связи необходимо подчеркнуть некоторую некорректность расчета объемов «островов» по Брёдеру из-за того, что список веб-ресурсов был получен из базы данных системы AltaVista, полученный в результате работы программы-робота, сканирующего веб-ресурсы, переходя от одного к другому по гиперссылкам. В настоящее время широкое распространение получили каталоги «скрытого» веб. Также осуществляются попытки доступа к объектам «скрытого» веб через специализированные системы поиска.

Л. Бйорнеборном (L. Björneborn) была предложена модель «мятого веб», которая ассоциируется с мятой бумагой. При этом путь между выбранными точками на мятой бумаге чаще всего короче, так как противоположные части листа бумаги соединены вместе. В соответствии с этой моделью каждая новая гиперссылка изменяет все существующие связи, создавая новые деформации «мятой» сети. Т.е. каждая новая гиперсвязь - «крючок», который растягивает или деформирует форму существующей сети WWW.

8.4.2. Сетевая структура новостного веб

Новостной веб представляет собой фрагмент веб-пространства, к которому можно отнести сайты информационных агентств, онлайновых СМИ, новостные разделы сайтов государственных учреждений и т.п.

В отличие от существующих моделей веб-пространства, при анализе сети, образованной новостными ресурсами, необходимо было учитывать чрезвычайно

высокую динамику информационных потоков, контекстные (не только гипертекстовые) ссылки, эффект содержательного дублирования. Кроме того, применение модели А. Бредера к новостной составляющей веб-пространства, по-видимому, нельзя считать корректным по ряду причин:

- на наиболее актуальные сообщения на протяжении определенного времени ссылок может вообще не существовать;
- модель Бредера не учитывает особенности «скрытого» веб, т.е. тех информационных веб-ресурсов, на которые не существует прямых гиперссылок;
- в новостных потоках необходимо учитывать не только гиперссылки, но и ссылки контекстные, причем не только на объекты из открытой части веб-пространства (это могут быть даже оффлайновые публикации из изданий, возможно и присутствующих в сети Интернет);
- модель Бредера не учитывает такого понятия, как содержательное дублирование информации;
- при построении модели структуры новостного веб наибольшее внимание должно уделяться именно веб-сайтам, на которых публикуются новостные сообщения, а не отдельным веб-страницам или самим сообщениям.

В качестве экспериментальной базы для построения модели новостного веб-пространства использовался информационный корпус системы InfoStream [31], обеспечивающей автоматизированный сбор информации с открытых веб-сайтов. Для построения модели для каждого из 2500 источников, охватываемых системой, был составлен запрос следующего вида:

<код источника><шаблоны для поиска>.

Совокупность подобных запросов была объединена в пакет, в результате специальной обработки которого для каждого сообщения, относящегося к определенному источнику - веб-сайту, были выявлены исходящие ссылки на другие источники (ссылки на собственный источник исключались). Было выявлено, что исходящие контекстные ссылки присутствовали на 484945 сообщениях с 2323 веб-сайтов.

Также было получено распределение новостных источников по количеству веб-сайтов, имеющих на них ссылки. Всего за месяц ссылки указывали на 1459 источников (без самоцитирования). Оказалось, что на 100 источников ведет свыше 80% ссылок.

Ниже приведен начальный фрагмент ранжированного списка новостных источников, на которые ведут ссылки с максимального количества веб-сайтов:

Web-сайт	**Количество ссылающихся веб-сайтов**
ИА «Интерфакс»	1051
«РосБизнесКонсалтинг»	983
"Reuters"	882
ИТАР-ТАСС	787
РИА «Новости»	773
УНИАН	675
Радио «Свобода»	662
НТВ	631
«Коммерсантъ»	623
BBC	598
«Комсомольская правда»	595

Следует отметить, что оценка уровня источника информации как «автора» преимущественно по количеству веб-сайтов, с которых на него ведут гиперссылки, вполне согласуется с предложенным Лемпелем и Мораном алгоритмом Salsa [108].

Интересным оказался график двумерного сечения значений $\log(N_{OUT}+1)$, $\log(N_{IN}+1)$, где N_{OUT} - количество входящих ссылок, N_{IN} - количество исходящих ссылок для каждого из источников (рис. 37). Этот график послужил основой идеальной схемы представления областей модели в зависимости от количества исходящих и входящих ссылок (рис. 38).

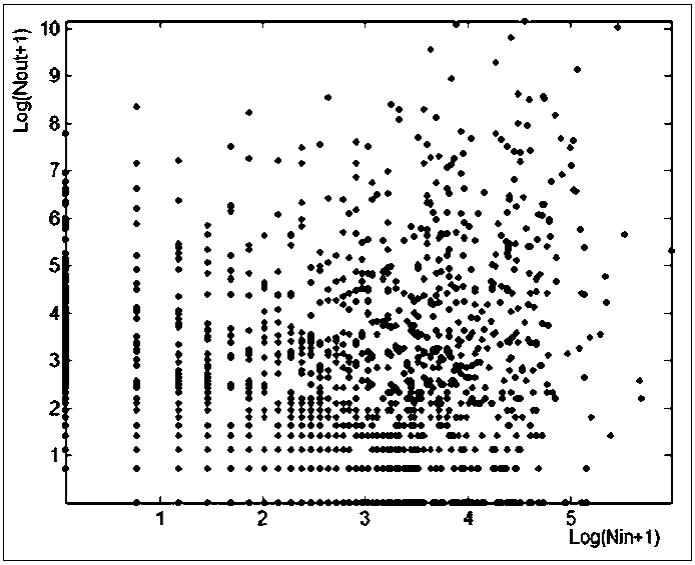


Рис. 37. График распределения зоны ядра в координатах «логарифм количества исходящих сообщений – логарифм количества входящих сообщений»

Рис. 38. Представление областей модели в зависимости от количества исходящих и входящих ссылок

В результате проведенных исследований была построена модель новостного веб-пространства, основанная на контекстных ссылках. Также предложены подходы к выявлению основных зон модели новостного веб-пространства и рассчитаны числовые соотношения различных зон модели.

Разработанная модель новостного веб позволила выявить те же основные подсети, что и в модели веб-пространства, но процентные соотношения между ними оказались различными.

8.5. Визуализация сложных сетей

Одним из направлений анализа сложных сетей является их визуализация. Визуализация имеет большое значение, поскольку чаще всего позволяет получать важную информацию о характере взаимодействия узлов, не прибегая к точным методам анализа. При отображении модели сложной сети можно считать целесообразным:

- размещение узлов сети в двух измерениях;
- пространственное приведение в порядок объектов в одном измерении в соответствии с некоторым количественным свойством;
- использование общих для всех сетевых диаграмм методов отображения количественных и качественных свойств объектов и отношений.

Так, например, TouchGraph Amazon отображает сеть, образованную книгами и связями между ними (по тематикам, авторам, издательствам). Одним из самых динамических новостных ресурсов Интернет сегодня можно считать также блоги. Компания TouchGraph, в частности, реализовала интерфейс для построения социальной сети на основе Livejournal - TouchGraph LiveJournal Browser.

В случае визуализации WWW средствами TouchGraph Google Browser (http://www.touchgraph.com/TGGoogleBrowser.html) узлами выступают веб-сайты, а ребрами - не гиперссылки, а отношения содержательного подобия. Google Browser представляет собой Java-апплет, позволяющий визуализировать связи подобия между веб-сайтами, которые рассчитываются в поисковой системе Google. В этом интерфейсе можно увидеть все сайты, связанные отношением подобия (реализованного в Google) с исходным заданным, при этом пользователь может задавать глубину связей и отображать взаимосвязи разных сайтов. TouchGraph Google Browser довольно полезный инструмент также при поиске сайтов, связанных с исходным общей тематикой (рис. 39).

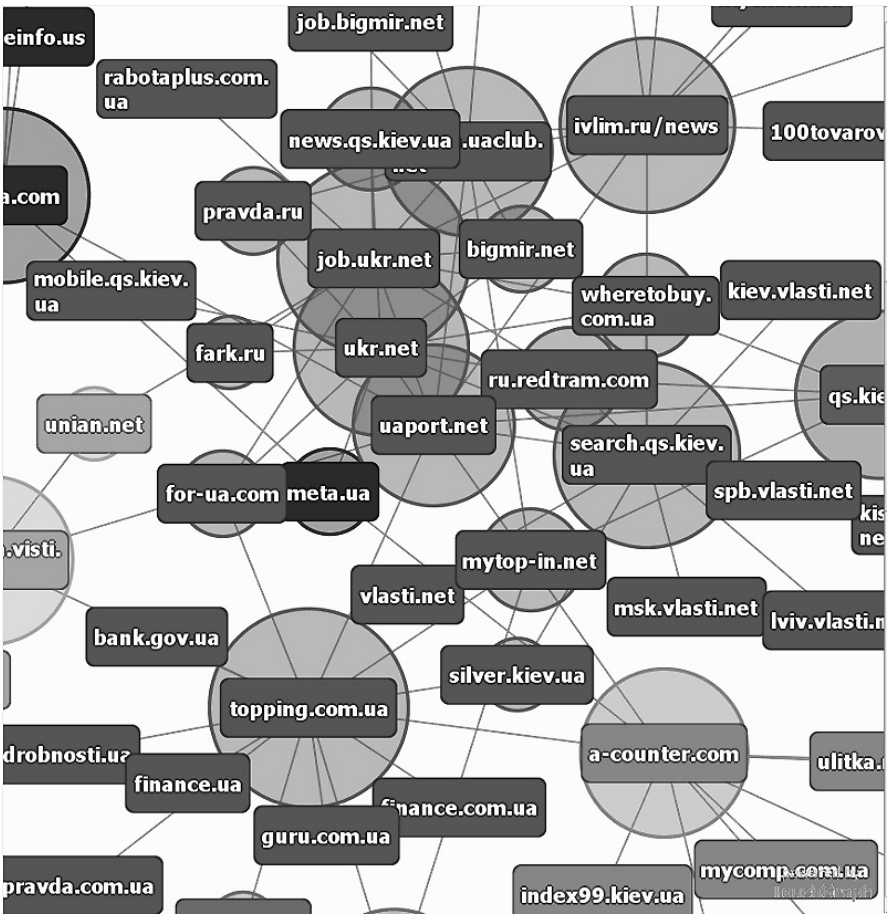


Рис. 39. Отображение связей веб-серверов (от TouchGraph)

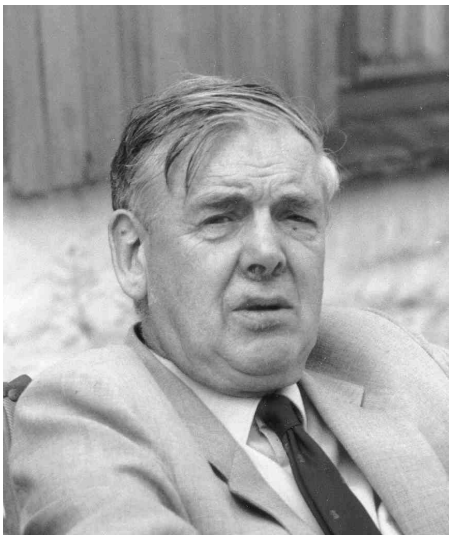
В качестве еще одного инструмента для анализа и визуализации сложных сетей можно привести программу NetVis (http://www.netvis.org), которая использует online-данные и импортированные файлы. Также широко известны программы визуализации и анализа социальных/организационных сетей InFlow (текущая версия 3.1 доступна по адресу http://www.orgnet.com/ inflow3.html) и система анализа социальных сетей UCINET (http://www.analytictech.com/ ucinet/ucinet.htm) с интегрированной в ней свободно распространяемой программой визуализации NetDraw.

9. ЭЛЕМЕНТЫ ТЕОРИИ ПЕРКОЛЯЦИИ

«- Подожди-ка Хуанита. Ты уж реши. Эта «Лавина», она что: наркотик, вирус или религия?
Хуанита пожала плечами. - А что есть разница?»
Нил Стивенсон

9.1. Задача теории перколяции

Одной из важных характеристик сложных сетей является возможность протекания по их ребрам тока, жидкости, информации (трафика) и т.п. Впервые задача перколяции (от англ. percolation – просачивание, протекание) была сформулирована в 1957 г. в работе С.Р. Бродбента (S.R. Broadbent) и Дж.М. Хаммерсли (J.M. Hammersley) [81]. В последствии была развита целая область исследований (в теории графов, теории вероятностей, физике, химии...), названная теорией перколяции, имеющая многочисленные применения на практике. Оказывается, что многие вопросы, которые возникают при анализе сетевой безопасности в Интернет, также непосредственно относятся к теории перколяции [132].



Дж.М. Хаммерсли (1920-2004)

Перед теорией перколяции стоят многие вопросы, которые выходят за стандартные рамки дискретной математики и теории вероятностей [48]. Самая простая формулировка задачи теории перколяций следующая. Дана решетка из

связей, случайная часть которых $(p)-$ «черная» - проводящая, а остальная – «белая», не проводящая поток. Необходимо найти такую минимальную концентрацию p_c «черных» связей, при которой еще есть связный путь по «черным» связям сквозь всю решетку. То есть такую концентрацию, когда решетка в целом проводит.

При $p=0$ все связи решетки «белые» - решетка не проводит. При увеличении концентрации «черных» - проводящих связей, при $p=p_c$ в решетке возникает перколяционный, проникающий, кластер из «черных» связей, соединяющий противоположные края сетки. При размере сетки, стремящемся к бесконечности, размер этого кластера также бесконечен, в связи с чем и был введен термин «бесконечный кластер». Другие совокупности соединенных между собой связей конечного размера называются конечными кластерами.

При переходе через порог протекания, т.е. при возникновении бесконечного кластера, свойства системы, характеризующие ее в целом, резко меняются. Если, например, «черная» связь проводит ток, а «белая» нет, то проводимость всей системы σ вблизи p_c имеет вид, представленный на рис. 39.

Сопротивление же всей системы $R\sim 1/G$ резко падает. Учитывая логарифмический масштаб на рис. 40, ясно, что вблизи порога протекания пропускная способность сети может при очень небольшом уменьшении p резко упасть.

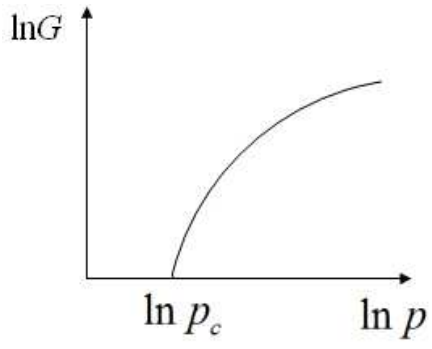


Рис. 40. Проводимость системы вблизи порога протекания (в логарифмическом масштабе)

Решетка может быть упорядоченной - квадратной, треугольной, кубической… Она может быть и случайной, со случайным числом связей,

приходящихся на данный узел. Естественно, что для различных случаев величина порога протекания будет разной. Удивительным оказалось то, что задача определения p_c, казалось бы простая задача, «не поддалась» точным методам теории вероятностей. За редкими исключениями p_c не удается вычислить аналитически – необходимо довольно сложное численное моделирование.

В настоящее время известно много важных обобщений перколяционной задачи, например, рассматриваются случаи, когда «непроводящие» связи проводят, но намного хуже проводящих ведущих; можно говорить о разных значениях проводимости для разных связей; можно рассматривать однонаправленные «диодные связи» и т.п.

К задачам, решаемым в рамках теории перколяции и анализа сложных сетей относятся такие, как определение предельного уровня проводимости (пропускной способности), изменения длины пути и его траектории (извилистости, параллельности) при приближении к предельному уровню проводимости, количества узлов, которые необходимо удалить, чтобы нарушить связанность сети.

9.2. Характеристики перколяционных сетей

Существуют важные характеристики «черных» связей, по которым происходит прохождение тока, жидкости, информации (имеющие одинаковый характер поведения).

Приведем некоторые из них:

$P_\infty(p)$ – вероятность того, что случайным образом выбранный узел принадлежит бесконечному кластеру, соединяющему две противоположные стороны сети – эта характеристика имеет смысл выше порога протекания, когда существует бесконечный кластер;

$S(p) = \dfrac{\sum_s s^2 n_s(p)}{\sum_s s\, n_s(p)}$ – средний размер кластера (от англ. mean cluster size), где n_s - число кластеров из s узлов, приходящихся на один узел решетки – эта

характеристика имеет смысл ниже порога протекания, когда все существующие в системе кластеры имеют конечный размер;

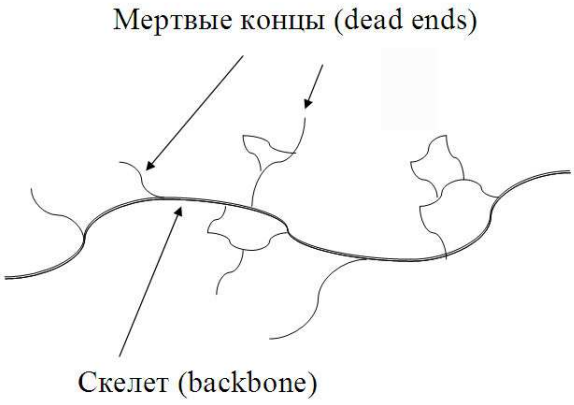
$\xi(p)$ - корреляционная длина, характеризующая быстроту спадания корреляций в решетке. Чтобы ввести ее необходимо определить парную корреляционную функцию $G(r, p \approx p_c) \equiv G\left(\left|r_i - r_j\right|, p\right) = \left\langle g\left(r_i, r_j\right)\right\rangle$, где функция $g\left(r_i, r_j\right)$от радиусов векторов узлов i и j равна единице, если узлы связаны «черными» связями, принадлежащими одному конечному кластеру, и равной нулю во всех остальных случаях, а угловые скобки означают усреднение по всем узлам. При $r \to \infty$ корреляционная функция $G(r, p)$ убывает экспоненциально, характерным масштабом при этом, как раз и является корреляционная длина $G(r, p) \sim \exp\left(-r / \xi(p)\right)$.

Отметим, что при $p \to p_c$ корреляционная длина $\xi(p)$ расходится – и это полностью соответствует качественным представлениям. Чем ближе к порогу протекания ($p \geq p_c$), тем меньше «черных» проводящих связей, тем существеннее каждая часть кластера, соединяющая две «бесконечности», обрыв одной из них может сказаться далеко от нее. На самом пороге протекания одна оборванная связь может разрушить весь проводящий путь, называемый бесконечным кластером. Структура бесконечного кластера не простая, на самом пороге протекания он является фрактальным объектом, состоящим из связей, входящих в этот кластер, с размерностью $d_f(d = 2) = d - \beta / \nu \approx 1.896,\ d_f(d = 3) \approx 2.54.$

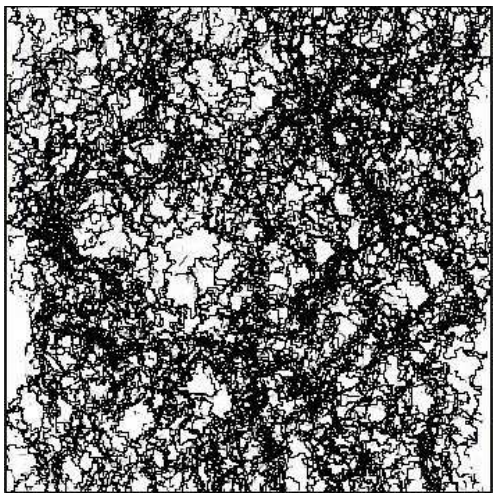
Вводят также и другие характеристики бесконечного кластера (рис. 41) – остов кластера (от англ. backbone) – та часть кластера, по которой происходит перенос, а также отсеченные «мертвые концы» (от англ. dead ends), и многие другие.

В более развитой и реалистичной теории протекания и «белые» связи считаются проводящими, с проводимостью много меньшей «черных» (принято говорить, что в $1/h$ раз меньше, где $0 < h << 1$). При введении параметра h теория протекания может быть сформулирована в терминах теории фазовых

переходов II рода, одним из самых сложных разделов теоретической физики. В этой теории вводят так называемую близость к точке фазового перехода τ.



а)



б)

Рис. 41. Бесконечный кластер: а) – часть бесконечного кластера; б) – распределение проводящих связей (черный цвет)

Если речь идет, например, о переходе магнита из ферромагнитного в парамагнитное состояние, то $\tau = (T - T_c)/T_c$, где T – температура, а T_c – критическая температура фазового перехода, в данном случае - температура Кюри. Параметр порядка η – это величина, существенно изменяемая при прохождении T_c, которая характеризует свойства системы в целом (в рассматриваемом случае – это намагниченность). Вблизи T_c этот параметр ведет себя степенным образом (рис. 41):

$$\eta \sim h\,|\tau|^{-\gamma},\ \tau > 0,\ (T > T_c),\ \gamma > 0;$$
$$\eta \sim |\tau|^{\beta},\ \tau < 0,\ (T < T_c),\ \beta > 0,$$

где γ, β – так называемые критические индексы, а h в данном случае безразмерное внешнее магнитное поле $(h << 1)$. Наличие степенной зависимости у параметра порядка η вблизи критической точки означает наличие нетривиальных корреляций в системе.

$$\eta \sim h\,|\tau|^{-\gamma},\ \tau > 0,\ (T > T_c),\ \gamma > 0;$$

$$\eta \sim |\tau|^{\beta},\ \tau < 0,\ (T < T_c),\ \beta > 0.$$

Внешнее поле h «размазывает» фазовый переход. В том случае, когда $h = 0,$ параметр порядка выше температуры фазового перехода T_c равен нулю. На рис. 42, в этом случае зависимость η от температуры при увеличении до T_c падает в ноль. Существование небольшого значения параметра порядка выше T_c (размытие) связано с существованием ненулевого значения h.

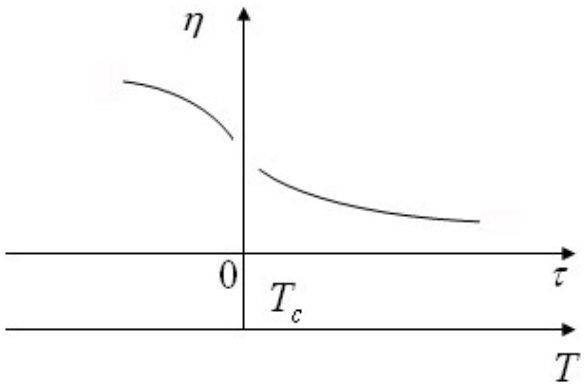


Рис. 42. Зависимость намагниченности от температуры

Существует характерное значение $h_\tau,$ зависящее от близости к точке перехода $T_c,$ такое, что при слабых полях $(h << h_\tau)$ его влиянием на поведение параметра порядка можно пренебречь. И наоборот, в сильных полях (при $h >> h_\tau$) влияние внешнего поля становится определяющим. При произвольных τ и h можно сформулировать так называемое скейлинговое соотношение, которое включает в себя все частные случаи:

$$\eta = h^{\frac{1}{\gamma}} f(\tau / h^{\frac{1}{\beta\delta}}), \qquad \tau = (T - T_c)/T_c$$

Скейлинговая функция $f(z)$ зависящая от одной переменной $z = \tau / h^{\frac{1}{\beta\delta}}$ ведет себя следующим образом:

$$f(z) \sim \begin{cases} z^{-\gamma}, & z \to +\infty, \\ const, & z \to 0, \\ |z|^{\beta}, & z \to -\infty. \end{cases}$$

Описание проводимости перколяционной сети также может быть сформулировано в терминах теории фазовых переходов II рода. Для

определенности будем говорить о так называемой эффективной проводимости σ_e, которая характеризует проводимость системы в целом. Роль параметра порядка в этом случае играет σ_e, а близость к критической точке $\tau = (p - p_c)/p_c$, где p_c – порог протекания. С точностью до обозначений скейлинговое соотношение для эффективной проводимости σ_e совпадает с соответствующим соотношением для параметра порядка η:

$$\sigma_e = h^{\frac{t}{\varphi}} f(\tau / h^{\frac{1}{\varphi}}),$$

$$f(z) \sim \begin{cases} z^t, & z \to +\infty, \\ const, & z \to 0, \\ |z|^{-q}, & z \to -\infty. \end{cases}$$

В этих формулах t и q - так называемые критические индексы проводимости, $\varphi = t + q$.

При этом необходимо учесть, что в силу «исторических» причин оси p - концентрации и T - температуры направлены в противоположные стороны поэтому с увеличением p эффективная проводимость растет (рис. 43).

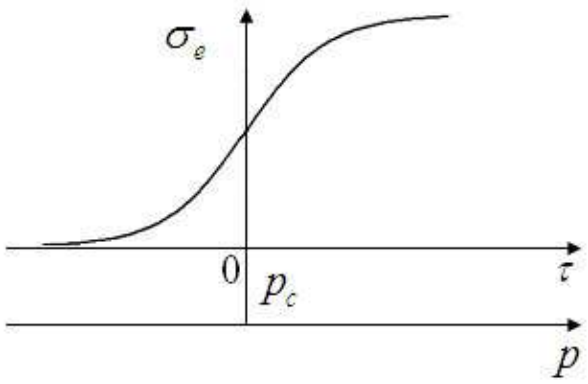


Рис. 43. Эффективная проводимость

Необходимо уточнить, что универсальное скейлинговое поведение σ_e имеет место только вблизи порога протекания, т.е. при $|\tau| << 1$.

9.3. Сеть с экспоненциально широким распределением

Имеется множество обобщений задачи протекания. В описанном выше, первом и простейшем варианте речь идет о двухфазной системе – существуют две фазы, два типа связей, узлов. «Черные» - хорошо проводящие и «белые» - плохо. Функция распределения при этом представляет собой две дельта-функции Дирака $f(\sigma) = p\delta(\sigma_1 - \sigma) + (1-p)\delta(\sigma_2 - \sigma)$. Возможен случай непрерывного распределения величины проводимости связей (рис. 44). Как и в двухфазной задаче, здесь существует малый параметр $\sigma_{\min} / \sigma_{\max} << 1$. Именно при его наличии имеет место универсальное поведение зависимости параметра порядка в такой сети.

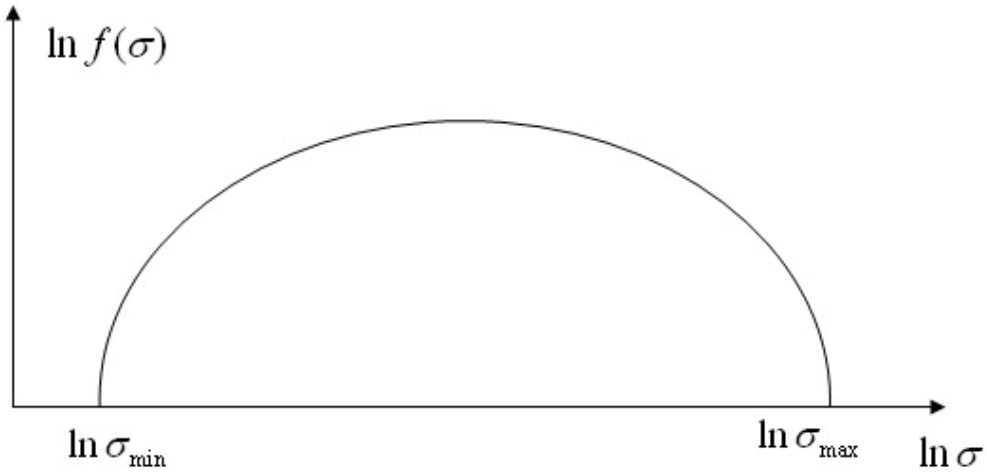


Рис. 44. Двухфазная перколяционная система

Одним из примеров непрерывного распределения проводимости связей - σ (или их сопротивления $r \sim 1/\sigma$) является экспоненциальное распределение (в частности, оно реализуется в так называемой высокотемпературной прыжковой проводимости в легированных полупроводниках):

$$r = r_0 e^{-\lambda x}, \quad \lambda >> 1,$$

где r - сопротивление данной связи, $x \in (0,1)$ - случайная переменная, имеющая гладкую функцию распределения - $f(x)$. Естественно, при таком определении случайной переменной x имеем $r_{\min} = r_0 \exp(-\lambda)$ и $r_{\max} = r_0$.

При $\lambda >> 1$ справедливо: $r_{\max} >> r_{\min}$ и $\sigma_{\max} >> \sigma_{\min}$. Сетка, узлы которой соединены между собой связями, с распределением сопротивлений $r = r_0 e^{-\lambda x}$ называется сеткой Абрахамса-Миллера.

На первый взгляд задача об определении σ_e сетки с экспоненциально широким спектром сопротивлений не является перколяционной – нет порога протекания, при достижении которого одна из фаз образует бесконечный кластер, так как нет самих фаз. Существует, однако, способ, который позволяет в определенном приближении свести такую задачу к стандартной перколяционной. Чтобы показать, как работает этот прием, проведем следующий мысленный эксперимент. Извлечем из сетки Миллера–Абрахамса все сопротивления, запомнив их положения, и начнем последовательно вставлять их на свои места, начиная процесс с самых минимальных ($r_{\min}$). Все включенные на данный момент сопротивления будем считать хорошо проводящими. При достижении хорошо проводящими связями пороговой концентрации в системе появится бесконечный кластер – последнее включенное сопротивление $r_c = r_0 \exp(-\lambda x_c)$ замкнет мостик – кластер из хорошо проводящих ($r < r_c$) связей. Поскольку при экспоненциально широком спектре сопротивлений каждое следующее включенное сопротивление намного больше предыдущего, сопротивление образовавшегося бесконечного кластера практически, с точностью до предэкспоненциального множителя, определяется последним включенным сопротивлением. Добавка следующих, шунтирующих, сопротивлений уже ничего не изменяет, так как их значения намного больше критического r_c.

Критическое сопротивление r_c легко определить из следующего условия: оно «появляется» при пороговой концентрации:

$$\int_{x_c}^{1} f(x)\,dx = p_c .$$

Для простейшего случая равномерного распределения, когда $f(x) = 1, \quad x \in (0,1)$ отсюда следует $x_c = 1 - p_c$. Таким образом, имеем:

$$r_c = r_0 e^{-\lambda x_c} = r_0 e^{-\lambda(1-p_c)},$$

или в терминах эффективной проводимости:

$$\sigma_e = \sigma_0 e^{\lambda x_c} = \sigma_0 e^{\lambda(1-p_c)}.$$

Таким образом, если речь идет о проводимости (пропускной способности) сети с экспоненциально широким распределением проводимости связей, то практически все определяется одной единственной связью, той, которая замыкает кластер (в рассмотренном эксперименте), назовем ее пороговой. Те связи, которые были включены «ранее» пороговой проводят много лучше и не лимитируют проводимость всей сети. Включенные же «позже» проводят много хуже и их вкладом в проводимость сети в целом можно пренебречь. В определенном смысле проводимость устойчива относительно повреждений. Вырывание практически любой связи (кроме пороговой) ни к чему не приводит. Вырывание же пороговой связи приводит к экспоненциальному уменьшению проводимости сети, при этом, естественно, появляется новая пороговая связь с экспоненциально более высоким сопротивлением.

9.4. Диодные перколяционные сети

Расположение диодных связей (направление их пропускания) в различных вариантах диодной перколяции может быть различным, например, как в случае с сетью WWW - почти хаотичным. На рис. 45 показан случай такой двумерной диодной перколяции – так называемый случай «полностью направленной» (от англ. fully directed) перколяции. Здесь направление пропускания связей направлено вдоль положительного направления. Стрелки на сетке показывают направление пропускания диодных связей.

Перколяционный кластер при «полностью направленной» перколяции уже не изотропен в среднем. Качественно, на больших размерах, его можно изобразить так как, это показано на рис. 46 – в виде расширяющегося вдоль OX потока. Ось OX соответствует направлению среднего протекания, а $\xi_{\parallel} \neq \xi_{\perp}$ корреляционные длины, соответственно, вдоль и поперек направления протекания.

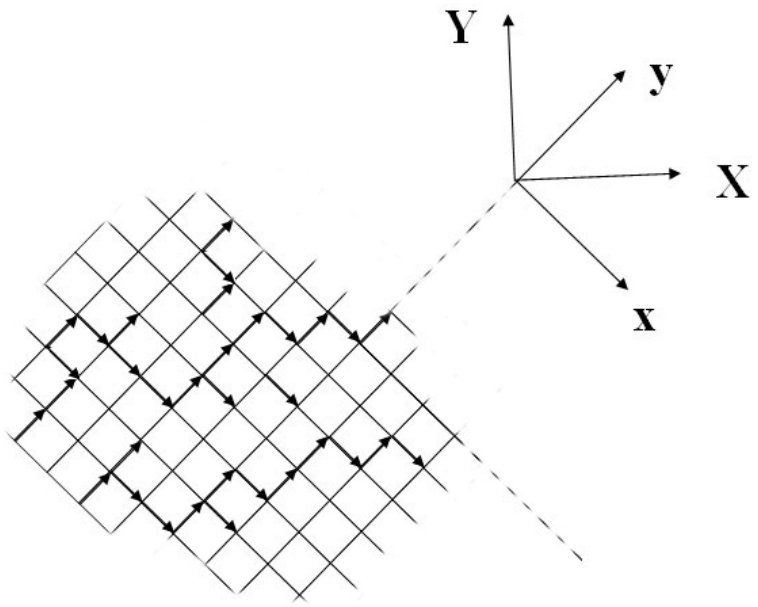


Рис. 45. «Полностью направленная» перколяция

Задача «полностью направленной» перколяции применяется для пространственно-временного описания распространения сигналов в реальных биологических, нейронных, логических сетях. В этом случае OX – это ось времени; естественно, протекание вспять по времени в модели «полностью направленной» перколяции запрещено.

Возможны и другие диодные модели перколяции, представленные, например, на рис. 47. Здесь p_+ и p_- - вероятности появления проводящей связи с положительной и отрицательной проекцией направления проводимости вдоль OX (направления среднего протекания, направления «времени»), p - вероятность появления связи проводящей равным образом в обе стороны – т.е. обычной «черной» связи. Вероятность непроводящих связей равна: $q = 1 - p - p_+ - p_-$.

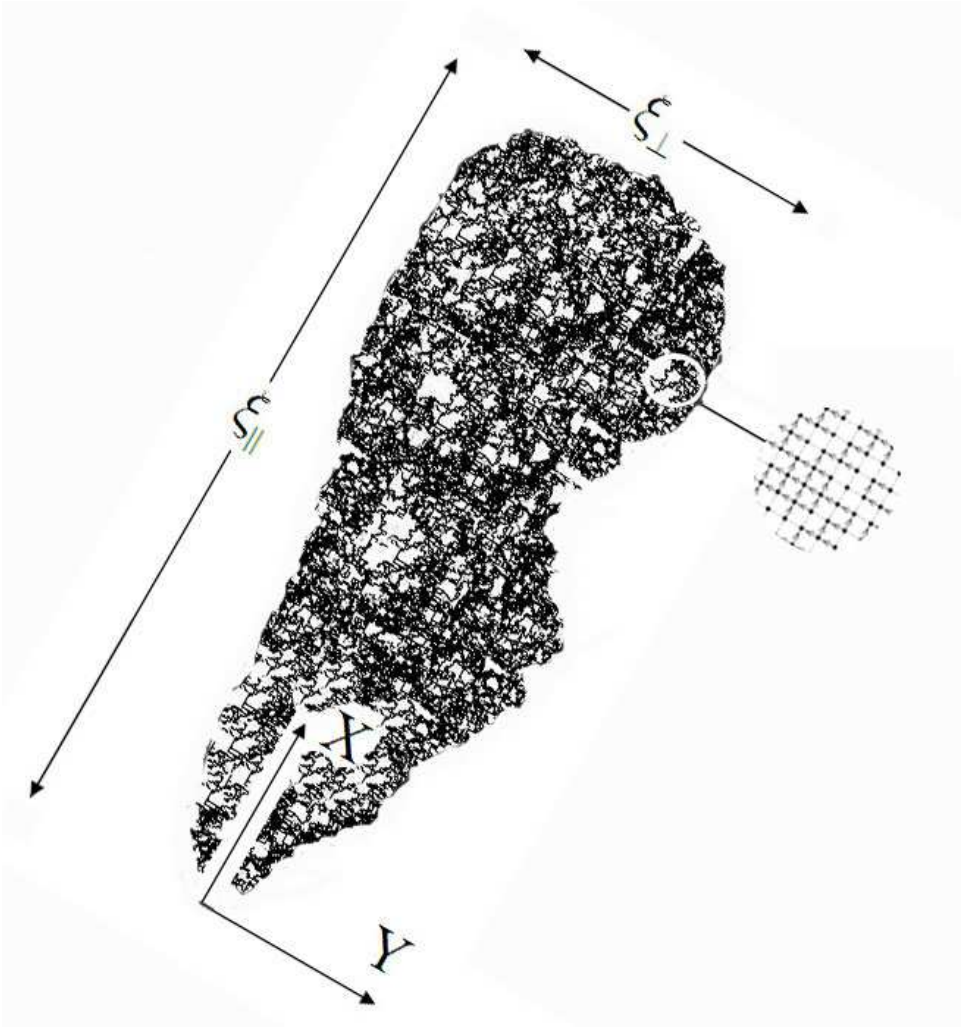


Рис. 46. Качественная картина перколяционного кластера при полностью направленной перколяции

Конфигурационное пространство такой перколяционной модели изображено на рис. 48. Здесь, например, точка p означает, что все связи «черные», q - белые, точка A- все связи в системе диодные направленные с равной вероятностью во все стороны и т.д. Линия $q-p$ - пример обычной двухфазной перколяции, движение по этой линии от q к p проходит через порог протекания p_c обычной «черно-белой» перколяции. Линия $q-p_+$ - рассмотренная выше fully directed перколяционная задача. Задачи, соответствующие линиям $q-p_+$ и $q-p_-$ отличаются только направлением оси «времени». Линия p_+-p означает направленное протекание, но в отличие от fully directed задачи, в сетке существуют обычные «черные» связи.

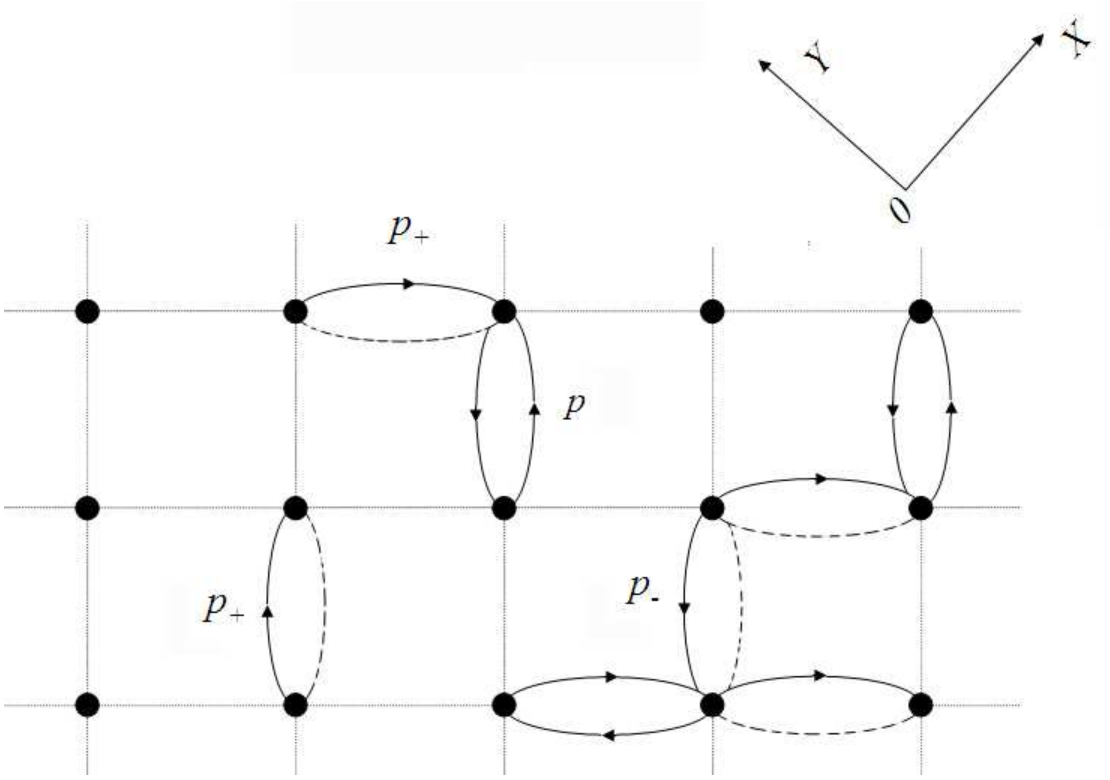


Рис. 47. Вероятностная диодная модель

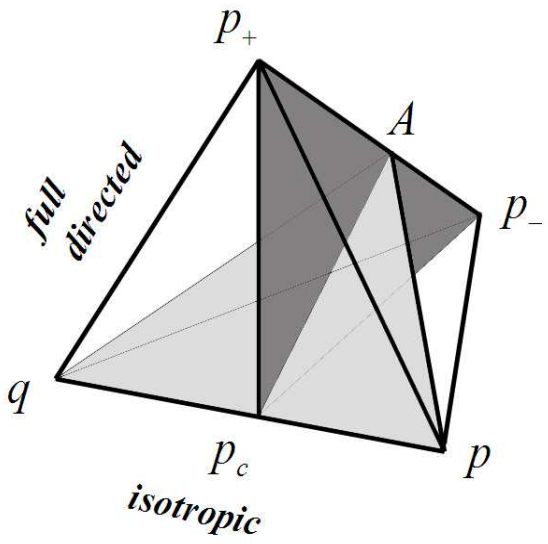


Рис. 48. Конфигурационное пространство диодной перколяционной задачи

9.5. Перколяция на случайных сетях

Перколяцию можно рассматривать и на случайных сетях Эрдоша-Реньи, Уаттса-Строгатса, и т.п. В этом случае вместо бесконечного перколяционного

кластера принято говорить о гигантской связной компоненте (giant connected component).

Для случайной сети (графа) Эрдоша-Реньи из N узлов было показано, что порог протекания $p_c \approx 1/N$, т.е. что протекание наступает когда средняя степень узла $\langle k \rangle \geq 1$.

М. Ньюманом (M. Newman) и Д. Уаттсом [117] была рассмотрена задача перколяции на сетях малого мира. При этом была использована модификация сети малого мира. В отличие от того как малый мир вводился в [146] новые связи, далее называемые shortcuts, набрасывались в начальную сеть дополнительно, все старые связи между соседями не прерывались, оставаясь на своих местах. Как и в стандартной теории протекания может быть введена характерная длина (в теории протекания она называется корреляционной длиной):

$$\xi = \frac{1}{(\varphi k d)^{1/d}},$$

где φ - вероятность встретить переброшенную связь (shortcuts), φN - число этих связей, N - число узлов сети, k - число соседей, d - мерность сети.

Смысл этой длины простой – это характерное расстояние между концами разных наброшенных связей (shortcuts).

Для кратчайшего пути l (среднего, конечно) существует скейлиговая функция:

$$l = \frac{N}{k} f\left(\frac{N}{\xi}\right),$$

$$f(z \ll 1) \to const, \qquad f(z \gg 1) \to \frac{\log z}{z}.$$

Так как $\xi \sim 1/(\varphi k)^{1/d}$, асимптотика скейлинговой функции означает при $z = N/\xi \ll 1$, что $N\varphi^{1/d} \ll 1$, т.е. что число переброшенных связей (shortcuts) очень мало. Тогда:

$$l = \frac{N}{k} f\left(\frac{N}{\xi}\right) \sim N ,$$

т.е. сеть представляет собой большой мир, среднее расстояние между двумя узлами растет пропорционально числу узлов.

Если же $z = N/\xi \gg 1$, т.е. $N\varphi^{1/d} \gg 1$, то выполняется:

$$l = \frac{N}{k} f\left(\frac{N}{\xi}\right) \sim \log\left[N\left(\varphi k d\right)^{1/d}\right] \sim \log N ,$$

сеть является малым миром, расстояние между узлами растет много медленнее, как логарифм числа узлов.

Вернемся теперь к вопросу о перколяционных свойствах таких сетей. Ответим на несколько вопросов. Что будет происходить, если некоторая доля $q = 1 - p$ узлов выпадет, т.е. не будет проводить (информацию, ток, …)? Чему равно критическое значение $q_c = 1 - p_c$, когда в сети еще существует гигантская связная компонента (giant connected component), т.е. когда значительное (сравнимое с полным) число узлов сети еще соединено между собой? В работе [117] показано, что порог протекания p_c (Np_c- число неразорванных связей) связан с долей переброшенных связей (shortcuts) φ следующим образом (рис. 48)

$$\varphi = \frac{\left(1 - p_c\right)^k}{2kp_c\left[1 + kp_c\left(1 - p_c\right)^k\right]}.$$

На рис. 49 показаны зависимости величины порога протекания p_c от φ, при которой еще есть гигантская связная компонента. Как видно, при протекании, чем меньше переброшенная доля связей, тем больше число узлов должно быть переброшено.

Перколяция на безмасштабных сетях, с распределением степеней узла $P(k) \sim k^{-\gamma}$ обладает своей спецификой, отличной от перколяции в малых мирах. Выражение для величины порога протекания p_c различное для различных диапазонов параметра γ. Так, например, при $\gamma > 3$:

$$q_c = 1 - p_c = 1 - \frac{1}{\frac{\gamma-2}{\gamma-3}k_0 - 1},$$

где степень узла лежит в диапазоне $k_0 \le k \le K_0$.

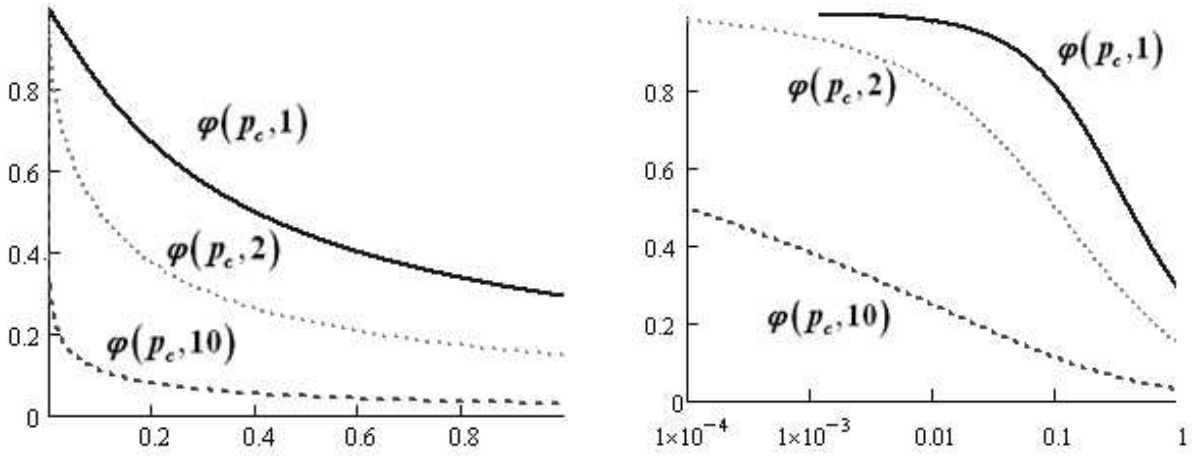


Рис. 49. Зависимость порога протекания - p_c от доли переброшенных связей - φ: сверху вниз – число соседей один, два и десять. На правой части ось абсцисс - в логарифмическом масштабе

При $k_0 = 0$ выражение для q_c упрощается $q_c = 4 - \gamma$. Т.о. образом, если доля выкинутых (разрушенных) узлов больше чем $4 - \gamma$ гигантский кластер отсутствует, т.е. можно считать, что сеть разрушена. Подробно об этом и о том, что происходит в других диапазонах параметра γ можно прочитать в [85].

9.6. Теория перколяции и моделирование атак на сети

Результаты, приведенные в [85], вполне пригодны для моделирования атак на такую глобальную сеть, как Интернет. Известно, что узлы этой сети связываются линиями связи с различной пропускной способностью. Отдельные сегменты Интернет соединены в точки обмена трафиком, существуют опорные (backbone) каналы. Можно констатировать, что сегодня само существование нашей цивилизации уже во многом зависит от связности этой сети. Моделирование атак, т.е. разрывов отдельных связей или удаление отдельных, как показывают расчеты,

очень немногих узлов может нарушить такую связность и дают большие шансы террористам. Нахождение «слабых» участков глобальной сети, проектирование резервных узлов и каналов требуют точных расчетов на базе теории перколяции.

Следует отметить, что выше все время речь шла вырезании (выключении, разрушении …) узлов с вероятность q случайным образом. В тоже время можно вырезать узлы, как в стандартной перколяционной задаче, так в других типах сложных сетей, целенаправленным образом, выбирая такие узлы, при вырезании которых сеть разрушается максимально быстро. В сети Интернет такое направленное вырезание узлов (серверов) называется «запланированной атакой». Под вырезением узлов понимается выведение сервера из его работоспособного состояния, например путем DDos-атаки или атаки на переполнение буфера [136]. При этом выведение из строя порядка 1% целенаправленных узлов уменьшает производительность интернета в два раза.

Возможно также обобщение перколяционной задачи на так называемую диодную или направленную (от англ. directed) перколяцию. В этой модификации некоторые из хорошо проводящих («черных») связей в двухфазной версии перколяции заменяются на «диодные», пропускающие только в одну сторону.

Отдельные веб-сайты (узлы) и гиперсвязи в сети WWW образуют диодную перколяционную сеть, так как в соответствии с протоколом HTTP не предусматривается, чтобы гиперсвязи, ведущие с веб-страниц, вели в обратную сторону. Робот традиционной поисковой системы, такой как Яндекс или Google, сканирует веб-страницы одна за другой, осуществляя переходы по гиперссылкам. Относительно небольшое количество веб-сайтов (например, каталоги веб-ресурсов) содержат большое количество гиперссылок, в то время как большая часть веб-сайтов содержит крайне мало ссылок на другие веб-ресурсы (чаще всего на различные веб-сервисы). Если допустить, что будет произведена атака на веб-сайты, имеющие наибольшее количество ссылок на другие веб-сайты, то, в зависимости от мощности такой атаки, может быть нарушена связность этой диодной сети. В результате поисковые серверы не смогут охватить некоторую часть веб-пространства, что автоматически переведет последнюю в зону «скрытого» веб [141].

10. МОДЕЛИ ИНФОРМАЦИОННЫХ ПОТОКОВ

«Прекрасных видений живой поток
щелчок выключателя не прервет!»
Иосиф Бродский

Анализ динамики информационных потоков, генерируемых в веб-пространстве становится сегодня одним из наиболее информативных методов исследования актуальности тех или иных тематических направлений [10]. Эта динамика обусловлена факторами, многие из которых не поддаются точному анализу. Однако общий характер временной зависимости количества тематических публикаций в Интернет все же допускает построение математических моделей.

10.1. Линейная модель

В некоторых случаях динамика тематических информационных потоков, которая может быть выражена количеством публикаций за определенный период, ее интенсивностью, обусловленной, например, изменением активности тематики (ее повышением или старением), происходит линейно, т.е. количество сообщений в момент времени *t* можно, соответственно, представить формулой:

$$y(t) = y(t_0) + v(t - t_0),$$

где t_0 - некоторое стартовое время отсчета, $y(t)$ – количество сообщений на время t, v – средняя скорость увеличения (уменьшения) интенсивности тематического информационного потока.

Важные характеристики информационного потока могут быть количественно оценены флюктуацией этого потока – изменением среднеквадратичного отклонения $\sigma(t)$, которое вычисляется по формуле:

$$\sigma(t_n) = \sqrt{\tfrac{1}{n}\sum_{i=0}^{n}\left[\, y(t_i) - \left(y(t_0) + v(t_i - t_0)\right)\right]^2}.$$

Если эта величина изменяется пропорционально квадратному корню от времени, то процесс изменения количества публикаций по выбранной теме можно считать процессом с независимыми приращениями. При этом связями с предыдущими тематическими публикациями можно пренебречь.

В случае, когда среднеквадратичное отклонение пропроционально некоторой степени от времени: $\sigma(t) \propto t^{\mu}$ $(\frac{1}{2} \le \mu \le 1)$, чем большее значение μ, тем выше корреляция между текущими и предыдущими сообщениями в информационном потоке.

10.2. Экспоненциальная модель

В некоторых случаях процесс изменения актуальности тематики (увеличение или уменьшение количества тематических сообщений в информационном потоке в единицу времени) аппроксимируется экспоненциальной зависимостью, которую можно выразить формулой:

$$y(t) = y(t_0)e^{\lambda \cdot (t - t_0)},$$

где λ - среднее относительное изменение интенсивности тематического информационного потока.

В реальности актуальность тематики является дискретной величиной, измеряемой в моменты времени t_0, ..., t_n, которая лишь аппроксимируется приведенной выше зависимостью. В рамках данной модели справедливо:

$$y(t_i) = y(t_0)e^{\lambda \cdot (t_i - t_0)} = y(t_0)\mathrm{e}^{\lambda \cdot (t_i - t_{i-1} + t_{i-1} - t_0)} = y(t_{i-1})\mathrm{e}^{\lambda \cdot (t_i - t_{i-1})}.$$

Откуда:

$$\frac{y(t_i)}{y(t_{i-1})} = \mathrm{e}^{\lambda \cdot (t_i - t_{i-1})}.$$

Введем обозначение: $\lambda(t_i)$ - относительное изменение интенсивности тематического информационного потока в момент времени t_i :

$$\lambda(t_i) = \lambda \cdot (t_i - t_{i-1})$$

и прологарифмируем приведенное выше уравнение:

$$\lambda(t_i) = \ln \frac{y(t_i)}{y(t_{i-1})}.$$

Относительное изменение интенсивности в момент времени t_i на практике также часто вычисляется как соотношение:

$$\lambda(t_i) = \ln \frac{y(t_i)}{y(t_{i-1})} \approx \frac{y(t_i) - y(t_{i-1})}{y(t_{i-1})}.$$

Изменение флюктуаций величины $\lambda(t_i)$ относительно среднего значения может оцениваться по стандартному отклонению:

$$\sigma(t_n) = \sqrt{\frac{1}{n}\sum_{i=0}^{n}\left(\lambda(t_i) - \lambda\right)^2}.$$

В этом случае также, если $\sigma(t)$ изменяется пропорционально корню квадратному от времени, то можно говорить о процессе с независимыми приращениями, корреляции между отдельными сообщениями несущественные. В случае наличия значительной зависимости сообщений наблюдается соотношение: $\sigma(t) \propto t^{\mu}$, причем μ превышает ½, но ограниченно 1.

Значения μ, которое превышает ½, говорит о наличии долгосрочной памяти в информационном потоке. Такой класс процессов получил название автомодельных, для которых предполагается корреляция между количеством публикуемых сообщений в разные моменты времени.

10.3. Логистическая модель

Логистическую модель [5, 6, 11] можно рассматривать как обобщение экспоненциальной модели Мальтуса, которая предусматривает пропорциональность скорости роста функции $y(t)$ в каждый момент времени ее значению :

$$\frac{dy(t)}{dt} = ky(t),$$

где $k-$ некоторый коэффициент.

Главная идея логистической модели заключается в том, что для ограничения скорости роста на функцию $y(t)$ накладывается дополнительное условие,

согласно которому ее значение не должно превышать некоторой величины [4, 5]. С этой целью в правую часть уравнения вводится дополнительный множитель вида:

$$N - ry(t),$$

где $N -$ пороговое значение, которое функция $y(t)$ не может превысить, а r – коэффициент, который описывает отрицательные для данной тенденции процессы.

Уравнение, полученное таким способом, называется логистическим и в общем случае (вместе с начальным условием) имеет следующий вид:

$$\begin{cases} \dfrac{dy(t)}{dt} = ky(t)(N - ry(t)), \\ y(t_0) = y_0. \end{cases}$$

Информационную динамику в общем случае можем представить как процесс, обусловленный возникновением и исчезновением отдельных тематик, происходящими на фоне общих тенденций информационного пространства.

Зафиксируем определенную тематику и предположим, что в момент времени $t = 0$ существует n_0 фоновых публикаций. Вследствие того, что (в рамках принятой модели) актуальность тематики сохраняется на протяжении промежутка времени λ, можно рассматривать отдельно две временные области: $0 < t \le \lambda$ с $D > 0$ и $t > \lambda$ с $D = 0$ (в рамка данной модели $D = const$ для каждой области - уровень актуальности темы) и, соответственно, функции $u(t)$ и $v(t)$, которые являются решениями для этих областей и “сшиваются” в точке λ:

$$y(t) = \begin{cases} u(t), \ \ 0 < t < \lambda, \\ v(t), \ \ t > \lambda, \\ u(t) = v(t), \ \ t = \lambda. \end{cases}$$

Первой области соответствует процесс роста количества публикаций в условиях ненулевой актуальности темы ($D > 0$) и, возможно, переход к состоянию насыщения.

Реакция медийных средств никогда не бывает мгновенной: всегда существует определенная задержка во времени. Этот аспект учитывается в модели путем введения фактора запаздывания τ.

Соответствующая динамика описывается уравнением, которое после переопределения коэффициентов и их нормирования к N, для функции $u(t)$ можно представить в виде:

$$\frac{du(t-\tau)}{dt} = pu(t-\tau)(1-qu(t-\tau)) + Du(t-\tau),$$

$$u(0) = n_0.$$

Подчеркнем, что содержательно величина p определяет нормированную вероятность появления публикации в единицу времени независимо от актуальности темы. Этот фактор отражает фоновые механизмы генерации информации (типичным примером может быть механическая перепечатка материалов из престижных информационных источников). Величина же D характеризует непосредственное влияние актуальности данной темы. Параметр q характеризует уменьшение скорости роста количества публикаций и является величиной, обратной к асимптотическому значению зависимости $u(t)$ при $D = 0$.

Для второй области, описываемой функцией $v(t)$, соответственно, имеем:

$$\frac{dv(t-\lambda)}{dt} = pv(t-\lambda)(1-qv(t-\lambda)).$$

При этом должно учитываться условие равенства функций $u(t)$ и $v(t)$ в момент $t = \lambda$:

$$v(\lambda) = u(\lambda).$$

Приведенные выше нелинейные дифференциальные уравнения являются вариантами записи уравнения Бернулли:

$$y' = ay^2 + by,$$

которое линеаризируется стандартной заменой $z = 1/y$:

$$z' + bz + a = 0.$$

Общее решение этого уравнения имеет вид:

$$z = \frac{1}{\mu(x)}[C - a\int \mu(x)dx]$$

с интегрирующим множителем:

$$\mu(x) = e^{bx}.$$

Переменные C определяются: для первой области из начальных условий, а для второй – из условия «сшивания». Путем несложных преобразований находим решение для первой области:

$$u(t) = \frac{u_s}{1 + (\frac{u_s}{n_0} - 1)\exp[-(p + D)(t - \tau)]},$$

где u_s – асимптотическое значение u, величина которого определяет область насыщения:

$$u_s = \frac{p + D}{pq}$$

Таким образом, мы видим, что модель правильно описывает зависимость, которая имеет *S*-подобную (логистическую) форму, представленную на рис. 50.

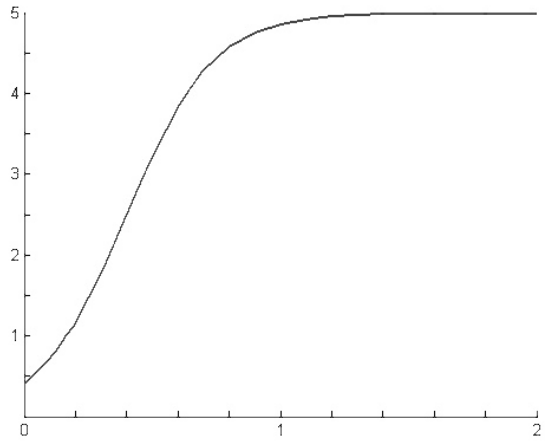


Рис. 50. Функция роста $u(t)$

Заметим, что решение не зависит от значения n_0, что свидетельствует о несущественности начальных условий для информационной динамики. Каким бы ни было начальное количество публикаций, насыщение будет определяться исключительно параметрами, которые характеризуют фоновую скорость роста

количества публикаций, количественную меру актуальности и отрицательные для процесса факторы.

Кривая, представленная на рис. 50 обладает точкой перегиба:

$$t_{\inf} = \frac{1}{p+D}\ln(\frac{u_s}{n_0} - 1) + \tau. \qquad (10)$$

Таким образом, для первой области имеем так называемую S-подобную зависимость, а при $t \sim t_{\inf}$ поведение $u(t)$ приближается к линейному и соответствует линейной модели.

Представим теперь для удобства выражение для $u(t)$ несколько в ином виде:

$$u(t) = \frac{u_s \exp[(p+D)t]}{\exp[(p+D)t] + (\frac{u_s}{n_0} - 1)\exp[(p+D)\tau]},$$

откуда видно, что при условии

$$t < \frac{1}{p+D}\ln(\frac{u_s}{n_0} - 1) + \tau = t_{\inf}$$

зависимость $u(t)$ имеет экспоненциальный характер, т.е. для значений t, значительно менших $t_{\inf}$, модель совпадает с экспонентциальной моделью.

Для второй области, соответственно, имеем (рис. 51):

$$v(t) = \frac{v(\lambda)}{qv(\lambda) + (1 - qv(\lambda))\exp[-p(t-\lambda)]},$$

учитывая условие «сшивки»:

$$v(\lambda) = u(\lambda).$$

Если зависимость $u(t)$ успевает достичь насыщения за промежуток времени $t < \lambda$, то приведенное выше уравнение можно упростить, представив его следующим образом:

$$v(t) = \frac{v_s(p+D)}{p + D(1 - \exp[-p(t-\lambda)])},$$

где $v_s = 1/q$ - асимптотическое значение зависимости $v(t)$.

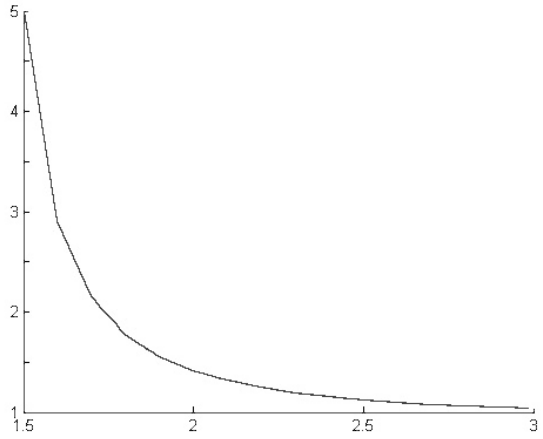


Рис. 51. Функция спада $v(t)$

Как и ожидалось, величина v_s также не зависит ни от начального условия, ни от условия “сшивания” с функцией $u(t)$ на границе областей. Как видно, полученная зависимость имеет область насыщения u_s (при $t \le \lambda$) и асимптотику v_s, которая описывает постепенное уменьшение числа публикаций до фонового уровня. А это означает, что она, по крайней мере, на качественном уровне, согласуется с общими представлениями о характере информационной динамики, полученными на основе опытных данных. Кроме того, на локальных участках она неплохо аппроксимируется линейной и экспоненциальной моделями.

Типовая полная зависимость $y(t)$ приведена на рис. 52.

В случае информационных потоков, которые ассоциируются с конкретными темами, необходимо описывать динамику каждого из таких потоков отдельно, принимая во внимание то, что рост одного из них автоматически приводит к уменьшению остальных и наоборот. Поэтому ограничение на количество сообщений по всем тематикам распространяется и на совокупность всех монотематических потоков.

В случае изучения общего информационного потока наблюдается явление “перетекания” публикаций из одних, теряющих актуальность тематик, к другим. Действительно, каждый информационный ресурс, веб-сайт, имеет определенные мощности, которые зависят как от технических аспектов, так и от конъюнктуры предметной области, т.е. каждый ресурс публикует более или менее стандартное количество сообщений в единицу времени.

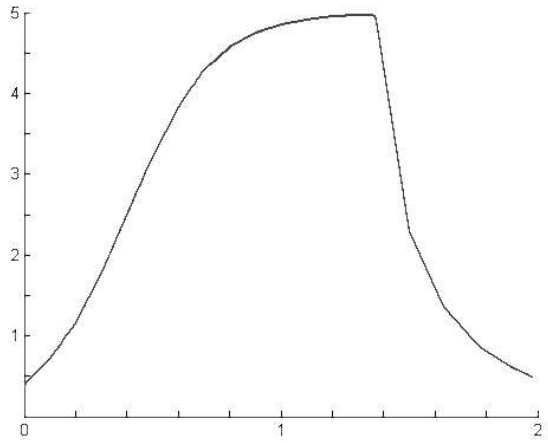


Рис. 52. Обобщенный график динамики тематического информационного потока

В результате на практике для локальных временных промежутков можно наблюдать «тематический баланс»:

$$\int_0^T \sum_{i=1}^{N_m} m_i(t)dt = MT ,$$

где $m_i(t)$ – плотность публикаций по i-й тематике, N_m – количество тематик, а M – константа, которая характеризует имеющиеся объемы информации.

Общая динамика должна описываться системой уравнений, каждое из которых относится к отдельному монотематическому потоку. Подчеркнем, что общие политематические потоки являются стационарными по количеству публикаций, динамика же в основном определяется «конкурентной борьбой» отдельных тематик.

В литературе описано много разновидностей систем «конкурентной борьбы» для разных модификаций модели в зависимости от целого ряда предположений относительно реальных условий протекания процессов. В простейшем виде такие уравнения могут иметь такой вид:

$$\frac{dm_i(t)}{dt} = p_i \cdot m_i(t) - \sum_{j=1}^{N_m} r_{ij} \cdot m_i(t) \cdot m_j(t),$$

где N_m – количество тематик.

Приведенная система уравнений описывает перераспределение публикаций между тематиками, образующими фиксированный набор. Но в реальной жизни тематики (сюжеты) появляются и со временем исчезают, поэтому необходимо ввести в эти уравнения соответствующие коррективы. Это можно сделать по-разному, например, определив коэффициенты p_i и r_{ij} зависимыми от времени так, чтобы каждый сюжет имел собственный максимум активности на определенном промежутке времени. По аналогии такие уравнения можно представить в таком виде:

$$\frac{dm_i(t)}{dt} = \left(p_i + D_i(t, \lambda_i) \right) \cdot \left(m_i(t) - \sum_j r_{ij} \cdot m_i(t) \cdot m_j(t) \right).$$

В этом соотношении коэффициенты p_i и D_i имеют тот же смысл, что и ранее, а λ_i являются точками, в которых соответствующие D_i достигают максимальных значений.

10.4. Модель диффузии информации

Обратимся к еще одному направлению в изучении процессов, связанных с информационными потоками - к диффузии информации.

Напомним, что в естественных науках под диффузией понимают взаимное проникновение друг в друга соприкасающихся веществ, вызванное, например, тепловым движением их частиц.

Для понимания сути дела необходимо, прежде всего, учитывать, что информация также в определенном смысле состоит из «частиц» - документов (сообщений). Множество процессов, близких к динамике информационных потоков, можно моделировать довольно точно, если четко параметризовать и установить их предельные параметры.

Процессы диффузии информации, как и процессы диффузии в физике, достаточно точно моделируются с помощью методов клеточных автоматов.

Концепция клеточных автоматов была впервые предложена больше полстолетия тому назад Дж. Фон Нейманом (J. Von Neumann) [41] и развита

С. Вольфрамом (S. Wolfram) в фундаментальной монографии “Новый вид науки” [149].


Дж. Фон. Нейман (1903-1957) и С. Вольфрам

Клеточные автоматы являются полезными дискретными моделями для исследования динамических систем [53]. Дискретность модели, а точнее, возможность представить модель в дискретной форме, может считаться важным преимуществом, поскольку открывает широкие возможности использования компьютерных технологий. Клеточные автоматы в этом смысле занимают особое место, поскольку их дискретность объединяется с другими преимуществами.

Главным достоинством клеточных автоматов является их абсолютная совместимость с алгоритмическими методами решения задач. Оконченный набор формальных правил, заданный на ограниченном множестве элементов (клеток), допускает точную реализацию в виде алгоритмов. Однако отсюда вытекает и главный недостаток клеточных автоматов: вычислительные трудности, которые возникают при расчетах соответствующих масштабов. Ведь на каждой итерации необходимо сканировать весь набор клеток и для каждой из них выполнять необходимые операции. Когда и клеток, и итераций действительно много, требуются значительные ресурсы, в том числе вычислительные и временные.

Поэтому продолжительное время клеточные автоматы воспринимались в основном как забавная, хотя и поучительная игра, которая не имеет практической ценности. Но в последние годы, в связи с бурным развитием компьютерных

технологий, они начинают быстро входить в арсенал инструментальных средств, которые используются на практике в различных областях науки и техники.

Клеточный автомат представляет собой дискретную динамическую систему, совокупность одинаковых клеток, одинаковым образом соединенных между собой. Все клетки образуют сеть (решетку) клеточных автоматов. Состояние каждой клетки определяются состоянием клеток, входящих в ее локальную окрестность и называемых ближайшими соседями. Окрестностью конечного автомата с номером j называется множество его ближайших соседей. Состояние j-го клеточного автомата в момент времени $t+1$, таким образом, определяется следующим образом:

$$y_j(t+1) = F\left(y_j,\ O(j),\ t\right),$$

где F – некоторое правило, которое можно выразить, например, языком булевой алгебры. Во многих задачах считается, что сам элемент относится к своим ближайшим соседям, т.е. $y_j \in O(j)$, в этом случае формула упрощается: $y_j(t+1) = F\left(O(j),\ t\right)$. Клеточные автоматы в традиционном понимании удовлетворяют таким правилам:

- изменение значений всех клеток происходит одновременно (единица измерения - такт);
- сеть клеточных автоматов однородная, т.е. правила изменения состояний для всех клеток одинаковые;
- на клетку могут повлиять лишь клетки из ее локальной окрестности;
- множество состояний клетки конечное.

Теоретически клеточные автоматы могут иметь любую размерность, однако чаще всего рассматривают одномерные и двумерные системы клеточных автоматов.

Модель диффузии информации, которую будем рассматривать в дальнейшем, является двумерной, поэтому дальнейший формализм касается этого случая. В двумерном клеточном автомате решетка реализуется двумерным массивом. Поэтому в этом случае удобно перейти к двум индексам, что вполне корректно для двумерных конечных решеток.

В случае двумерной решетки, элементами которой являются квадраты, ближайшими соседями, входящими в окрестность элемента $y_{i,j}$, можно считать или только элементы, расположенные вверх-вниз и влево-вправо от него (так называемая окрестность фон Неймана: $y_{i-1,j}$, $y_{i,j-1}$, $y_{i,j}$, $y_{i,j+1}$, $y_{i+1,j}$), либо добавленные к ним еще и диагональные элементы (окрестность Мура (G. Moore): $y_{i-1,j-1}$, $y_{i-1,j}$, $y_{i-1,j+1}$, $y_{i,j-1}$, $y_{i,j}$, $y_{i,j+1}$, $y_{i+1,j-1}$, $y_{i+1,j}$, $y_{i+1,j+1}$).

В модели Мура каждая клетка имеет восемь соседей. Для устранения краевых эффектов решетка топологически «сворачивается в тор» (рис. 53), т.е. первая строка считается продолжением последней, а последняя – предшествующей первой. То же самое относится и к столбцам.

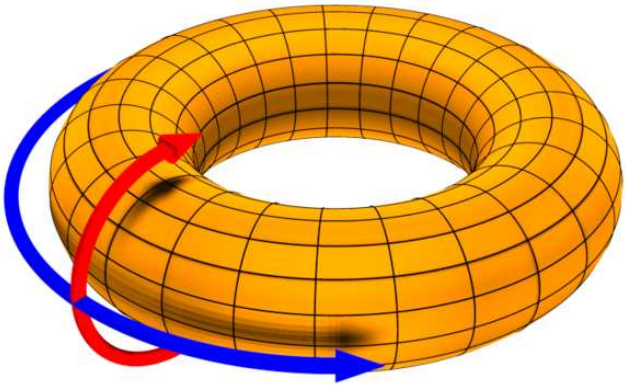


Рис. 53. Свертывание плоскости в тор. Источник: wikimedia.org

Это позволяет определять общее соотношение значения клетки на шаге $t+1$ по сравнению с шагом t:

$$y_{i,j}(t+1) = F\left(y_{i-1,j-1}(t), y_{i-1,j}(t), y_{i-1,j+1}(t), y_{i,j-1}(t), y_{i,j}(t), y_{i,j+1}(t), y_{i+1,j-1}(t), y_{i+1,j}(t), y_{i+1,j+1}(t)\right).$$

С. Вольфрам, классифицируя различные клеточные автоматы, выделил те, динамика которых существенным образом зависит от начального состояния. Подбирая различные начальные состояния, можно получать разнообразнейшие конфигурации и типы поведения. Именно к таким системам относится классический пример - игра "Жизнь", изобретенная Дж. Конвеем (J. Conway) и известная широкому кругу читателей благодаря публикации в книге М. Гарднера (M. Gardner) [12].

Клеточные автоматы с успехом применяются при моделировании процессов распространения инноваций [76]. Клеточные автоматы также используются при моделировании электоральных процессов, в этом случае предполагается, что избирательные преференции человека определяются установками его ближайшего окружения [43].

В одной из моделей предполагается, что индивид принимает решение голосовать в момент $t+1$ за республиканцев или демократов в соответствии с правилом простого большинства. В этой модели учитывались взгляды индивида и четырех его ближайших соседей в момент t (окрестность фон Неймана). Модель исследовалась на большом временном отрезке - до 20 000 тактов. Оказалось, что партийная борьба приводит к очень сложным конфигурациям, которые существенным образом зависят от исходного распределения.

Как упрощенную модель диффузии информации сначала рассмотрим признанную модель распространения инноваций [76]. Подобная модель функционирует по следующим правилам: каждый индивид, который способен принять инновацию, соответствует одной квадратной клетке на двумерной плоскости. Каждая клетка может находиться в двух состояниях: 1 - новинка принята; 0 - новинка не принята. Предполагается, что автомат, восприняв инновацию один раз, запоминает ее навсегда (состояние 1, которое не может быть измененным). Автомат одобряет решение относительно принятии новинки, ориентируясь на мнение восьми ближайших соседей, т.е. если в окрестности данной клетки (используется окрестность Мура) есть m приверженцев новинки, p - вероятность ее принятия (генерируется в ходе работы модели) и если $pm > R$, (R - фиксированное предельное значение), то клетка принимает инновацию (значение 1). По мнению авторов этой модели, клеточное моделирование позволяет строить значительно более реалистические модели рынка инноваций, чем традиционные подходы.

Вместе с тем динамике распространения информации присущи некоторые дополнительные свойства, которые были учтены в представленной ниже модели. В модели диффузии информации, наряду с теми же условиями, которые относятся к клеточному пространству, окрестности Мура и вероятностному правилу принятия

новости, дополнительно к у условиям диффузии инноваций предполагалось, что клетка может быть в одном из трех состояний: 1 - «свежая новость» (клетка окрашивается в черный цвет); 2 - новость, которая устарела, но сохраненная в виде сведений (серая клетка); 3 - клетка не имеет информации, переданной новостным сообщением (клетка белая, информация не дошла или уже забыта). В модели приняты такие правила распространения сообщений:

- сначала все поле состоит из белых клеток за исключением одной, черной, которая первой «приняла» новость (рис. 54 а);
- белая клетка может перекрашиваться только в черный цвет или оставаться белой (она может получать новость или оставаться «в неведении»);
- белая клетка перекрашивается, если выполняется условие, аналогичное модели диффузии инноваций: $pm > 1$ (это условие несколько модифицируется для $m \le 2$: $1.5 \cdot pm > 1$);
- если клетка черная, а вокруг нее исключительно черные и серые, то она перекрашивается в серые цвета (новость устаревает, но сохраняется как сведения);
- если клетка серая, а вокруг нее исключительно серые и черные, то она перекрашивается в белый цвет (происходит старение новости при ее общеизвестности).

Описанная система клеточных автоматов вполне реалистично отражает процесс распространения сообщений среди отдельных информационных источников. Авторами были реализован приведенный выше алгоритм на поле размером 40 х 40 (размеры были выбраны исключительно с целью наглядности). Выяснилось, что состояние системы клеточных автоматов полностью стабилизируется за ограниченное количество ходов, т.е. процесс эволюции оказался сходящимся. Пример работы модели приведен на рис. 54.

Многочисленные эксперименты с данным клеточным автоматом, доступным в настоящее время в Интернет по адресу http://edu.infostream.ua/newsk.pl показывают, что период его сходимости составляет от 80 до 150 шагов.

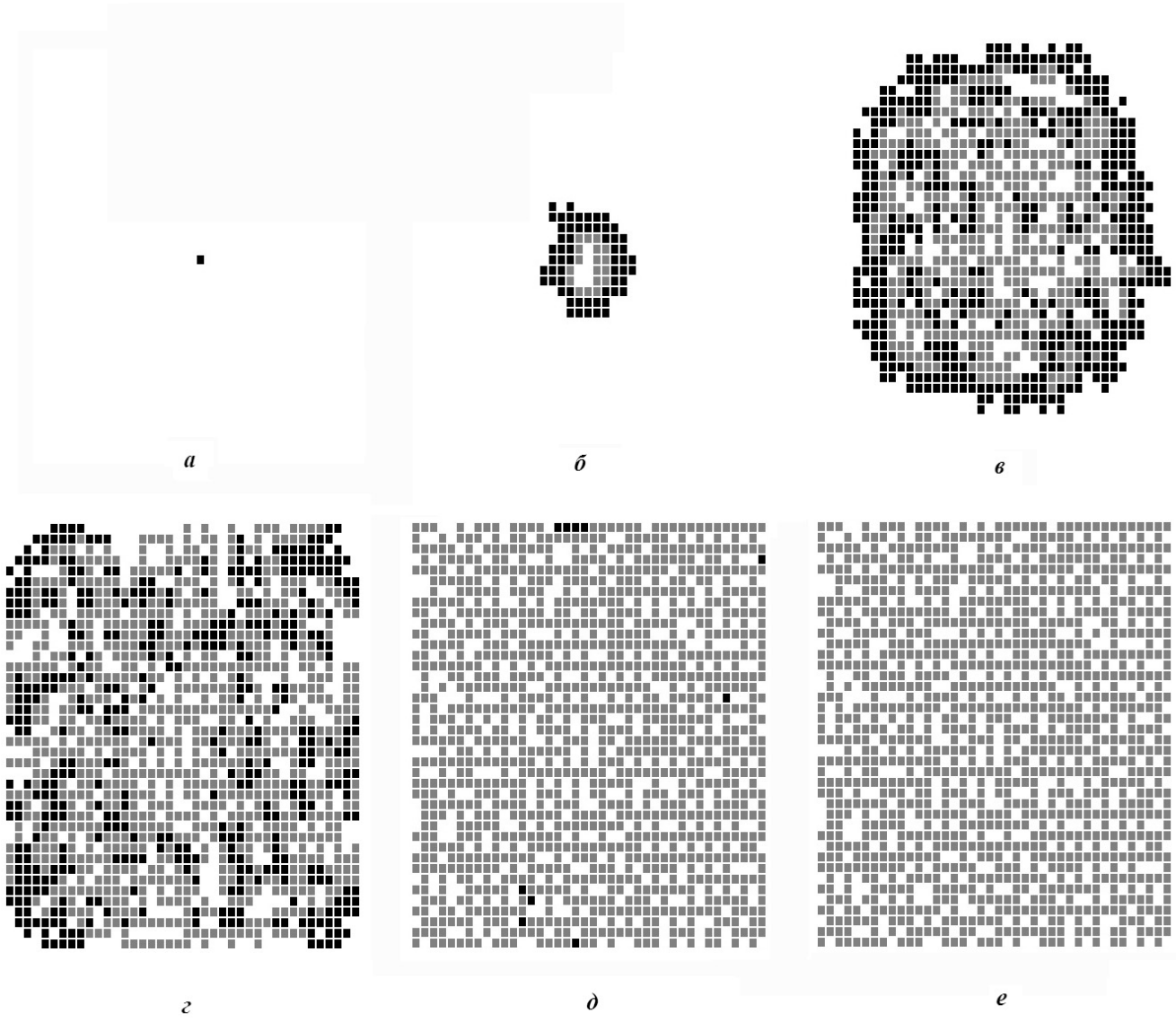


Рис. 54. Процесс эволюции системы клеточных автоматов «диффузии новостей»: а - исходное состояние; б-д - промежуточные состояния; е - конечное состояние

Типичные зависимости количества клеток, которые находятся в разных состояниях в зависимости от шага итерации приведены на рис. 55. При анализе приведенных графиков следует обратить внимание на такие особенности: 1 - суммарное количество клеток, которые находятся во всех трех состояниях на каждом шагу итерации постоянно и равно количеству клеток, 2 - при стабилизации клеточных автоматов соотношения серых, белых и черных клеток приблизительно составляет: 3:1:0; существует точка пересечения всех трех кривых.

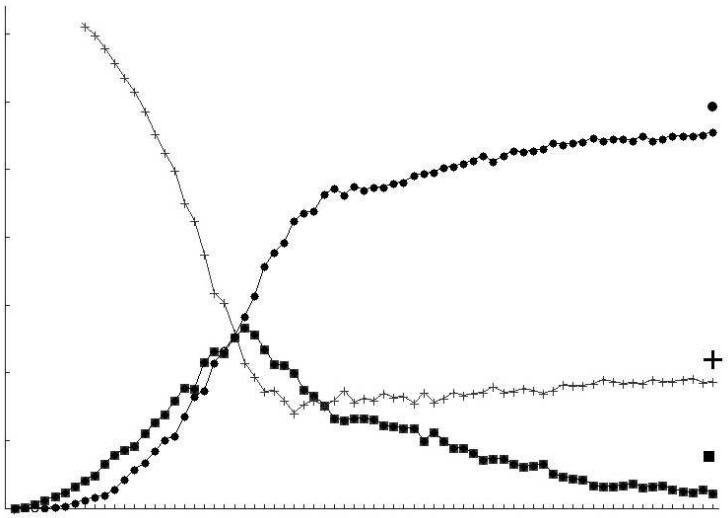


Рис. 55. Количество клеток каждого из цветов в зависимости от шага эволюции: белые клетки - (+); серые клетки - (•); *черные клетки* – (■)

Детальный анализ полученных зависимостей позволил провести аналогии данной модели «диффузии информации» с некоторыми аналитическими соображениями [107]. Результаты моделирования дают основания предположить, что эволюция серых клеток описывается некоторой непрерывной функцией:

$$x_g = f\left(t, \tau_g, \gamma_g\right),$$

где t - время (шаг эволюции), τ_g - сдвиг по времени, обеспечивающий получение необходимого фрагмента аналитической функции, γ_g - параметр крутизны данной функции.

Соответственно, динамика белых клеток x_w (количество клеток в момент t) может моделироваться «перевернутой» функцией x_g с аналогичными параметрами:

$$x_w = 1 - f\left(t, \tau_w, \gamma_w\right).$$

Поскольку, как было сказано выше, всегда выполняется условие баланса, т.е. общее количество клеток в любой момент времени всегда постоянно, то условие нормирования можно записать следующим образом:

$$x_g + x_w + x_b = 1,$$

где x_w - количество черных клеток в момент времени t.

Таким образом, получаем:

$$x_b = 1 - x_g - x_w = f\left(t, \tau_w, \gamma_w\right) - f\left(t, \tau_g, \gamma_g\right).$$

Вид представленной на рис. 55 зависимости позволяет предположить, что в качестве функции $f\left(t, \tau, \gamma\right)$ может быть выбрано следующее выражение (логистическая функция):

$$f\left(t, \tau, \gamma\right) = \frac{C}{1 + e^{\gamma(t-\tau)}},$$

где C - некоторая нормирующая константа.

На рис. 56 приведены графики зависимости x_g, x_w, x_b от шага эволюции системы клеточных автоматов, полученные в результате аналитического моделирования.

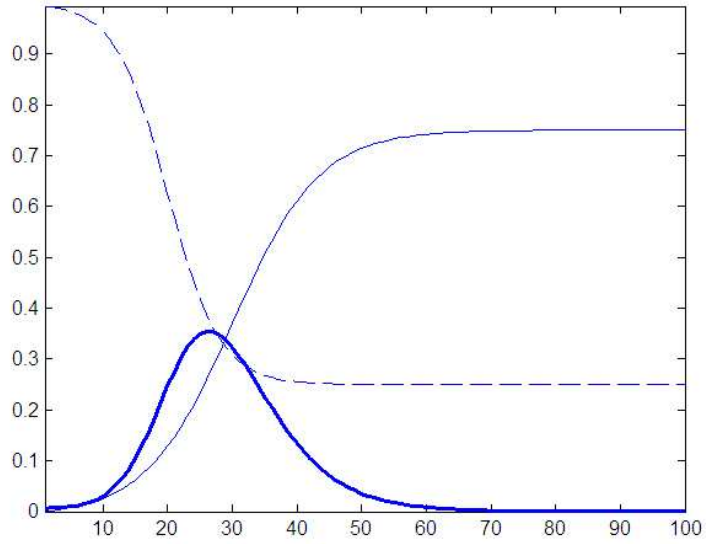


Рис. 56. Непрерывные зависимости, полученные в результате аналитического моделирования, в зависимости от шага эволюции: сплошная линия – серые (x_g); пунктирная линия – белые (x_w); сплошная жирная линия – черные (x_b)

Следует отметить, что зависимость диффузии новостей, полученная в результате моделирования, хорошо согласуется с «жизненным» поведением тематических информационных потоков на интернет-источниках (веб-сайтах), а на локальных временных промежутках - с традиционными моделями.

10.5. Модель самоорганизованной критичности

Как было показано на примере в предыдущем пункте, методы описания процессов переноса, подробно рассматриваемые в физике, могут быть применимы и к информационным процессам. Интуитивно ясно, что у каждого тематического информационного потока есть «характер», в связи с чем, можно говорить и о наиболее адекватной модели описания распространения информации по выбранной теме.

Под системой, порождающей информацию, чаще всего предполагают реальную социальную или экономическую систему, ожидать от которой простой предсказуемой информации или единообразного поведения не приходится. В реальной системе информационное событие можно рассматривать в каком-то смысле как катастрофу, поскольку оно неожиданно. Если в большинстве случаев предсказание отдельного события кажется невозможным, то поведение системы в целом, ее отклик на воздействие или возмущение частично предсказуемо и является объектом научного исследования.

Термин “самоорганизация” (“self-organizing”), связанный с общей теорией систем, был введен В. Ашби (V. Ashby) в 1947 году и воспринят новой тогда кибернетикой, ее создателями Н. Винером (N. Winner), Г. Форстером (G. Forster) и др. В настоящее время это понятие чаще всего ассоциируется с именем П. Бака (P. Bak) [7, 69]. В 1987-1988 году П. Бак, Ч. Танг (C. Tang) и К. Визенфельд (K. Wiesenfeld) в своих работах [70, 71] впервые детально описали клеточный автомат, приводивший систему к статистически одному и тому же «критическому» состоянию, названному ими состоянием самоорганизованной критичности. Типичная стратегия физики заключается в уменьшении количества степеней свободы в исследуемой задаче, например, в теории среднего поля, где окружение воздействует на оставшуюся степень свободы системы как некоторое среднее поле, оставляя для исследования только одну переменную.

Пер Бак (1948 – 2002)

В настоящее время, появилась возможность исследовать системы, сохраняя в них все степени свободы, не объединять их в одну. Таким образом, любое возмущение будет изменять состояние системы, но поскольку на практике она постоянно находится в некотором состоянии равновесия, то этот баланс между стабильностью и изменчивостью можно назвать критическим состоянием системы. В этом смысле можно говорить об экономической, социальной, экологической системе и, конечно же, системе потоков информации.

Самой наглядной моделью, демонстрирующей самоорганизованную критичность, является куча песка, знакомая каждому с детства. Если песок сухой, то никакой кулич из него не построить, всё тут же осыпается. В детские годы мало кто задумывался о том, как это происходит. Какой бы высоты куча не была, угол наклона конуса осыпания оставался неизменным, Это еще раз в эксперименте доказали в эксперименте выросшие дети, в Чикагском университете под руководством Х. Ягера (H. Yager) они экспериментировали с самой настоящей кучей песка.

Состояние этой кучи можно назвать критическим, поскольку приложив минимальное возмущение, бросив сверху одну песчинку, поверхность кучи выйдет из равновесия, вниз сойдет лавина. А после ее схода останется снова куча песка поменьше, новые падающие песчинки достроят кучу до того же критического наклона, а новая брошенная песчинка вновь вызовет лавину. Куча

всегда находится в критическом состоянии – малые возмущения вызывают реакцию, непредсказуемую по размеру, и всегда самоорганизуется – сохраняет угол наклона поверхности (рис. 57).

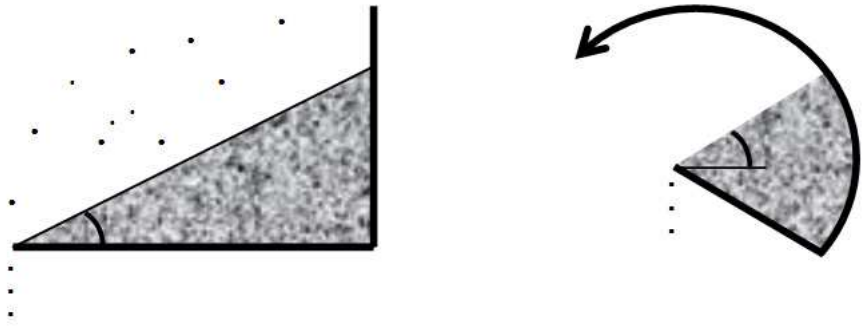


Рис. 57. Коробка и вращающийся барабан с песком с одинаковым углом наклона боковой плоскости

При моделировании самоорганизованной критичности исследуется статистика схода лавин, когда одна брошенная песчинка вызывает лавину из других, лежащих на поверхности.

Рассмотрим дискретную систему, аналогичную куче песка в одномерном случае. Пусть $h(x)$ – высота кучи в точке x $(x=1,\ 2,\ ...,\ N)$. Кучу удобно изображать в двух видах – исходном (рис. 58 *а*) как функцию высоты от координат $h=h(x)$ и в виде приращений $z(x)=h(x)-h(x+1)$, которые показывают отличие высот в соседних точках – рис. 58 *б*. Левые части рис. 57 показывают начальное состояние кучи.

Введем правило 1: если разница высот в точке x больше некоторого критического значения $h(x)>h_c$, то лишние песчинки скатываются на соседние точки. Выбирая критическое значение $h_c=3$ правило 1 можно записать следующим образом:

$$\left.\begin{array}{l} z(x)\to z(x)-2, \\ z(x\pm1)\to z(x\pm1)+1 \end{array}\right|_{z(x)>2}$$

Первое соотношение означает, что высота кучи в точке x уменьшается на две песчинки, второе, что в соседних точках (левой и правой) высота увеличится на одну песчинку.

На границах кучи будут выполняться граничные условия 1:

$$\left.\begin{array}{l} z(1)=0, \\ z(N)\to z(N)-1, \\ z(N-1)\to z(N-1)+1 \end{array}\right|_{z(N)>2}$$

Первое из условий 1 можно назвать «закрытым», поскольку наружу системы частица при этом никогда не выйдет, в противоположность «открытым» условиям на другой стороне, когда частица скатывается наружу и падает.

На рис. 58 слева изображено начальное состояние кучи – высоты $h(x)$ и приращения $z(x)$, справа – после осыпания. Так, например, при $x=6$ приращение до осыпания было равно $z(6)=3>2$. После осыпания, две песчинки с позиции $x=6$, переходят по одной налево $x=5$ и направо $x=7$ - рис. 58 слева.

С одной стороны правило *1*, это дискретное нелинейное уравнение диффузии а, с другой стороны, это клеточный автомат, в котором состояние ячейки x в момент времени $t+1$ определяется состоянием/соседних ячеек в предыдущий момент времени t. Графически действие правила 1 можно представить так, как это сделано на рис. 58.

Очевидно, что у одномерной кучи, когда она осыпается согласно правилу 1 и условиям 1 из неравновесного состояния с $z(x)\gg 2$, есть одно критическое состояние $z(x)=2$ для любого x. В одномерном случае любое стабильное состояние является в определенном смысле критическим, поскольку любое малое возмущение будет приводить к тому, что оно пройдет по всей системе, а любое уменьшение наклона до $z(x)<2$ в любой точке остановит его. Это очень похоже на другие одномерные критические явления, такие как перколяция. Также следует отметить, что у такого состояния в одномерной модели нет пространственной структуры.

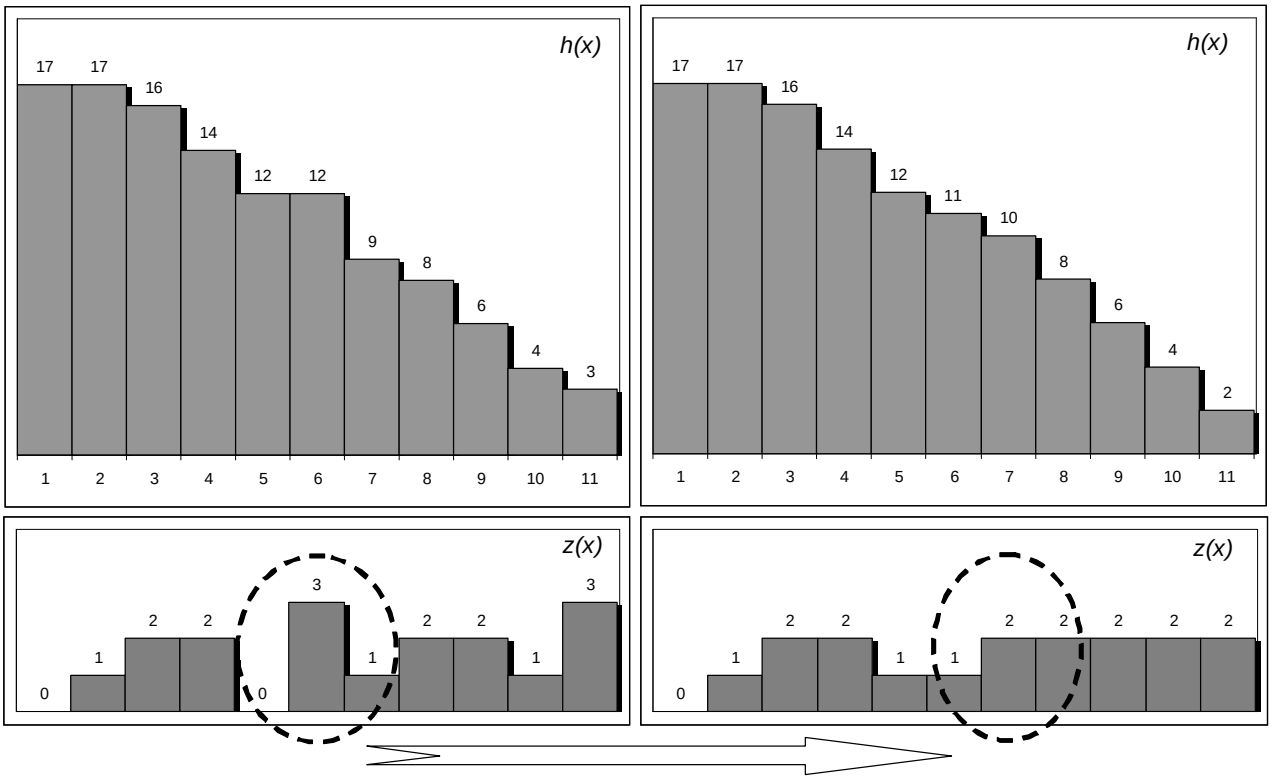


Рис. 58. Одномерная куча до и после осыпания. Осыпался столбец 6 и крайний правый столбец 11 с «открытыми» граничными условиями

Аналогично одномерному П. Бак предложил правило для двумерной кучи (правило 2 и условия 2). В такой системе сохраняется действие правил 1 для каждого из направлений x и y, критическое значение традиционно выбирается равным 3:

$$\left.\begin{aligned} &z(x,y) \to z(x,y) - 4, \\ &z(x,y \pm 1) \to z(x,y \pm 1) + 1, \\ &z(x \pm 1,y) \to z(x \pm 1,y) + 1 \end{aligned}\right| \; {\scriptstyle z(x,y)>3} \qquad (\textit{Правило 2})$$

$$z(0,y) = z(x,0) = z(N+1,y) = z(x,N+1) = 0 \qquad (\textit{Условия 2})$$

Указан вариант «закрытых» граничных условий 2 по всем направлениям. Конечно возможна любая комбинация «открытых» и «закрытых» условий. Условия 2 также могут быть модифицированы для «настоящей» кучи, насыпанной в углу, скажем, обувной коробки. Решетки, на которых строились самоорганизующиеся критические системы также разнообразны, например, проводились эксперименты на квадратных решетках в больших размерностях, на

гексагональных решетках были даже получены точные аналитические результаты. Аналогичные клеточные автоматы возможно построить и на нерегулярных решетках.

Соответствие между величиной $z(x,y)$ и наклоном кучи не такое однозначное, как в 1D, поскольку теперь значение наклона $z(x,y)$ представляет собой средний наклон по диагонали системы и при осыпании частицы начнут двигаться в обоих направлениях x и y. В двумерном случае уже нельзя говорить о том, что из неустойчивого состояния с $z(x,y) \gg 4$ система перейдет в одно и то же состояние, поскольку неустойчивость будет распространяться по обоим направлениям взаимозависимо. Конечное состояние системы будет существенно зависеть от начальных условий, но свойства этого получившегося состояния, например, наклон, будут всегда одинаковыми.

Получить систему в состоянии самоорганизованной критичности можно двумя различными способами. Либо осыпанием системы из случайного состояния с $z(x,y) \gg 4$ до равновесного состояния, либо насыпая на ровную поверхность $z(x,y) = 0$ песчинки одну за другой в случайно выбранных точках и выполняя процедуру согласно правилу 2 тогда, когда это будет необходимо. Определить момент когда система достигает критического уровня можно по тому, средний наклон кучи перестанет изменяться. Эксперимент показывает, что свойства систем полученных обоими способами не отличаются друг от друга.

Результаты, приведенные ниже, получены на квадратной решетке в двумерном случае согласно определенным ранее правилам и условиям. Сначала случайным образом выбирались $z(x,y) \gg 4$, после чего проводилась «релаксация» кучи и она осыпалась до устойчивого состояния.

На рис. 59 представлено одно из таких стабильных состояний на двумерной решетке 500 x 500, где цвета от черного до белого соответствуют значениям $z(x,y)$ от 0 до 3.

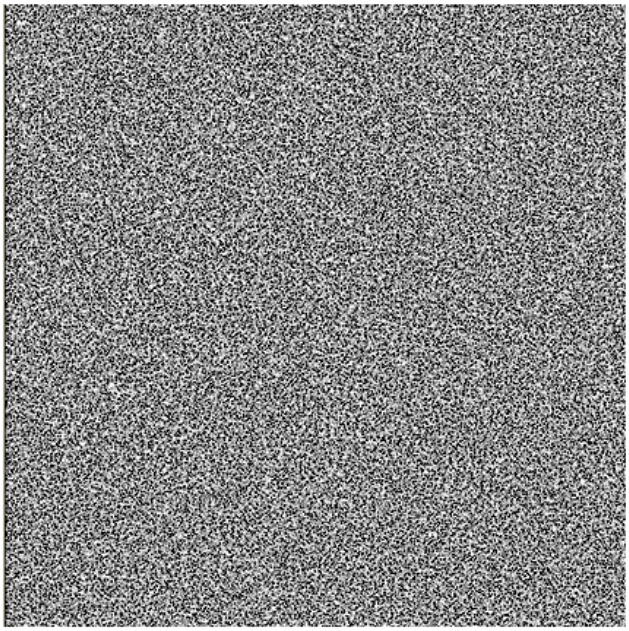


Рис. 59. Стабильное состояние

Если в одной из наиболее неустойчивых точек системы (в нашем случае $z(x,y)=3$), запустить процесс по *правилу 2* с *условиями 2*, положив $z(x,y)=4$, т.е. добавить одну песчинку, то система начнет осыпаться, покатится лавина песка. Для каждой такой точки x, y системы, область, затронутая осыпанием, будет различна. На рис. 60 представлены несколько таких лавин, полученных осыпанием. Исходной послужила система, изображенная на рис. 59. Центры лавин отмечены на фоне белых снежных лавин черными точками, а пары цифр в скобках обозначают время осыпания и размер лавины.

Определим $D(s)$ - функцию распределения размеров возникающих лавин. Чтобы получить эту функцию, в каждой точке системы, где $z(x,y)=3$ положим $z(x,y)=4$ и запустим сход лавины, определим ее площадь - s, для получения достаточного количества лавин обработаем таким образом несколько систем в самоорганизованном состоянии, получая их из случайного начального состояния с $z(x,y)\gg 4$. На рис. 61 *а*) представлен вид зависимости $D(s)$, полученный обработкой набора систем размера 500 x 500.

Распределение размеров лавин подчиняется степенному закону:

$D(s)\sim s^{-\tau}, \quad \tau_{2D}\approx 1{,}1.$

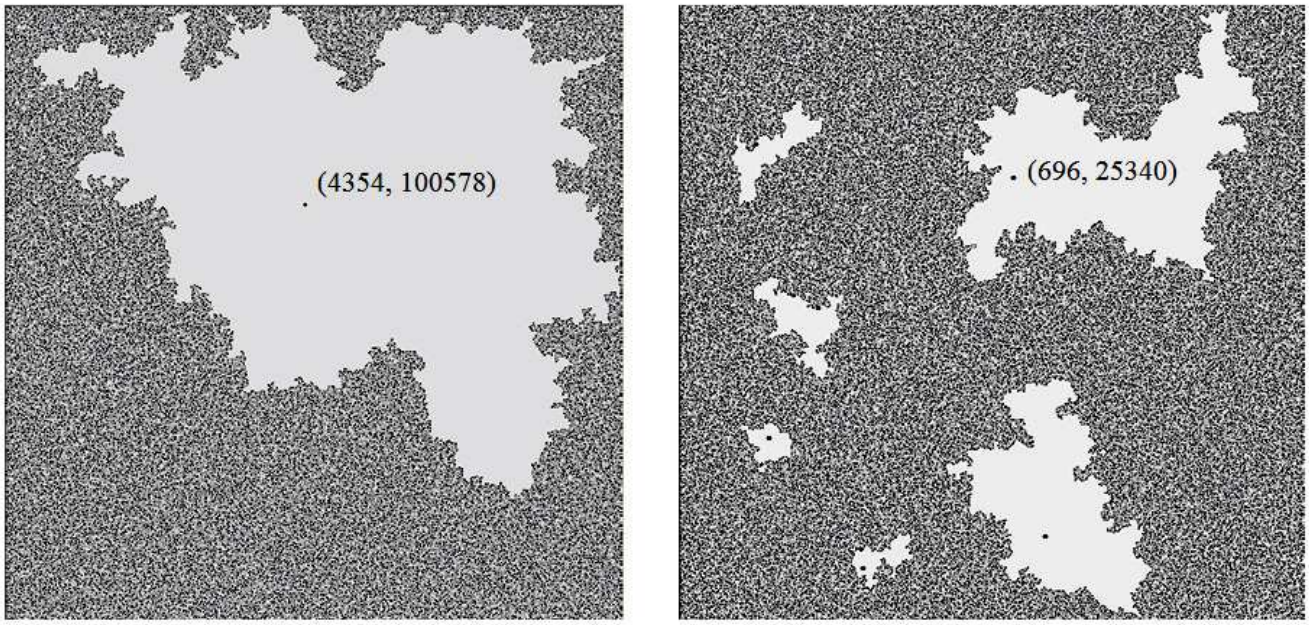


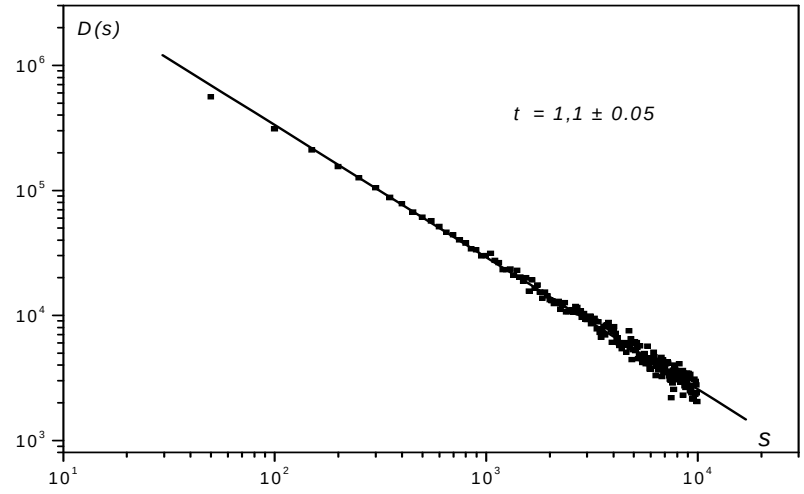
Рис. 60. Лавины, полученные осыпанием

Аналогично возможно исследовать и временные характеристики этого процесса, введя $D(t)$ - функцию распределения времен осыпания этих лавин. В общем случае площадь s лавины больше, чем время ее осыпания t, поскольку в один момент осыпаются несколько песчинок. Распределение развития лавин также подчиняется степенному закону:

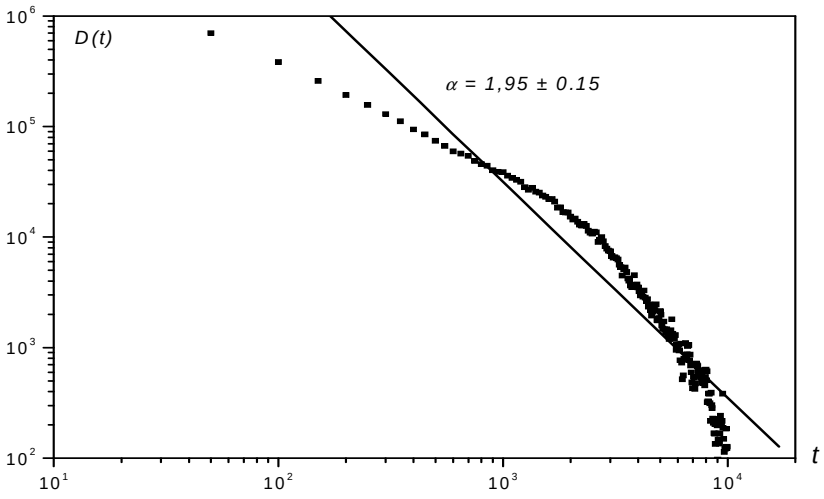
$$D(t) \sim t^{-\alpha}, \quad \alpha_{2D} \approx 0,43.$$

Безусловно, индексы τ и α связаны как друг с другом, так и с другими индексами, характеризующими самморганизованное критическое состояние.

Размер «лавины» новостей, возникающих в информационных потоках при появлении новой темы порой кажется непредсказуемым, однако вполне поддается моделированию. Степенные распределения количества тематических публикаций, которые будут приведены в следующей главе (см. п. 11.3), соответствуют приведенным выше распределениям размеров рассматриваемых лавин «песка». В последнее десятилетие моделирование информационных процессов с помощью методов клеточных автоматов и теории самоорганизованной критичности получили большое распространение.



а)



б)

Рис. 60. Распределения (а) размеров лавин D(s) и (б) времен их осыпаний D(t)

11. ЭЛЕМЕНТЫ ФРАКТАЛЬНОГО АНАЛИЗА

«Порядок творенья обманчив,
Как сказка с хорошим концом.»
Борис Пастернак

11.1. Фракталы и фрактальная размерность

Термин фрактал был предложен Бенуа Мандельбротом (B. Mandelbrot) в 1975 году для обозначения нерегулярных самоподобных математических структур. Основное определение фрактала, данное Мандельбротом, звучало так: "Фракталом называется структура, которая состоит из частей, которые в каком-то смысле подобны целому" [37]. Следует признать, что это определение, ввиду своей нестрогости, не всегда верно. Можно привести много примеров самоподобных объектов, не являющихся фракталами, например, сходящиеся к горизонту железнодорожные пути.

Бенуа Мандельброт

В самом простом случае небольшая часть фрактала содержит информацию обо всем фрактале. Строгое определение самоподобных множеств было дано Дж. Хатчинсоном (J. Hutchinson) в 1981 году. Он назвал множество самоподобным, если оно состоит из нескольких компонент, подобных всему

этому множеству, т.е. компонент получаемых афинными преобразованиями - поворотом, сжатием и отражением исходного множества.

Однако самоподобие – это хотя и необходимое, но далеко не достаточное свойство фракталов. Ведь нельзя же, в самом деле, считать фракталом точку, или плоскость, расчерченную клетками. Главная особенность фрактальных объектов состоит в том, что для их описания недостаточно «стандартной» топологической размерности d_τ, которая, как известно, для линии равна 1 ($d_\tau = 1$ - линия одномерный объект), для поверхности $d_\tau = 2$, и т.д. Фракталам характерна геометрическая «изрезанность». Поэтому используется специальное понятие фрактальной размерности, введенное Ф. Хаусдорфом (F. Hausdorf) и А.С. Безиковичем. Применительно к идеальным объектам классической евклидовой геометрии она давала те же численные значения, что и топологическая размерность, однако новая размерность обладала более тонкой чувствительностью ко всякого рода несовершенствам реальных объектов, позволяя различать и индивидуализировать то, что прежде было безлико и неразличимо. Размерность Хаусдорфа - Безиковича как раз и позволяет измерять степень «изрезанности». Размерность фрактальных объектов не является целым числом, характерным для привычных геометрических. Вместе с тем, в большинстве случаев фракталы напоминают объекты, плотно занимающие реальное пространство, но не использующие его полностью.

Феликс Хаусдорф (1868-1942)

Пусть есть множество G в евклидовом пространстве размерности d_τ. Это множество покрывается кубиками размерности d_τ, при этом длина ребра любого кубика не превышает некоторого значения δ, т.е. $\delta_i < \delta$.

Вводится зависящая от некоторого параметра d и δ сумма по всем элементам покрытия:

$$l_d(\delta) = \sum_i \delta_i^d .$$

Определим нижнюю грань данной суммы:

$$L_d(\delta) = \inf \sum_{i,\ \delta_i < \delta} \delta_i^d .$$

При уменьшении максимальной длины δ, если параметр d будет достаточно велик, очевидно, будет выполняться:

$$\lim_{\delta \to 0} L_d(\delta) \to 0.$$

При некотором достаточно малом значении параметра d будет выполняться:

$$\lim_{\delta \to 0} L_d(\delta) \to \infty.$$

Промежуточное, критическое значение d_x , для которого выполняется:

$$\lim_{\delta \to 0} L_d(\delta) = \begin{cases} 0, & d > d_x, \\ \infty, & d < d_x, \end{cases}$$

и называется размерностью Хаусдорфа-Безиковича (или фрактальной размерностью). Для простых геометрических объектов размерность Хаусдорфа-Безиковича совпадает с топологической (для отрезка d_x=1, для квадрата d_x=2, для куба d_x=3 и т.д.)

Несмотря на то, что размерность Хаусдорфа-Безиковича с теоретической точки зрения определена безупречно, для реальных фрактальных объектов расчет этой размерности является весьма затруднительным. Поэтому вводится несколько упрощенный показатель - емкостная размерность d_c. При определении этой размерности используются кубики с гранями одинакового размера. В этом случае, естественно, справедливо:

$$L_d(\delta) = N(\delta)\delta^{d_c},$$

где $N(\delta)$ - количество кубиков, покрывающего область G. Путем логарифмирования и перехода к пределу при уменьшении грани кубика ($\delta \to 0$) получаем:

$$d_c = -\lim_{\delta \to 0} \frac{\log N(\delta)}{\log \delta},$$

если этот предел существует. Следует отметить, что в большинстве численных методов определения фрактальной размерности используется именно d_c, при этом необходимо учитывать, что всегда справедливо условие: $d_x \le d_c$. Для регулярных самоподобных фракталов емкостная размерность и размерность Хаусдорфа-Безиковича совпадают, поэтому терминологически их часто не различают и говорят просто о фрактальной размерности объекта [13].

При проведении практических вычислений фрактальной размерности для реальных объектов используют следующий методический прием. Пусть на некотором этапе покрытия фрактала пришлось использовать $N(\delta)$ кубиков с гранями размера δ, а на другом – $N(\delta')$ элементов с гранями размера δ'. Ввиду предполагаемой степенной зависимости справедливо:

$$N(\delta) \sim \frac{1}{\delta^{d_c}}, \qquad N(\delta') \sim \frac{1}{\delta'^{d_c}},$$

откуда значение d_c может оцениваться как:

$$d_c = -\frac{\log\left(N(\delta)/N(\delta')\right)}{\log\left(\delta/\delta'\right)}.$$

11.2. Абстрактные фракталы

Рассмотрим принципы формирования нескольких абстрактных фрактальных объектов, которые обладают выраженными свойствами самоподобия.

Построение фрактального множества, снежинки Хельге фон Коха (H. Von Koch) (рис. 62), начинается с правильного треугольника, длина стороны которого равна 1. Сторона треугольника считается базовым звеном. Далее, на любом шаге итерации каждое звено заменяется на образующий элемент – ломаную, которая состоит по краям исходного звена из отрезков длиной 1/3 от

длины этого звена, между которыми размещаются две стороны правильного треугольника со стороной также в 1/3 длины звена. Все отрезки - стороны полученной ломаной считаются базовыми звеньями для следующей итерации. Кривая, получаемая в результате *n*-ой итерации при любом конечном *n*, называется предфракталом, и лишь при *n*, стремящимся к бесконечности, кривая Коха становится фракталом. Полученное в результате итерационного процесса фрактальное множество представляет собой линию бесконечной длины, которая ограничивает конечную площадь. Действительно, при каждом шаге число сторон результирующего многоугольника увеличивается в 4 раза, а длина каждой стороны уменьшается только в 3 раза, т.е. длина многоугольника на *n*-ой итерации равна $3\times\left(4/3\right)^n$ и стремится к бесконечности с ростом n.

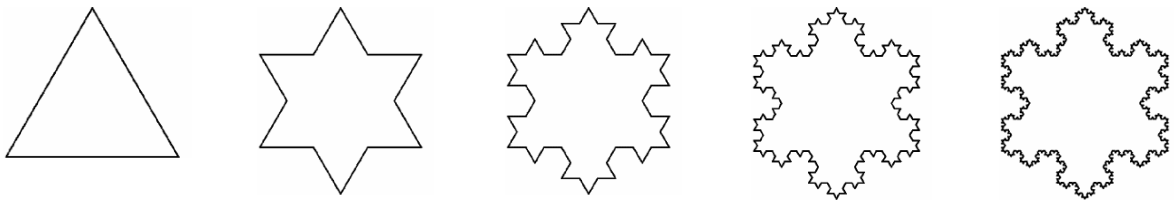


Рис. 62. Первые 5 поколений снежинки Коха

Площадь под кривой, если принять площадь первого образующего треугольника за единицу, равна:

$$S = 1 + \frac{1}{3}\sum_{k=0}^{\infty}\left(\frac{4}{9}\right)^k = 1,6.$$

Подсчитаем фрактальную размерность снежинки Коха. Пусть длина стороны исходного треугольника равна единице. В данном случае роль кубиков, покрывающих рассматриваемую фигуру играют отрезки прямой. Тогда на нулевом шаге имеем: $\delta = 1,\ N(\delta) = 3$. Для второго шага справедливо: $\delta' = 1/3,\ N(\delta') = 12$. Этих данных достаточно для оценки фрактальной размерности:

$$d_c = -\frac{\log\left(N(\delta)/N(\delta')\right)}{\log\left(\delta/\delta'\right)} = -\frac{\log\left(3/12\right)}{\log\left(3\right)} = \frac{\log 4}{\log 3} \sim 1,26.$$

Самоподобный фрактал, предложенный в 1915 г. В. Серпинским (V. Serpinski), формируется по следующим правилам. Исходным множеством, соответствующим нулевому шагу, является равносторонний треугольник (рис. 63). Затем он разбивается на четыре области путем соединения серидины сторон исходного треугольника отрезками прямых. Затем удаляется внутренность центральной области исходного треугольника – малый внутренний «перевернутый треугольник». Затем, на следующем шаге итерации, этот процесс повторяется для каждого из трех оставшихся треугольников. Продолжая описанную процедуру до бесконечности, образуется множество, называемое салфеткой Серпинского.

Очевидно, фрактальная размерность салфетки Серпинского составляет:

$$d_c = -\frac{\log\left(N(\delta)/N(\delta')\right)}{\log\left(\delta/\delta'\right)} = \frac{\log 3}{\log 2} \sim 1{,}58.$$

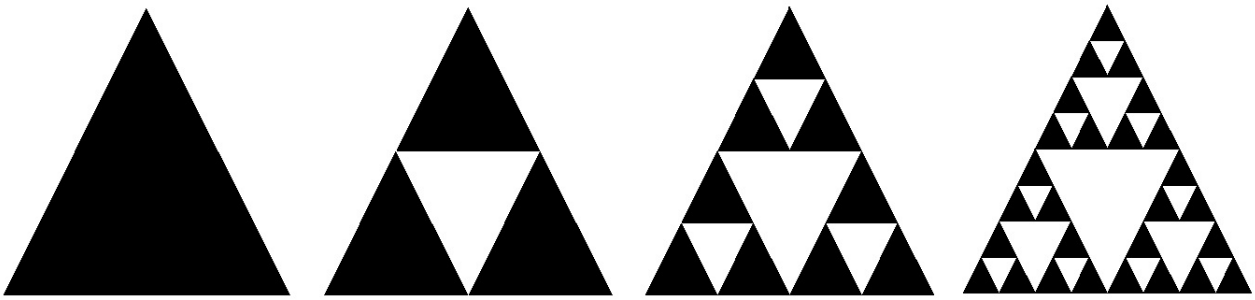


Рис. 63. Первые шаги алгоритма построения салфетки Серпинского

Этот фрактал интересен тем, что занимаемая им площадь равна нулю. Для обоснования этого факта подсчитаем суммарную площадь частей, ислюченную при построении. На первом шаге выбрасывается четвертая часть площади исходного треугольника, на втором шаге у каждого из трех треугольников удаляется четвертая часть площади и т.д. Таким образом, полная удаленная площадь вычисляется как сумма ряда:

$$S = \frac{1}{4} + \frac{3}{4}\cdot\frac{1}{4} + \frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{4} + \cdots = \frac{1}{4}\cdot\left[1 + \left(\frac{3}{4}\right) + \left(\frac{3}{4}\right)^2 + \left(\frac{3}{4}\right)^3 + \cdots\right] = \frac{1}{4}\cdot\frac{1}{1-3/4} = 1.$$

Таким образом, исключенная площадь совпадает с площадью исходного треугольника.

Алгоритм построения множества Мандельброта (рис. 64) основан на итеративном вычислении по формуле:

$$Z_{i+1} = Z_i^2 + C, \quad i = 0,\ 1,\ 2,\ ...$$

где Z_{i+1}, Z_i и C - комплексные переменные.

Итерации выполняются для каждой стартовой точки C прямоугольной или квадратной области – подмножестве комплексной плоскости. Итерационный процесс длится до тех пор, пока Z_i не выйдет за пределы окружности заданного радиуса, центр которой лежит в точке (0,0), или после достаточно большого количества итераций. В зависимости от количества итераций, в течение которых Z_i остается внутри окружности, устанавливаются цвета точки C. Если Z_i остается внутри окружности в течение достаточно большого количества итераций, то эта точка растра окрашивается в черный цвет. Множеству Мандельброта принадлежат именно те точки, которые в течение бесконечного числа итераций не уходят в бесконечность. Так как количество итераций соответствует номеру цвета, то точки, которые находятся ближе к множеству Мандельброта, имеют более яркую окраску.

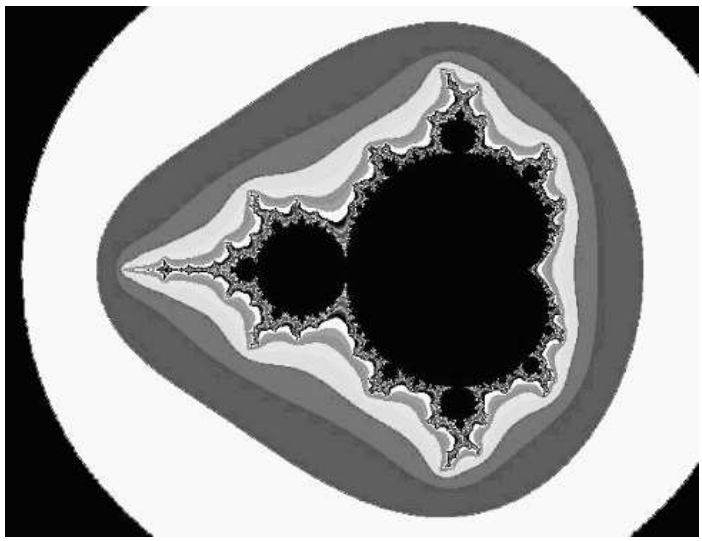


Рис. 64. Множество Мандельброта

В 80-х годах прошлого столетия появился метод "Систем Итерационных Функций" (Iterated Functions System - IFS), представляющий собой простое средство получения фрактальных структур. IFS представляет собой систему

функций, которые отображают одно многомерное множество на другое. Наиболее простая реализация IFS представляет собой афинные преобразования плоскости:

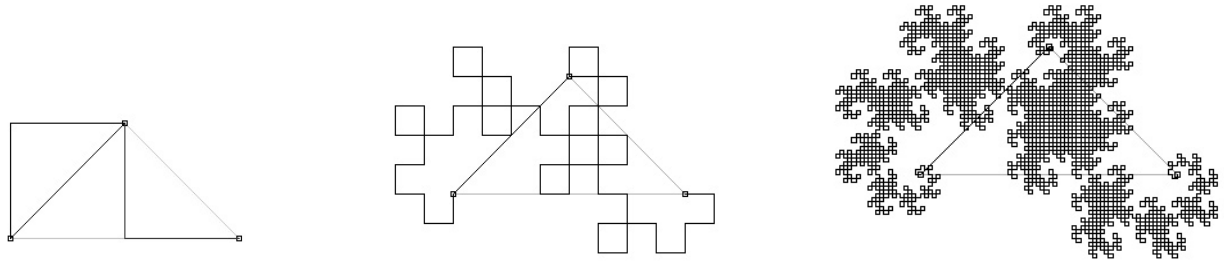
$$X' = AX + BY + C,$$

$$Y' = DX + EY + F,$$

где X, Y - предыдущие значения координат, X', Y' - новые значения, A, B, C, D, E и F - коэффициенты.

В качестве примера использования IFS для построения фрактальных структур, можно привести "дракона" Хартера-Хейтуея, сформированного с помощью Java-апплета, приведенного в Интернет по адресу http://www.fractals.nsu.ru/fractals.chat.ru/ifs2.htm (рис. 65). Использование IFS для сжатия обычных изображений, например, фотографии основаны на выявлении локального самоподобия (в отличие от фракталов, где наблюдается глобальное самоподобие).

В 80-х годах М. Барнсли (M. Barnsley) и А. Слоан (A. Sloane) предложили идею сжатия и хранения графической информации, основанную на соображениях теории фракталов и динамических систем. На основе этой идеи был создан алгоритм фрактального сжатия информации, который позволяет сжимать некоторые образцы графической информации в 500-1000 раз. При этом каждое изображение кодируется несколькими простыми афинными преобразованиями.



а) б) в)

Рис. 65. "Дракон" Хартера-Хейтуея: а) – второй шаг итерации; б) – шестой шаг; в) – двенадцатый шаг

По алгоритму Барнсли в изображении происходит выделение пар областей, меньшая из которых подобная большей, и сохранение нескольких коэффициентов, которые кодируют преобразование, переводящее большую область в меньшую. При этом требуется, чтобы множество таких областей покрывало все изображение.

Один из лучших примеров проявления фракталов в природе - структура береговых линий. Действительно, на километровом отрезке побережье выглядит настолько же порезанным, как и на стокилометровом. Опыт показывает, что длина береговой линии L зависит от масштаба l, которым проводятся измерения, и увеличивается с уменьшением последнего по степенному закону $L = \Lambda l^{1-\alpha}$, $\Lambda = const$. Так, например, фрактальная размерность береговой линии Великобритании (рис. 66) составляет $\alpha \approx 1.24$.

11.3. Информационное пространство и фракталы

Под информационным пространством принято понимать совокупность информационных ресурсов, технологий их сопровождения и использования, информационных и телекоммуникационных систем, образующих некую информационную инфраструктуру. В данном случае элементами информационного пространства могут быть документы, обобщающие самые различные виды информации – файлы, электронные письма, веб-страницы не зависимо от форматов их представления.

Естественно, приведенное определение информационного пространства является качественным. Конечно же, термин «пространство» в данном случае, вообще говоря, не совпадает с понятием «пространство» в математике или физике. В качестве примеров удачных моделей информационного пространства можно приветсти «векторно-пространственную» модель Г. Солтона [131] или модель старения информации Бартона-Кеблера [31]. Модель такого информационного пространства, как сеть WWW была построена А. Брёдером и его соавторами из компаний IBM и Altavista [82].

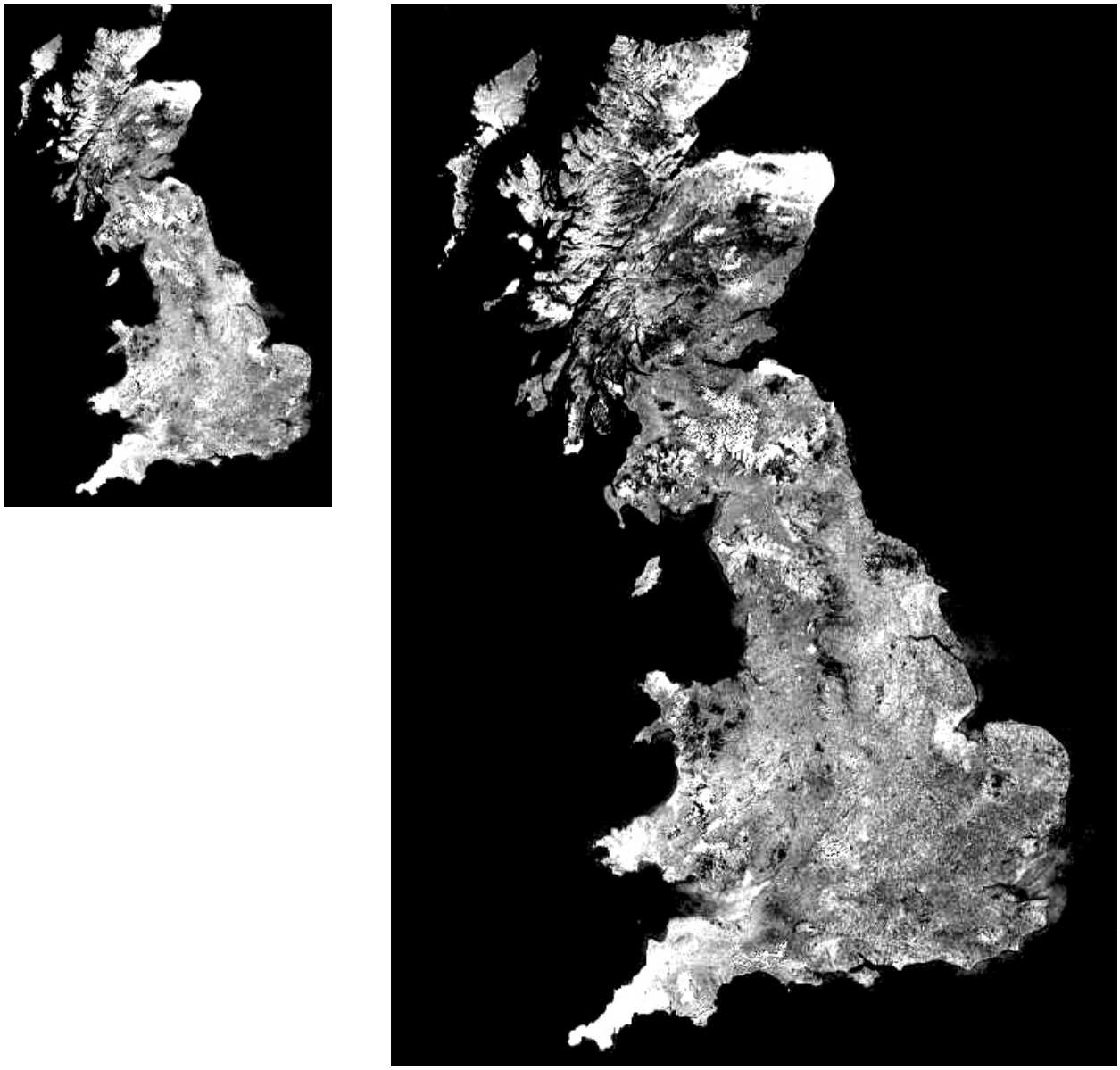


Рис. 66. Побережье Великобритании в разных масштабах (http://maps.google.com)

Во многих моделях информационного пространства изучаются структурные связи между тематическими множествами его элементов - документами. Многочисленные эксперименты, замеры параметров информационного пространства подтверждают тот факт, что при значительном возрастании объемов информационных ресурсов статистические распределения документов, получаемые в самых разнообразных содержательных разрезах (таких, например, как источники, авторы, тематики) практически не меняют своей формы [22, 23].

Применение теории фракталов при анализе информационного пространства позволяет с общей позиции взглянуть на закономерности, которые составляют

основы информатики [145, 20]. Известно, что многие информационно-поисковые системы, включающие элементы кластерного анализа, позволяют автоматически обнаруживать новые классы и распределяют документы по этим классам. Соответственно, показано, что тематические информационные массивы представляют собой самоподобные развивающиеся структуры, однако их самоподобие справедливо лишь на статистическом уровне (например, распределение тематических кластеров документов по размерам).

Чем же определяется природа фрактальных свойств информационного пространства, порождаемого такими кластерными структурами? С одной стороны, параметрами ранговых распределений, а с другой стороны, механизмом развития информационных кластеров. Появление новых публикаций увеличивает размеры уже существующих кластеров и является причиной образования новых.

Фрактальные свойства характерны и для кластеров информационных веб-сайтов, на которых публикуются документы, соответствующие определенным тематикам. Эти кластеры, как наборы тематических документов, представляют собой структуры, обладающие рядом уникальных свойств.

Свойства самоподобия фрагментов информационного пространства наглядно демонстрирует интерфейс, представленный на веб-сайте службы News Is Free (http://newsisfree.com). На этом сайте отображается состояние информационного пространства в виде ссылок на источники и отдельные сообщения. В этом интерфейсе можно наглядно наблюдать «дробление» групп источников при увеличении ранга популярности и «свежести» изданий (рис. 67).

Топология и характеристики моделей веб-пространства оказываются приблизительно одинаковыми для его разных подмножеств, подтверждая тем самым наблюдение о том, что «веб - это фрактал».

Как показано в роботах С. Иванова [20-23], для последовательности сообщений тематических информационных потоков количество сообщений, резонансов на события реального мира, пропорционально некоторой степени количества источников информации (кластеров).

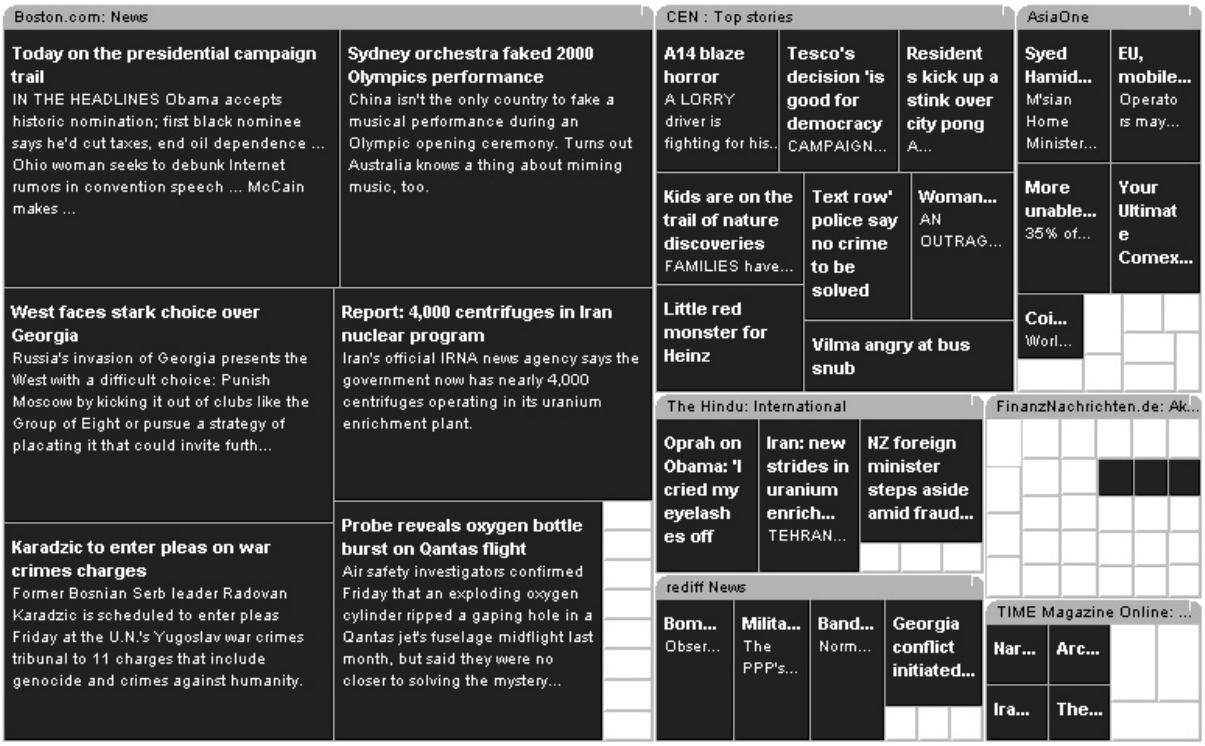


Рис. 67. Дробление групп источников при изменении ранга их популярности (http://newsisfree.com)

Известно, что все основные законы научной коммуникации, такие как законы Парето, Лотки, Бредфорда, Ципфа, могут быть обобщенные именно в рамках теории стохастических фракталов. Точно так же, как и в традиционных научных коммуникациях, множество сообщений в Интернете по одной тематике во времени представляет собой динамическую кластерную систему, которая возникает в результате итерационных процессов. Этот процесс обуславливается републикациями, односторонним или взаимным цитированием, различными публикациями - отражениями одних и тех же событий реального мира, прямыми ссылками и т.п.

Фрактальная размерность в кластерной системе, которая соответствует тематическим информационным потокам, показывает уровень заполнения информационного пространства сообщениями на протяжении определенного времени [20]:

$$N_{publ}(\varepsilon t) = \varepsilon^{\rho} N_k(t)^{\rho},$$

где N_{publ} - размер системы (общее количество сообщений в информационном потоке); N_k – размер – число кластеров (тематик или источников); ρ – фрактальная размерность информационного массива; ε – коэффициент масштабирования. В приведенном соотношении между количеством сообщений и кластеров проявляется свойство сохранения внутренней структуры множества при изменении масштабов его внешнего рассмотрения.

11.4. Фракталы и временные ряды

Объемы сообщений в тематических информационных потоках образуют временные ряды (например, ряд, состоящий из значений количества публикаций в отдельные дни, рис. 67). Для исследования временных рядов сегодня все шире используется теория фракталов.

Изучение характеристик временных рядов, порождаемых информационными потоками, сообщения которых отражают процессы, происходящие в реальном мире, дает возможность прогнозировать их динамику, выявлять скрытые корреляции, циклы и т.п.

В этом разделе будут описаны основные алгоритмы, применяемые при исследовании фрактальных свойств рядов измерений. В качестве иллюстраций приведены результаты реальных численных экспериментов. Как база для исследования фрактальных свойств рядов, отражающих интенсивность публикаций тематических информационных потоков, использовалась система контент-мониторинга новостей с веб-сайтов сети Интернет InfoStream. Тематика исследуемого информационного потока определялась запросом к этой системе. Данные для исследований были получены из интерфейса режима «Динамика появления понятий».

В ходе исследований обрабатывался информационный корпус, содержащий сообщения онлайновых СМИ - массив из 14069 документов, опубликованных с 1 января 2006 г. по 31 декабря 2007 г., по тематике компьютерной вирусологии, удовлетворяющих запросу:

«компьютерный вирус» OR «вирусная атака» OR (антивирус AND (программа OR утилита OR Windows OR Linux)).

Ниже анализируетсятся временной ряд из количества тематических публикаций за указанный период с определенной дискретностью по времени в сутки (рис. 68).

Остановимся подробнее на некоторых методах анализа подобного типа временных рядов, порождаемых, в частности, информационными потоками.

11.4.1. Метод DFA

Один из универсальных подходов к выявлению самоподобия основывается на методе DFA (Detrended Fluctuation Analysis) [121] – универсальном методе обработки рядов измерений.

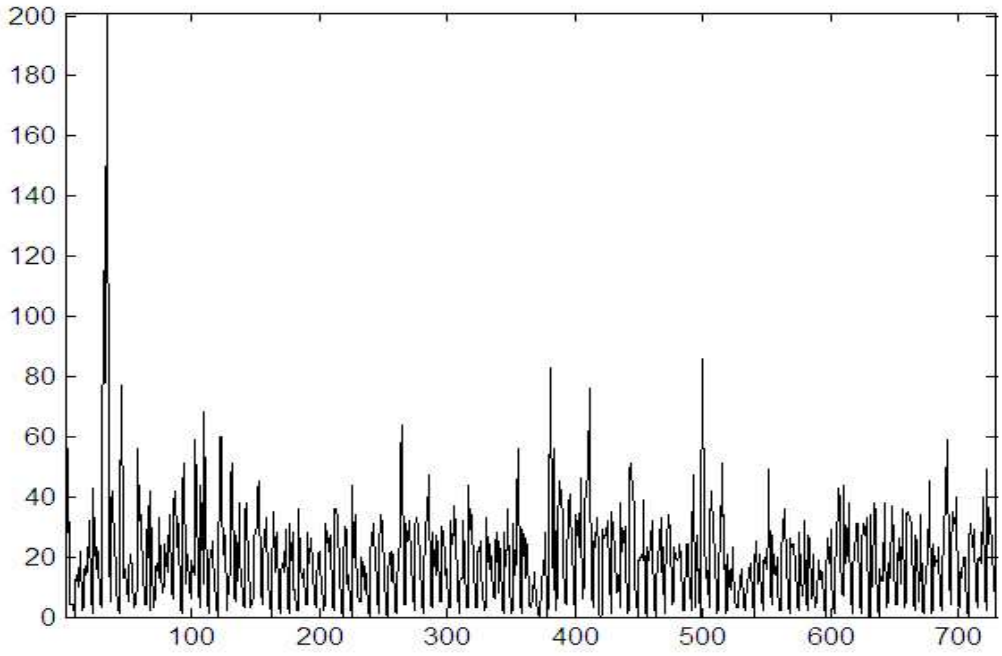


Рис. 68. Количество тематических публикаций (ось ординат) в разрезе дат (ось абсцисс)

Метод DFA представляет собой вариант дисперсионного анализа, который позволяет исследовать эффекты продолжительных корреляций в нестационарных рядах. При этом анализируется среднеквадратичная ошибка линейной аппроксимации в зависимости от размера отрезка аппроксимации. В рамках этого

метода сначала осуществляется приведение данных к нулевому среднему (вычитание среднего значения $\langle F \rangle$ из временного ряда F_n, $n = 1, ..., N$) и строится случайное блуждание $y(k)$:

$$y(k) = \sum_{i=1}^{k} \left[F_i - \langle F \rangle_N \right].$$

Потом ряд значений $y(k)$, $k = 1, ..., N$ разбивается на неперекрывающиеся отрезки длины *n*, в пределах каждого из которых методом наименьших квадратов определяется уравнение прямой, аппроксимирующей последовательность $y(k)$. Найденная аппроксимация $y_n(k)$ ($y_n(k) = ak + b$) рассматривается как локальный тренд.

Далее вычисляется среднеквадратичная ошибка линейной аппроксимации $D(n)$ при широком диапазоне значений *n*:

$$D(n) = \sqrt{\frac{1}{N} \sum_{k=1}^{N} \left[y(k) - y_n(k) \right]^2}.$$

В случае, когда зависимость $D(n)$ имеет степенной характер $D(n) \sim n^{\alpha}$, т.е. наличия линейного участка при двойном логарифмическом масштабе $\ln D \sim \alpha \ln n$, можно говорить о существовании скейлинга.

Как видно по рис. 69, значения $D(n)$ для выбранного информационного потока степенным образом зависят от n, т.е. в двойном логарифмическом масштабе эта зависимость близка к линейной.

11.4.2. Корреляциионный анализ

Если обозначить через X_t член ряда количества публикаций (количества электронных сообщений, поступивших, например, в день t, $t = 1, ..., N$), то функция автокорреляции для этого ряда *X* определяется как:

$$F(k) = \frac{1}{N-k} \sum_{t=1}^{N-k} (X_{k+t} - m)(X_t - m),$$

где m – среднее значение ряда *X*, которое в дальнейшем, не ограничивая общности, будем считать равным 0 (это достигается переприсвоением значению

X_t значения X_t - m). Предполагается, что ряд X может содержать скрытую периодическую составляющую.

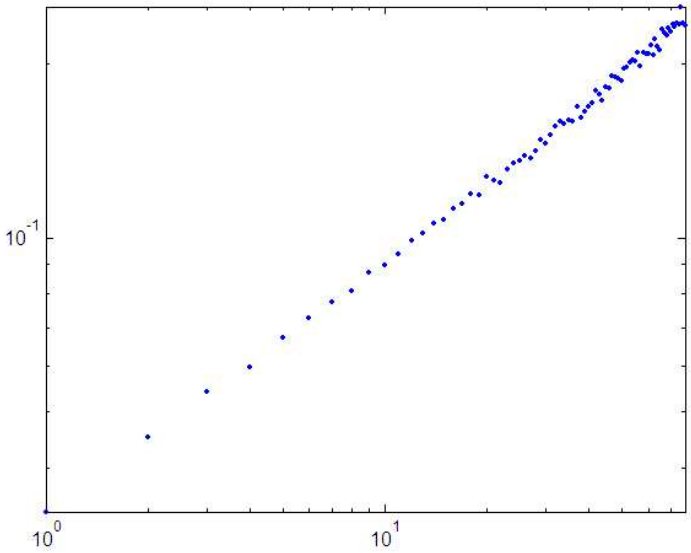


Рис. 69. Зависимость D(n) ряда наблюдений (ось ординат) от длины отрезка аппроксимации n (ось абсцисс) в логарифмической шкале

Известно, что функция автокорреляции обладает тем свойством, что если скрытая периодическая составляющая существует, то ее значение асимптотически приближается к квадрату среднего значения исходного ряда .

Если рассматриваемый ряд периодический, т.е. может быть представлен как:

$$X_t = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n\omega t + \theta_n),$$

то его функция автокорреляции будет равна:

$$F(k) = \frac{a_0^2}{4} + \frac{1}{2}\sum_{n=1}^{\infty} a_n^2 \cos n\omega k.$$

Этот результат [59] показывает, что функция автокорреляции периодического ряда также является периодической, содержит основную частоту и гармоники, но без фазовых углов θ_n.

Рассмотрим числовой ряд X, являющийся суммой некоторой содержательной составляющей N и синусоидальной сигнала S:

$$X_t = N_t + S_t.$$

Найдем функцию автокорреляции для этого ряда (значения приведены к среднему $m = 0$):

$$F(k) = \frac{1}{N-k}\sum_{t=1}^{N-k} X_{k+t}X_t =$$

$$= \frac{1}{N-k}\sum_{t=1}^{N-k}(N_{k+t} + S_{k+t})(N_t + S_t) =$$

$$= \frac{1}{N-k}\sum_{t=1}^{N-k} N_{k+t}N_t + \frac{1}{N-k}\sum_{t=1}^{N-k} S_{k+t}S_t + \frac{1}{N-k}\sum_{t=1}^{N-k} N_{k+t}S_t + \frac{1}{N-k}\sum_{t=1}^{N-k} S_{k+t}N_t.$$

Очевидно, первое слагаемое есть функция непериодическая, асимптотически стремящаяся к нулю. Так как взаимная корреляция между N и S отсутствует, то то третье и четвертое слагаемое также стремятся к нулю. Таким образом, самый значительный ненулевой вклад составляет второе слагаемое – автокорреляция сигнала S. Т.е. функция автокорреляции ряда X остается периодической.

Для экспериментального подтверждения рассмотренной гипотезы была сгенерирована последовательность, по своей природе напоминающая реальный информационный поток. Предполагалось, что ежедневное количество сообщений в сети растет по экспоненциальному закону (с очень небольшим значением экспоненциальной степени), и на это количество накладываются колебания, связанные с недельной цикличностью в работе информационных источников. Также принимается во внимание некоторый элемент случайности, выраженный соответствующими отклонениями.

Для получения соответствующего временного ряда были рассмотрены значения функции:

$$y = ae^{0.001x} + \sin(\pi x/7 + a),$$

которая реализует простейшую модель информационного потока – экспонента отвечает за рост количества публикаций во времени (общая тенденция), синус – за недельную периодичность, параметр a – за случайные отклонения. Количество публикаций y не может быть отрицательным числом. На рис. 70 представлен график модели (ось абсцисс – переменная x – день, ось ординат – переменная y – количество публикаций).

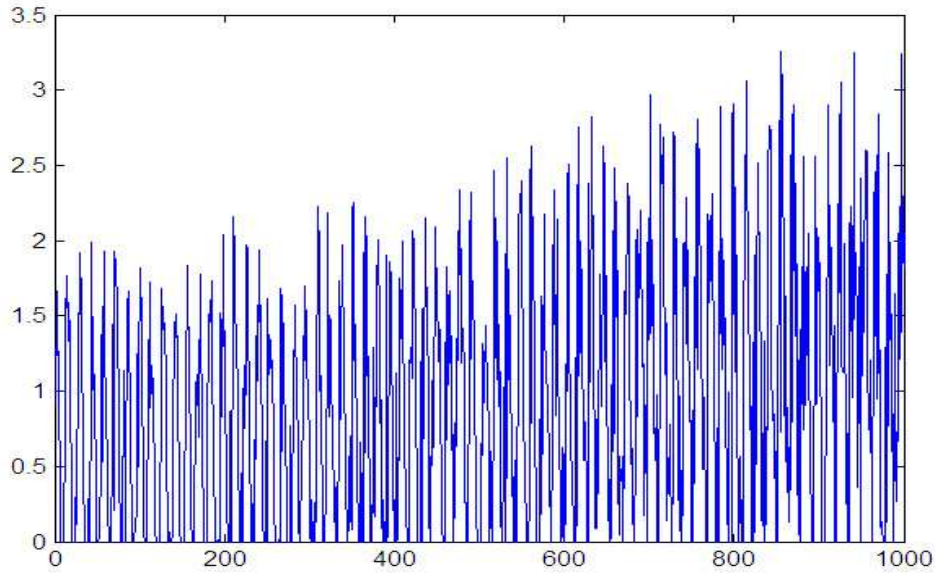


Рис. 70. Модель потока с экспоненциальным ростом

Исходный ряд был обработан: приведен к нулевому среднему и нормирован (каждый член разделен на среднее). После этого были рассчитаны коэффициенты корреляции, которые для рядов измерений X длиной N рассчитываются по формуле:

$$R(k) = \frac{F(k)}{\sigma^2},$$

где $F(k)$ – функция автокорреляции; σ^2 – дисперсия.

На рис. 71 приведен график значений коэффициентов корреляций (ось абсцисс – переменная *k*, ось ординат – коэффициент корреляции *R(k)*.

Графическое представление коэффициента корреляции для ряда наблюдений, соответствующего динамике реального информационного потока веб-публикаций свидетельствует о неизменности корреляционных свойств по дням недели (рис. 72). Вместе с тем коэффициенты корреляции ряда наблюдений, усредненного по неделям, аппроксимируются гиперболической функцией, которая характеризует долгосрочную зависимость членов исходного ряда (рис. 73).

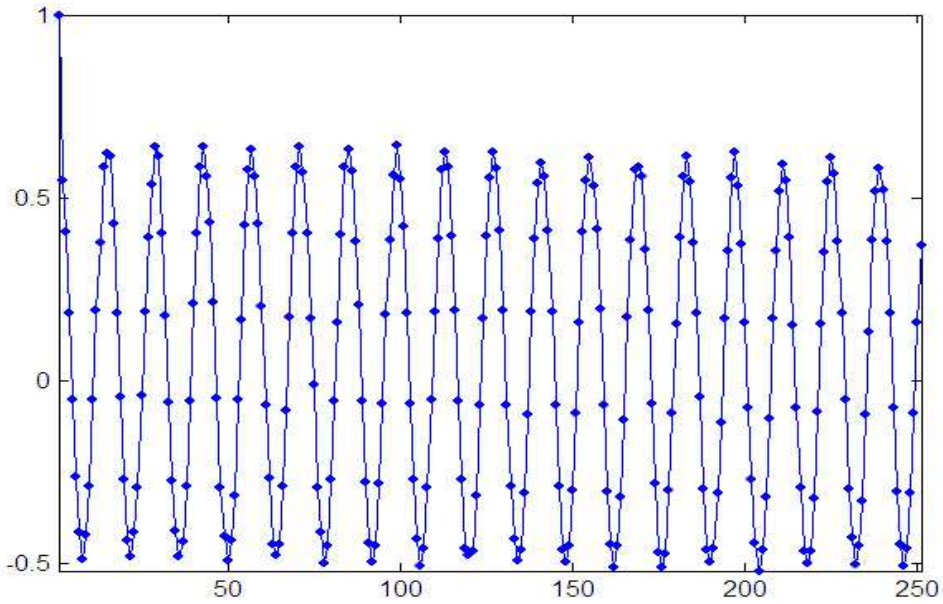


Рис. 71. Значения коэффициентов корреляции модели

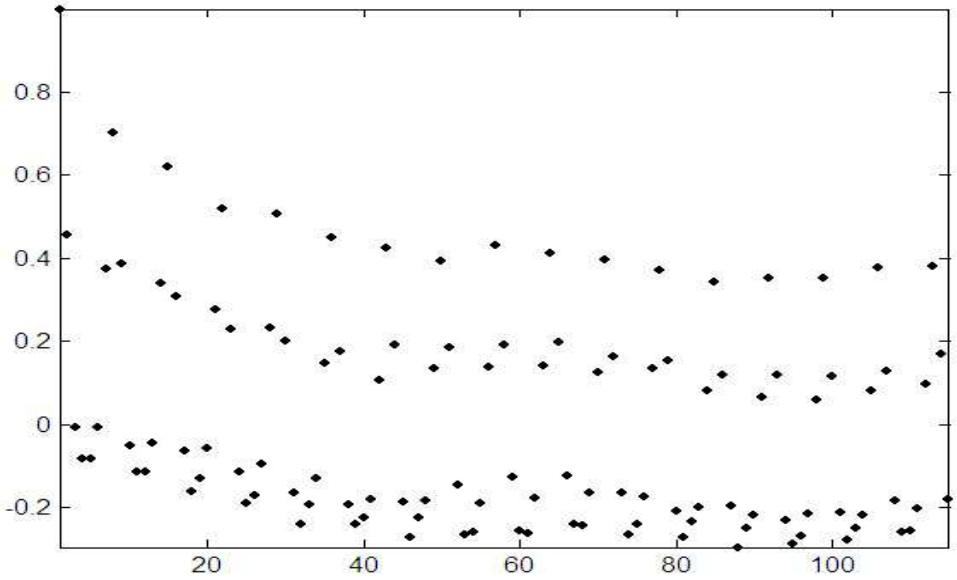


Рис. 72. Коэффициенты корреляции ряда наблюдений $R(k)$ *(ось ординат) в зависимости от* k *(ось абсцисс)*

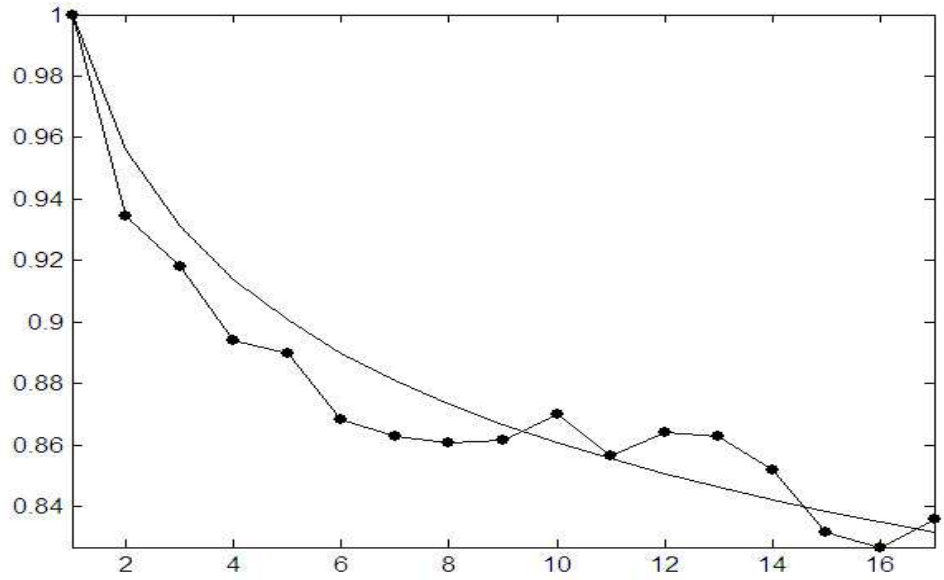


Рис. 73. Коэффициенты корреляции ряда наблюдений $R(k)$ (ось ординат), усредненного по неделям в зависимости от k (ось абсцисс)

11.4.3. Фактор Фано

Для изучения поведения процессов принято использовать еще один показатель – индекс разброса дисперсии (IDC), так называемый фактор Фано (U. Fano) [90]. Эта величина определяется как отношение дисперсии количества событий (в нашем случае – количества публикаций) на заданном окне наблюдений k к соответствующему математическому ожиданию:

$$F(k) = \sigma^2(k) / m(k).$$

Для самоподобных процессов выполняется соотношение:

$$F(k) = 1 + Ck^{2H-1},$$

где C и H – константы. На рис. 73 приведен график значений $F(k)$ в логарифмическом масштабе, при этом $C \approx 6.8$ и $H \approx 0.65$.

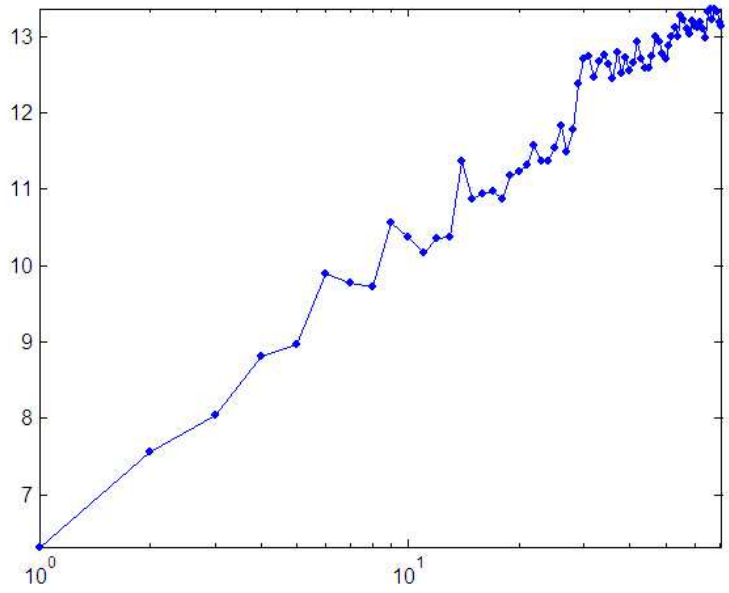


Рис. 74. Зависимость фактора Фано от ширины окна наблюдений

11.4.4. Показатель Херста

Показатель Херста (H.E. Hurst) - H связан с коэффициентом нормированного размаха R/S, где R - вычисляемый определенным образом «размах» соответствующего временного ряда, а S - стандартное отклонение [102]. Г.Э. Херст (1880 – 1978) экспериментально обнаружил, что для многих временных рядов справедливо: $R/S = (N/2)^H$. В [58] показано, что он связан с традиционной «клеточной» фрактальной размерностью D простым соотношением:

$$D = 2 - H.$$

Условие, при котором показатель Херста связан с фрактальной «клеточной» размерностью в соответствии с приведенной формулой, определено Е. Федером следующим образом: «… рассматривают клетки, размеры которых малы по сравнению как с длительностью процесса, так и с диапазоном изменения функции; поэтому соотношение справедливо, когда структура кривой, описывающая фрактальную функцию, исследуется с высоким разрешением, т.е. в локальном пределе». Еще одним важным условием является самоаффинность функции. Не вдаваясь в подробности, заметим, что для информационных потоков это свойство интерпретируется как самоподобие, возникающее в результате

процессов их формирования. Можно отметить, что указанными свойствами обладают не все информационные потоки, а лишь те, которые характеризуются достаточной мощностью и итеративностью при формировании. При этом временные ряды, построенные на основании мощных тематических информационных потоков, вполне удовлетворяют этому условию. Поэтому при расчете показателя Херста фактически определяется и такой показатель тематического информационного потока как фрактальная размерность.

Известно, что показатель Херста представляет собой меру персистентности - склонности процесса к трендам (в отличие от обычного броуновского движения). Значение $H > ½$ означает, что направленная в определенную сторону динамика процесса в прошлом, вероятнее всего, повлечет продолжение движения в том же направлении. Если $H < ½$, то прогнозируется, что процесс изменит направленность. $H = ½$ означает неопределенность — броуновское движение.

Для изучения фрактальных характеристик тематических информационных потоков за определенный период для временных рядов $F(n)$, $n = 1, ..., N$, составленных из количества относящихся к ним сообщений, изучалось значение показателя Херста, которое определялось из соотношения:

$$R/S = (N/2)^H, \quad N >> 1.$$

Здесь S – стандартное отклонение:

$$S = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(F(n) - \langle F \rangle_N\right)^2},$$

$$\langle F \rangle_N = \frac{1}{N}\sum_{n=1}^{N} F(n),$$

а R - так называемый размах:

$$R(N) = \max_{1 \le n \le N} X(n,N) - \min_{1 \le n \le N} X(n,N),$$

где

$$X(n,N) = \sum_{i=1}^{n}\left(F(i) - \langle F \rangle_N\right).$$

Исследования фрактальных свойств рядов измерений, получаемых в результате мониторинга тематических информационных массивов из Интернет, свидетельствуют о том, что при увеличении n показатель H принимает значения $0.65 \div 0.75$. Ввиду того, что значение H намного превышает ½, в этом ряду обнаруживается персистентность (существование долговременных корреляций, которые могут быть связаны с проявлением детерминированного хаоса). Если предположить, что ряд $F(n)$ является локально самоаффинным (этот вопрос в настоящее время открыт), то он имеет фрактальную размерность D, равную

$D = 2 - H \approx 1.35 \div 1.25$.

То есть, исследования тематических информационных потоков подтверждают предположение о самоподобии и итеративности процессов в веб-пространстве. Републикации, цитирование, прямые ссылки и т.п. порождают самоподобие, проявляющееся в устойчивых статистических распределениях и известных эмпирических законах.

В результате экспериментов было подтверждено наличие высокого уровня статистической корреляции в информационных потоках на продолжительных временных интервалах. На основе рассмотренного примера показана высокая персистентность процесса, что, в частности, свидетельствует об общей тенденции увеличения публикации по выбранной тематике.

Анализ самоподобия информационных массивов может рассматриваться как технология для осуществления прогнозирования.

11.5. Мультифрактальный анализ рядов измерений

Наиболее общее описание природы самоподобных объектов дает теория мультифракталов, характеризуемых бесконечной иерархией размерностей, и позволяющая отличить однородные объекты от неоднородных [58]. Концепция мультифрактального формализма [8, 15, 24, 42] дает эффективный инструмент для изучения и количественного описания широкого многообразия сложных систем.

Понятие мультифрактала можно пояснить с помощью специальной модели, описывающей процедуру распределения наследства (например, золота) между поколениями наследников. Данная модель реализует так называемое неоднородное множество Кантора (Y. Kantor), которое строится следующим образом. В начале предполагается, что все «золото» приписывается отрезку $[0,\ 1]$ - этому отрезку соответствует «предок» - персона, 100% наследства которой будет распределяться (рис. 75 а).

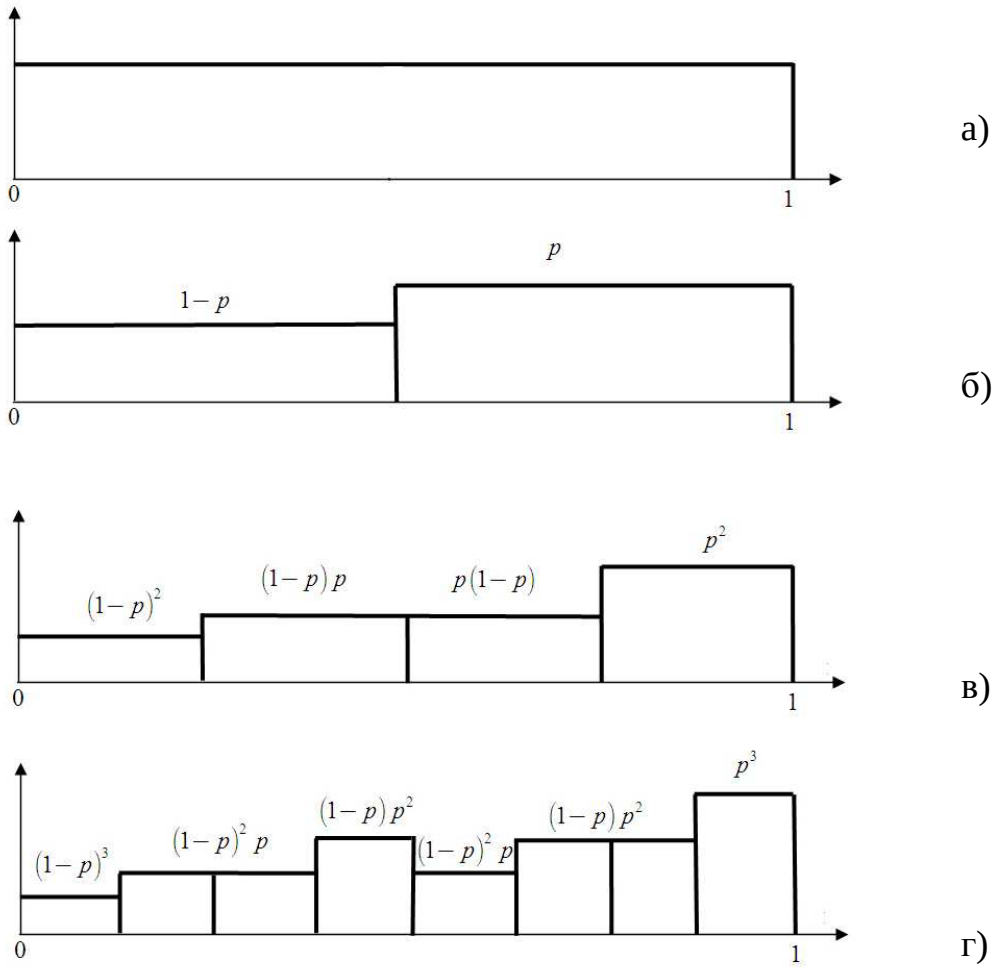


Рис. 75. Удельные части наследства (ось ординат): а – исходное состояние; б – первый шаг модели; в – второй шаг; г – третий шаг

На первой итерации наследство делится между двумя наследниками (в рамках всей этой модели у каждого «предка» имеется ровно по два наследника) на две неравные части – старший наследник получает p-ю часть, а младший - $(1-p)$-ю часть (эти пропорции фиксируются и для последующих поколений

наследников), при этом для определенности соблюдаются условия: $1 > p > (1-p) > 0$. Соответственно, отрезок $[0,\ 1]$ делится на две части - первый отрезок $[0,\ 0.5)$ соответствует младшему наследнику, а второй $[0.5,\ 1]$ - старшему (рис. 74 б).

Следующий, второй шаг, аналогичен первому. Каждый отрезок, соответствующий первому поколению наследников, снова делится на две равные (с точностью до одной средней точки) части, а наследники получают свои доли, старшие (правые отрезки) p-ю, младшие (левые отрезки) - $(1-p)$-ю. Соответственно, самый богатый наследник на этом шаге получает p^2-ю часть начального «золота», а самый бедный - $(1-p)^2$-ю (рис. 74 в).

Уже на третьем шаге (рис. 75 *г*) выясняется, что наследство распределяется достаточно сложным образом. Для анализа этого распределения удобно рассмотреть древовидную структуру, представленную на рис. 76.

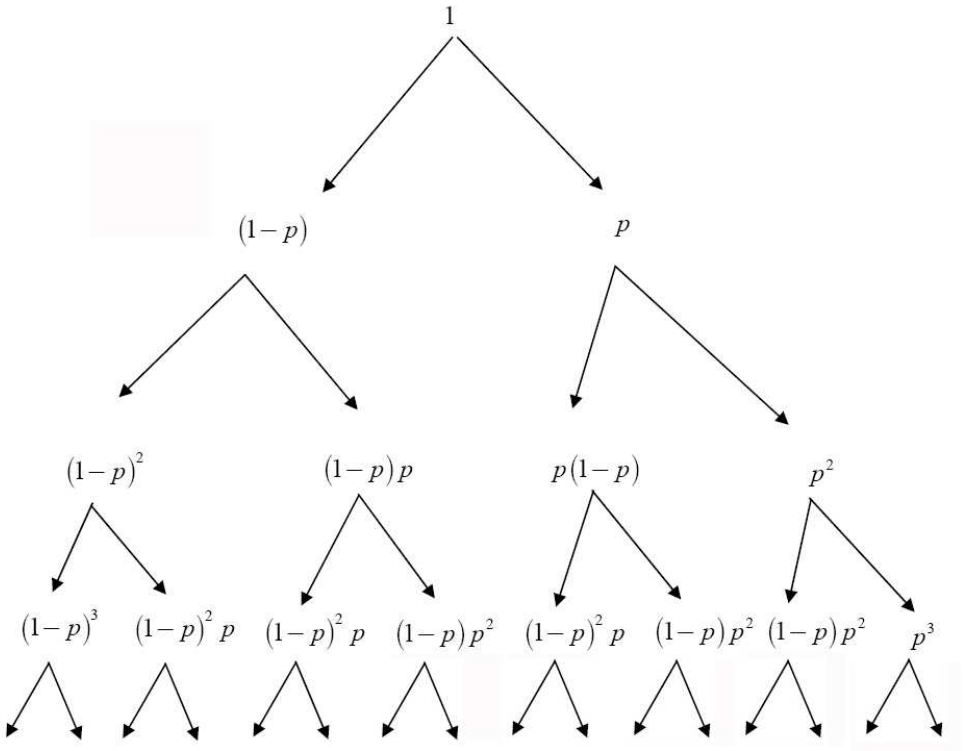


Рис. 76. Дерево пошагового распределения наследства

Несложно заметить существующую в этой модели связь между двоичной записью числа и величиной наследства. В самом деле, выберем на

рассматриваемом отрезке оси абсцисс некоторую точку x^*, например $x^*=1/5$, двоичное разложение которой следующее: $1/5\to 0.00110011...$ Каждый ноль в двоичной записи означает переход влево (L) по дереву, представленному на рис. 75, а единица – переход вправо (R). С одной стороны, выполняя эти переходы *LLRRLLRR...* мы приближаемся к $x^*=1/5$ в соответствии с процедурой, заданной рассматриваемой моделью, с другой стороны, каждый шаг вправо дает умножение величины наследства предыдущего поколения на p, а влево на $1-p$. Например, для шестого шага, для отрезка, внутри которого лежит $x^*=1/5$, доля первоначального наследства составляет $p^2(1-p)^4$, а для n-го шага:

$$x^*\to p^k(1-p)^{n-k},$$

где на n-ом шаге имеется k нулей и $n-k$ - единиц.

На этом же шаге имеется n отрезков размером $1/2^n$, из них $p^k(1-p)^{n-k}$ «золота» имеют $C_n^k=n!/(n-k)!k!$ отрезков (наследников). Последнее определяет количество способов, когда проходя по дереву, представленному на рис. 75, можно прийти к величине равной $p^k(1-p)^{n-k}$ (количество способов размещения k нулей и $n-k$ единиц).

Таким образом, полная вероятность встретить отрезок со значением $p^k(1-p)^{n-k}$ (или наследника с количеством «золота» $p^k(1-p)^{n-k}$) есть:

$$C_n^k p^k(1-p)^{n-k}.$$

При $n>>1$, используя формулу Стирлинга, можно дать следующую оценку величине C_n^k:

$$C_n^k\approx\left(\frac{k}{n}\right)^{-k}\left(1-\frac{k}{n}\right)^{k-n}=2^{nH\left(\frac{k}{n}\right)}=\left(\frac{1}{2^n}\right)^{-H\left(\frac{k}{n}\right)},$$

где

$$H(\xi)=-\xi\log_2\xi-(1-\xi)\log_2(1-\xi),\qquad \xi=k/n.$$

График функции $H(\xi)$ приведен на рис. 77.

Множество всех отрезков с заданным значением k/n является фрактальным (конечно, при $n \to \infty$) , для которого достаточно просто вычисляется фрактальная размерность D.

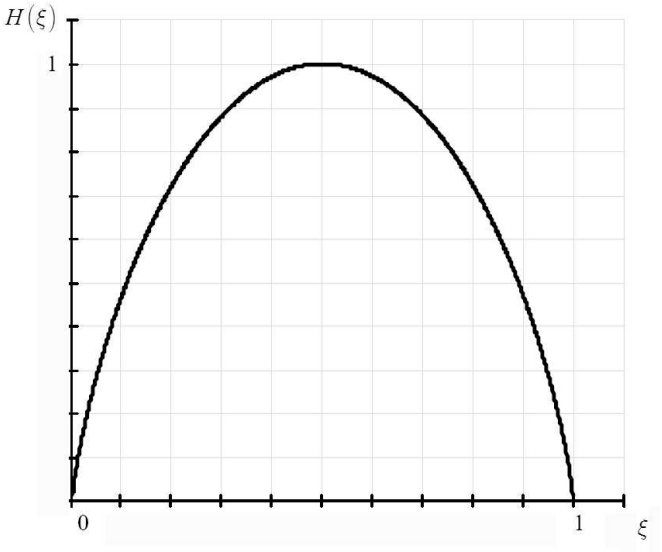


Рис. 77. График функции $H(\xi)$

Исходя из того, что на шаге n размер каждого отрезка равен $1/2^n$, а всего таких отрезков - C_n^k, получаем:

$$D = \lim_{n\to\infty} \frac{\log_2 C_n^k}{\log_2\left(1/2^n\right)} = H\left(\frac{k}{n}\right).$$

Таким образом, как следует из свойств функции $H(\xi)$ (см. рис. 76), существует целый спектр фрактальных размерностей.

В рамках рассматриваемой модели распределения наследства величина k/n означает всех наследников, имеющих одно и тоже количество «золота» (определенную часть начального наследства). Как показано выше, количество таких наследников при росте n увеличивается по фрактальному закону. Очевидно, что при крайних значениях $k = 0$ и $k = n$ фрактальная размерность равна нулю. Действительно, множество самых богатых наследников, имеющих

p^n «золота» состоит из одного человека, то же самое справедливо и для самого бедного, имеющего $(1-p)^n$ первоначального наследства.

Остановимся детальнее на формальном определении мультифрактальности. Носителем мультифрактальной меры является множество L – объединение фрактальных подмножеств L_α. Т.е. мультифрактал можно понимать как некое объединение различных однородных фрактальных подмножеств L_α исходного множества L, каждое из которых имеет свое собственное значение фрактальной размерности.

Для характеристики мультифрактального множества используют так называемую функцию мультифрактального спектра $f(\alpha)$ (спектр сингулярностей мультифрактала), к которой вполне подходил бы термин «фрактальная размерность». Величина $f(\alpha)$ равна хаусдорфовой размерности однородного фрактального подмножества L_α из исходного множества L, которое дает доминирующий вклад в некоторую статистическую сумму (как будет показано ниже, в моменты распределения при заданных значениях порядка моментов q).

Кроме того, для описания мультифрактала используют обобщенные фрактальные размерности D_q. В соответствии с мультифрактальным формализмом, обобщенные фрактальные размерности D_q определяются соотношением:

$$D_q = \lim_{r\to 1} \frac{1}{q-1} \frac{\ln \sum_{i=1}^{N} p_i^q}{\ln r},$$

где p_i – вероятность того, что случайная величина (нормированный по общей сумме элемент числового ряда) попадет в некоторый диапазон r.

Далее вводится показатель мультифрактального скейлинга τ, который определяется на основании значений D_q и q:

$$\tau(q) = (1-q)D_q.$$

Функции $f(\alpha)$ и $\tau(q)$ связаны друг с другом соотношением:

$$\tau(q) = f(\alpha) - q\alpha,$$

где α как функция от q определяется из решения уравнения:

$$\frac{d}{d\alpha}\left(q\alpha - f(\alpha)\right) = 0.$$

И наоборот, если известна фрактальная размерность D_q (или показатель мультифрактального скейлинга $\tau(q)$), то мультифрактальный спектр может быть найден по формуле:

$$f(\alpha(q)) = \tau(q) + q\alpha(q),$$

где

$$\alpha(q) = -\frac{d\tau(q)}{dq}.$$

Эти соотношения задают кривую $f(\alpha)$ параметрически (как функцию от параметра q) и представляют собой так называемое преобразование Лежандра от переменных q и τ к переменным α и f.

При анализе ряда динамики событий использовался следующий метод расчета мультифрактальных характеристик. Значения исследуемого ряда нормируются $Z_i = \frac{\xi_i}{\sum_k \xi_k}$ и ассоциируются с вероятностями p_i в рамках приведенной выше формуле для расчета обобщеных фрактальных размерностей D_q.

После нормирования весь диапазон значений ряда $\left[0,\ N\right]$ разбивался на $n = N/m$ ячеек (участков) длиной m. Затем определялась следующая сумма:

$$S_m^Z(q) = \sum_{k=1}^{n}\left(\overline{Z}_k^{(m)}\right)^q,$$

где

$$\overline{Z}_k^{(m)} = \sum_{l=1}^{m} Z_{(k-1)m+l}.$$

Как оказалось для рядов, задаваемых динамикой публикаций, $\log S_m^Z(q)$ хорошо аппроксимируется линейной зависимостью от $\log m$, в результате чего появилась возможность говорить [128], что числовой ряд - мультифрактал. Наклон аппроксимирующей линии, полученный методом наименьших квадратов - $\tau(q)$ определялся по формуле:

$$\log S_m^Z(q) \cong \tau^Z(q) \log m + \text{const}.$$

Приведенные ниже расчеты относятся к анализу числового ряда отражающих проблематику использования антивирусного программного обеспечения (посуточная динамика публикаций в интернет-новостях сообщений по данной теме, получаемая с помощью системы контент-мониторинга в течение всего 2007 г. и первого квартала 2008 г.), а также ряда, полученного по уточненной тематике (первоначальный запрос был расширен словом «тороянский»). Соответствующие посуточные диаграммы приведены на рис. 78.

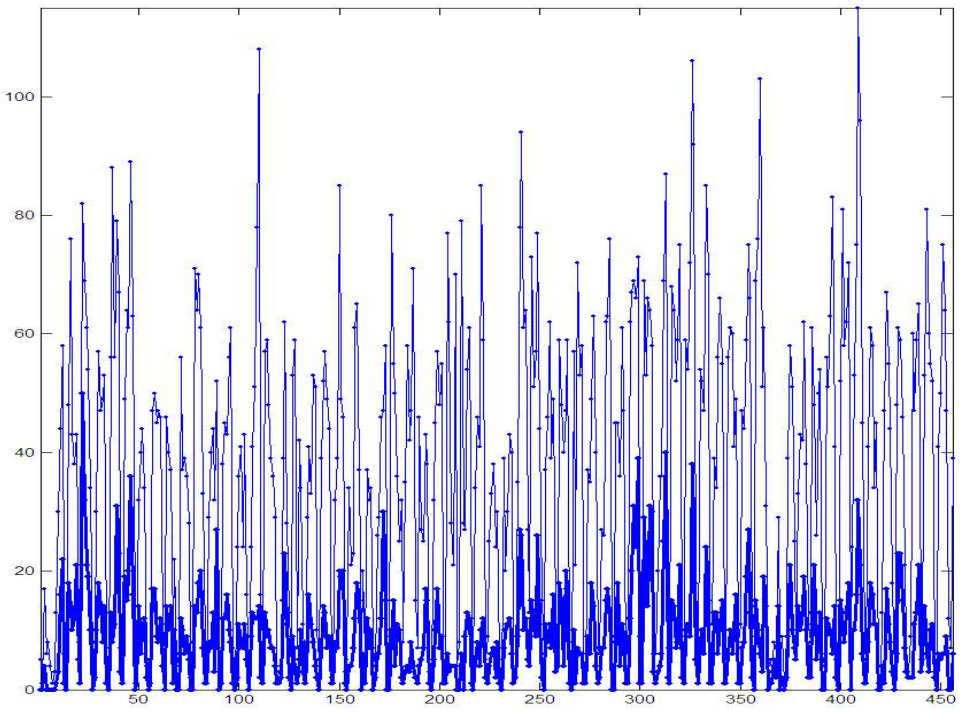


Рис. 78. Диаграммы интенсивности публикаций по основной (тонкая соединительная линия) и уточненной тематике (жирная линия): ось абсцисс – порядковые номера дней, ось ординат – количество публикаций

На рис. 79 показана поверхность – зависимость $\tau(m, q)$ от q и m для динамики появления документов. В соответствии с формулой:

$$f(\alpha(q)) = \tau(q) - q\tau'(q),$$

был определен мультифрактальный спектр исследуемого ряда (рис. 79).

Во многих мультифрактальных исследованиях основным объектом анализа является зависимость мультифрактального спектра f от индекса сингулярности (показателя Липшица-Гельдера) α. Данная зависимость для рядов, соответствующих основной и уточненной тематике представлена на рис. 80.

Итак, ряды, соответствующие динамике появления публикаций, в рассмотренных случаях обладают мультифрактальной природой. Вместе с тем соответствующие исследуемым рядам зависимости (рис. 81), имеют различные параметры кривизны. Этот факт свидетельствует, с одной стороны, о том, что ряд, соответствующий подтематике менее стабилен, чем ряд, соответствующий всей тематике, а с другой стороны, о том, что рассматриваемая подтематика не является репрезентативной для анализа потока публикаций по общей тематике.

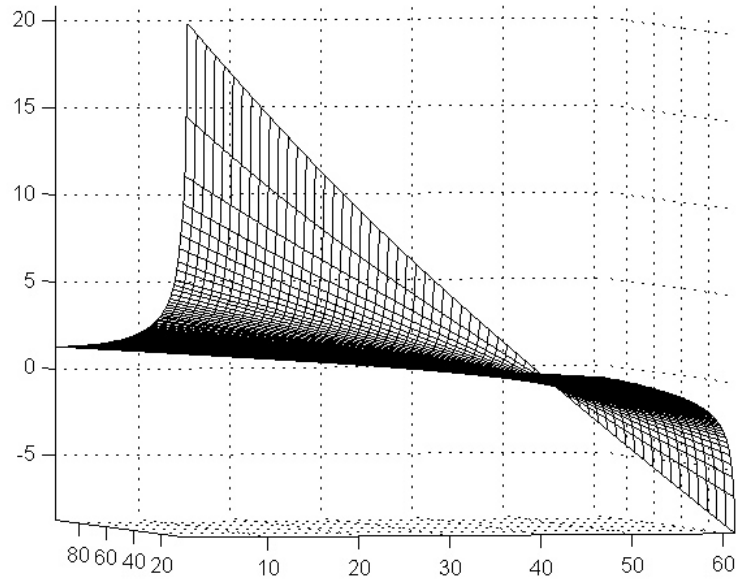


Рис. 79. Значения τ(q,m) для исследуемого ряда (запрос «банк»)

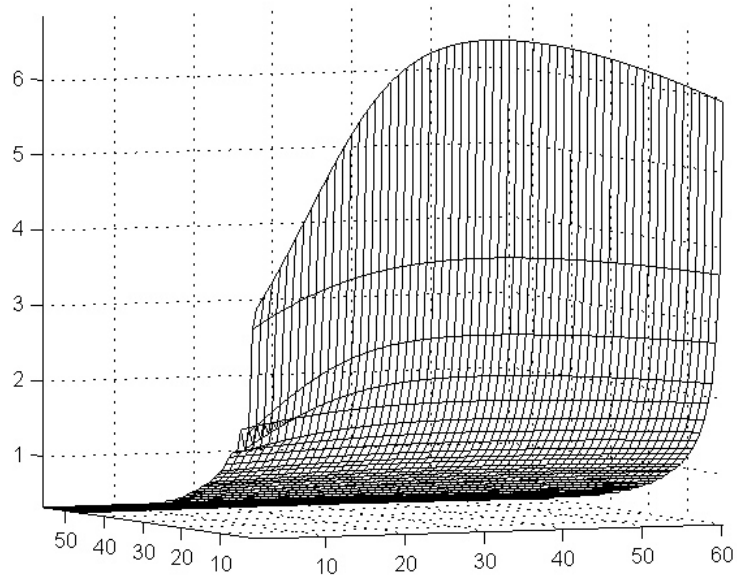


Рис. 80. Значения мультифрактального спектра $f(q,m)$ для исследуемого ряда

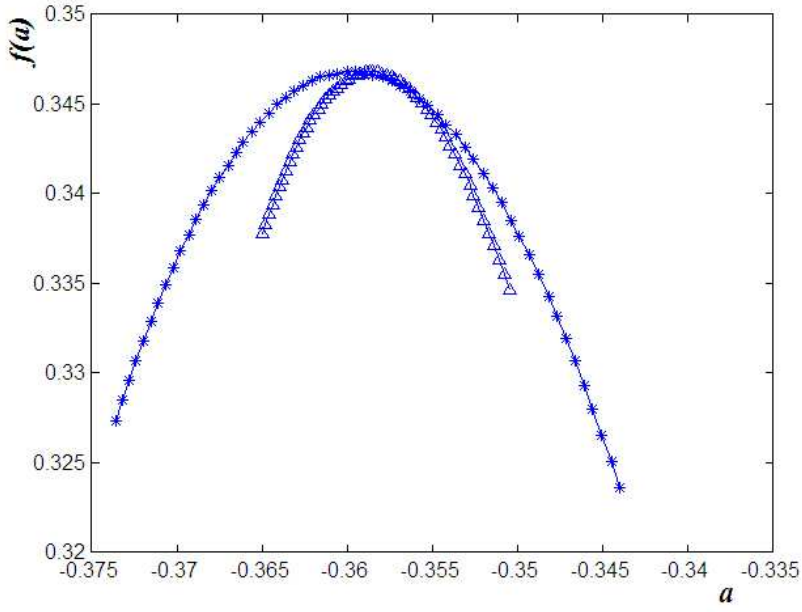


Рис. 81. Сравнение мультфрактальных спектров исследуемых рядов – по основной тематике (▲) и уточненного (*) от индекса сингулярности

В свою очередь, для формирования репрезентативных выборок из массивов документов, может быть применен подход, основанный на подобии мультифрактальных спектров, дополняющий традиционные методы, базирующиеся на выявлении содержательного подобия документов. Практическая ценность задачи выявления репрезентативных выборок, основанная на данном подходе, может быть выражена в таких приложениях, как предъявление

пользователю обозримых результатов поиска, отражающих весь спектр документального массива (с учетом колебаний интенсивности публикаций по дням) или выделение подмножеств документов для дальнейших детальных исследований.

ЗАКЛЮЧЕНИЕ

Сегодня развитие сетевых информационных ресурсов соответствует закономерности Мура, которая в начале была сформулирована как прогноз развития технологии микросхем, но сегодня все шире вторгается во все сферы жизни. В 1965 году Гордон Мур дал прогноз, что плотность транзисторов в интегральных схемах и, соответственно, производительность микропроцессоров будут удваиваться каждый год. На протяжении трех последних десятилетий этот прогноз, названный «законом Мура», более или менее выполнялся, хотя довольно быстро был скорректирован - удвоение должно происходить каждые два года [26].

Гордон Мур

В настоящее время прогноз Мура распространяется на все большее количество областей. Сегодняшнее расширение Интернет, стремительный рост объемов данных в Сети, развитие электронной коммерции и беспроволочной связи, а также внедрение цифровых технологий в бытовую технику, можно рассматривать как следствия этого закона. Было замечено, что рост документальной информации, целиком подчиняясь закону Мура, также носит экспоненциальный характер (рис. 82), а именно кривая роста числа документов может быть описана уравнениям вида $y = Ae^{kt}$, где y – количество документов,

t – время, A – количество документов в момент начала отсчета, k – коэффициент.

Развитие коммуникационных возможностей приводит к росту количества доступной информации, в частности, в Интернет. С другой стороны, увеличение объемов доступного контента способствует росту инновационной деятельности. Все больше знаний, необходимых для исследовательских работ, публикуется в Сети, тем самым способствуя технологическому прогрессу, на котором основывается прогноз Мура.

Предполагается, что новый уровень развития сетевого информационного пространства будет определяться технологиями работы с огромным объемом информации, накопившимся в Интернет.

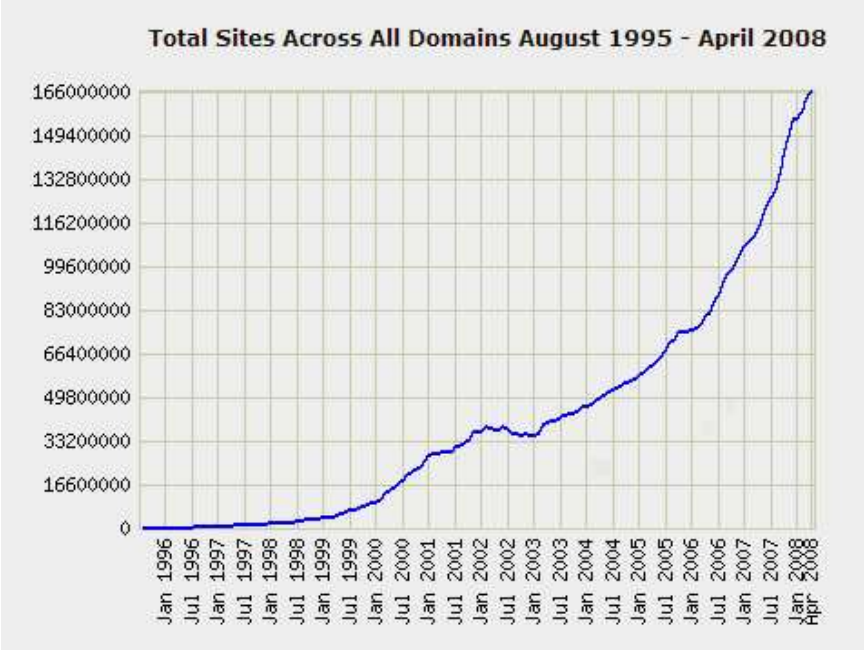


Рис. 82. Динамика роста количества веб-серверов Сети (по данным службы Netcraft за апрель 2008 года)

Информация, размещенная в компьютерных сетях, образует крупнейший распределенный информационный ресурс благодаря нескольким изначльно заложенным принципам, к которым, в частности, относится реализация гипертекста, позволяющего интегрировать неоднородные информационные ресурсы, естественную, адаптированную к человеческой логике систему навигации.

Вместе с тем возможности доступа к информации в WWW всегда ограничивались статичностью языка HTML, что обуславливало преимущественно навигационный доступ к документам, практическое отсутствие поддержки метаинформации, несовершенство идентификации информационных ресурсов, и, самое главное, тот факт, что разметка HTML относилась только к внешнему представлению документов, не касаясь их семантики.

По мере развития WWW его возможности расширялись, эволюционно были добавлены динамические компоненты, возможность управлять стилевыми решениями, были разработаны и некоторые принципы представления контента, зафиксированные как стандарты.

Наряду с этим традиционному WWW все же присущи такие недостатки, как высокий уровень информационного шума, невозможность гарантирования целостности документов, отсутствие возможности смыслового поиска, ограниченность доступа к «скрытому» веб. Над решением названных проблем работают многочисленные коллективы во всем мире, в частности, консорциум W3C, где под руководством основателя WWW Тима Бернерса-Ли разрабатывается концепция Семантического веб [74]. Основная идея этого проекта заключается в организации такого представления данных в сети, чтобы допускалась не только их визуализация, но и эффективная автоматическая обработка программами разных производителей. Путем таких радикальных преобразований предполагается превращение WWW в систему семантического уровня. По замыслу создателей Семантический веб должен обеспечить "понимание" информации компьютерами, выделение ими наиболее подходящих по тем или иным критериям данных, и уже после этого - предоставление информации пользователям [114].

Семантический веб будет представлять собой расширение существующей сети WWW, в котором информация будет представляться в соответствии со смысловым значением, что повысит уровень согласованности при взаимодействии людей и компьютерной сети. Это будет достигаться за счет объединения разнообразных видов информации в единую структуру, где каждому смысловому элементу данных будет соответствовать специальный

синтаксический блок (тег). Теги будут образовывать единую иерархическую структуру.

Тим Бернерс-Ли

В рамках Семантического веб разрабатываются языки для выражения информации в форме, доступной для машинной обработки, на которых можно будет описывать как данные, так и принципы трактования этих данных. Предполагается, что правила выводов, существующие в какой-либо одной системе представления знаний, будут доступны другим системам.

Сегодня очевидно, что центральную роль в представлении и обмене данными в Семантическом веб будет играть Расширяемый Язык Разметки (XML). Предполагается также использование нового принципа идентификации информационных ресурсов, формирование новой архитектуры веб-пространства на основе многоуровневого представления информационных ресурсов и стандартизированных веб-сервисов.

В процессе развития концепции Семантического веб получили широкое развитие синтаксические методы представления информации языковыми средствами XML и его дополнений, предназначенных для описания типичных свойств элементов XML-документов, их структуры и семантики: рекомендации W3C, регламентирующие DTD, XML Schema, XQuery (язык запросов к базам XML-данных) и т.д. К языкам представления данных относятся также средства описания ресурсов RDF.

Отдельная область Семантического веб названа онтологическим подходом. Этот подход включает средства аннотирования документов, которыми могли бы воспользоваться компьютерные программы - веб-сервисы и агенты при обработке сложных запросов пользователей. Модели предметных областей в терминологии Семантического веб называются онтологиями. В 2004 г. консорциумом W3C была утверждена и опубликована спецификация языка сетевых онтологий OWL (Web Ontology Language). Язык онтологий OWL выступает решающим компонентом интеллектуализации, базой для построения семантических сетей. При этом сами онтологии образуют систему, состоящую из наборов понятий и утверждений об этих понятиях, на основе которых можно строить классы, объекты и отношения. Отдельная онтология определяет семантику конкретной предметной области и способствует установлению связей между значениями ее элементов.

Итак, в Семантическом вебе используются три ключевых языка:

- XML, позволяющий определить синтаксис и структуру документов;
- механизм описания ресурсов RDF, обеспечивающий модель кодирования для значений понятий;
- язык онтологий OWL, позволяющий определять понятия и отношения между ними.

Если говорить о логических уровнях, на которых базируется технология Семантического веб (рис. 81), то самый нижний уровень - это Universal Resource Identifier (URI), унифицированный идентификатор, определяющий способ записи адреса произвольного ресурса.

Семантический веб, именуя всякое понятие просто с помощью URI-идентификатора, дает возможность каждому выражать те понятия, которыми он пользуется. Типичными примерами URI-идентификаторов являются URL-адреса, однако URI-идентификатор задавая или ссылаясь на некоторый ресурс, не обязательно при этом указывает на его местонахождение в Интернет.

В рамках Семантического веб особая роль отводится электронным агентам – программам, которые для достижения поставленных перед ними целей работают без непосредственного управления со стороны человека. Предполагается, что

эффективность программных агентов в Семантическом веб будет расти по мере увеличения количества доступного им веб-контента и автоматизированных сервисов.

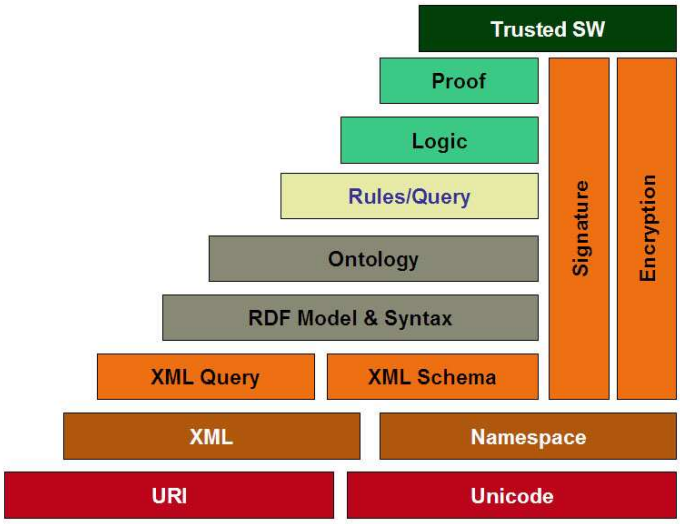


Рис. 81. Архитектура Семантического веб

Перспективы эффективного охвата информационного пространства будут зависеть как от создания и развития эффективной сетевой инфраструктуры, так и развития теоретических основ информатики. В этой связи одной из актуальнейших задач, стоящих перед исследователями различных специальностей, является построение адекватных моделей сетевого информационного пространства и информационного поиска, которые базируются на достижениях в областях лингвистики и информатики, строгом математическом инструментарии.

Успешное продвижение в изучении современного информационного пространства невозможно без общих представлений о структуре и свойствах динамики сетевых информационных процессов, что в свою очередь требует выявления и учета их устойчивых закономерностей в рамках нового научного направления – «Интернетики».

СПИСОК СОКРАЩЕНИЙ

ДНФ	Дизъюнктивная нормальная форма
ИА	Информационное агентство
ИПС	Информационно-поисковая система
ПОД	Поисковый образ документа
РОМИП	Российский семинар по Оценке Методов Информационного Поиска
СМИ	Средства массовой информации
СУБД	Система управления базами данных
ARPANET	Advanced Research Projects Agency Network, Сеть Управления Перспективных Исследований
BFS	Breadth First Search, метод широкого первичного поиска
DBMS	Database Management System, система управления базами данных
DDos	Distributed Denial of Service, распределённый отказ обслуживания
DFA	Detrended Fluctuation Analysis, анализ колебаний с исключенным трендом
DNS	Domain Name System, система доменных имен
DTD	Document Type Definition, язык описания структуры документа
HAC	Hierarchical Agglomerative Clustering, иерархическое группирование-объединение
HITS	Hyperlink Induced Topic Search, метод ранжирования
HTML	HyperText Markup Language, язык гипертекстовой разметки
HTTP	HyperText Transport Protocol, протокол передачи гипертекста
IETF	Internet Engineering Task Force, Группа по решению задач проектирования Интернет
IRS	Information Retrieval System, информационно-поисковая система

ISM	Intelligent Search Mechanism, интеллектуальный поисковый механизм
LSA	Latent Semantic Analysis, латентно-семантический анализ
OSI	Open Systems Interconnection Reference Model, открытая модель сетевых коммуникаций
OWL	Web Ontology Language, язык веб-онтологий
P2P	Peer-to-peer, пиринговые сети
PLSA	Probabilitstic Latent Semantic Analysis, вероятностный латентно-семантический анализ
RBFS	Random Breadth First Search, метод случайного широкого первичного поиска
RDF	Resource Description Framework, модель для описания ресурсов
RFC	Request for Comments, запрос для комментариев
RWA	Random Walkers Algjrithm, метод «случайных блужданий»
Salsa	Stochastic Approach for Link-Structure Analysis, стохастический алгоритм анализа структуры связей
SNA	Social Network Analysis, анализ социальных сетей
SCC	Strongly Connected Component, компонента сильной связности
SQL	Structured Query Language, язык структурированных запросов.
SVD	Singular Vector decomposition, сингулярное разложение матриц
SVM	Support Vector Mashine, метод опорных векторов
TCP/IP	Transmission Control Protocol/Internet Protocol, основные протоколы Интернет
TREC	Text Retrieval Conference, Конференция по оценке систем текстового поиска
URI	Universal Resource Identifier, универсальный идентификатор ресурсов
URL	Universal Resource Locator, универсальный локатор ресурсов
W3C	World Wide Web Consortium, Консорциум W3C

WWW	World-Wide Web, Всемираня паутина
WAIS	Wide Area Information Service, служба поиска распределенной информации
XML	Extensible Markup Language, расширяемый язык разметки

ГЛОССАРИЙ

Автоматическое реферирование (от англ. Summarization) - автоматическое формирование краткого изложения исходного текстового материала либо путем выделения фрагментов информационного наполнения и последующего их соединения, либо методом генерации текста на основании выявления знаний из оригинала.

База данных реляционная - база данных, построенная на основе реляционной модели данных.

Весовой коэффициент (англ. - Weighting) - коэффициент, приписываемый лексической единице в документе и учитываемый для вычисления числового значения релевантности. Весовой коэффициент может зависеть от расположения лексической единицы в документе, абзаце, предложении. Кроме того, весовой коэффициент непосредственно зависит от смысла лексической единицы, ее соответствия тематике поисковой системы, частоты встречаемости в документе. Весовые коэффициенты могут приписываться лексическим единицам как в индексе информационно-поисковой системы, так и в запросах пользователя.

Гипертекст (от англ. Hypertext) - документы, содержащие связи с другими документами (или имеющие внутренние связи). Гипертекстовый документ представляет собой специальным образом размеченную текстовую информацию. При отображении гипертекстовых документов отдельные элементы текста могут служить ссылками на другие документы. Механизм ссылок, дополняющий текстовую информацию, является неотъемлемой частью гипертекста. Веб-страницы, как правило, представляют собой гипертекстовые документы написанные с использованием языка гипертекстовой разметки HTML.

Гиперсвязь, гиперссылка (англ. - Hyperlink) - связь между отдельными компонентами информации. Применяется для ссылок, сделанных внутри одного объекта на другой объект. Ссылка, как правило, делается от объекта, размещенного на HTML-странице, на другой объект, который может находиться на произвольном FTP или WWW-сервере.

ДНФ (дизъюнктивная нормальная форма) - нормальная форма в булевой логике, в которой булева формула имеет вид дизъюнкции нескольких конъюнктивных компонент (пропозициональных формул, являющихся конъюнкцией одного или более элементов). Известно, что любая булева формула может быть приведена к ДНФ.

Индекс ИПС (англ. – IRS Index) - индекс информационно-поисковой системы представляет собой определенным образом организованную совокупность данных, где содержатся поисковые образы всех документов базы данных. Является основной составляющей архитектуры информационно-поисковой системы, обеспечивающей возможность оперативного поиска и доступа к информации.

Интернет, Сеть (Internet) - глобальная информационная сеть, части которой логически связаны единым адресным пространством, основанном на стеке протоколов TCP/IP. Интернет состоит из множества взаимосвязанных компьютерных сетей.

Информационное пространство (англ. - Information space) - совокупность информационных ресурсов, технологий их сопровождения и использования, информационных и телекоммуникационных систем, образующих информационную инфраструктуру.

ИПС, Информационно-поисковая система (англ. - Information Retrieval System, IRS) - система, предназначенная для обеспечения поиска и отображения документов, представленных в базах данных. Ядром информационно-поисковой системы является поисковый механизм - программный модуль, который осуществляет поиск по запросу. ИПС, интегрированные с веб-технологиями, являются основой построения информационно-поисковых веб-серверов.

Ключевое слово (англ. - Keyword):

1. Отдельный термин, используемый в запросах к информационно-поисковым системам.

2. Дескриптор, отдельное слово или словосочетание, используемое при ручном или автоматизированном индексировании документов перед погружением в ИПС.

Контент (от англ. Content) - содержание. Под "контентом" обычно понимают любое содержательное наполнение информационных ресурсов (например, веб-сайтов) - тексты, графику, мультимедиа.

Контент-анализ - метод получения выводов путем анализа содержания текстовой информации. Чаще всего реализуется как систематическая числовая обработка, оценка и интерпретация формы и содержания информационного источника.

Контент-мониторинг - систематическое, непрерывное во времени сканирование и контент-анализ информационных ресурсов.

Кэш (от англ. cache) - в подборка данных, дублирующих оригинальные значения, когда оригинальные данные труднодоступны из-за большого времени доступа или для вычисления. Кэш - это промежуточный буфер с быстрым доступом, который хранит в себе информацию, которая может быть запрошена пользователем.

Латентно-семантический анализ (от англ. Latent Semantic Analysis, LSA) - теория и метод для извлечения контекстно-зависимых значений слов при помощи статистической обработки больших наборов текстовых данных. Латентно-семантический анализ основывается на идее, что совокупность всех контекстов, в которых встречается и не встречается терм, задает множество ограничений, которые в значительной степени позволяют определить похожесть смысловых значений термов между собой. В качестве исходной информации LSA использует матрицу «термы-на-документы», содержащую весовые значения термов в документах.

Лемматизация (от англ. Lemmatization) - реконструкция основной формы изменяемых частей речи, приведение слов к исходной (канонической) форме - лемме. Если существительное - то к именительному падежу, если глагол - то к инфинитивной форме и т.д.

Метаданные – «данные о данных» - описание состава данных, их структуры представления, места хранения и других признаков.

Метаинформация - информация о способах и методах переработки информации или о том, где найти информацию. Так, интернет-каталог

представляет собой метаинформацию по отношению к информации, содержащейся на веб-сайтах.

Модель реляционная (от англ. Relation – отношение) - логическая модель данных, описывающая:

- структуры данных в виде наборов отношений;
- теоретико-множественные операции над данными: объединение, пересечение, разность и декартово произведение;
- специальные реляционные операции: селекция, проекция, соединение и деление;
- специальные правила, обеспечивающие целостность данных.

Мультифрактал – множество содержащие в себе одновременно бесконечное число фрактальных множеств, характеризуется спектром фрактальных размерностей.

Онтология - в рамках концепции Семантического веб - онтология определяет термины, с помощью которых можно описать предметную область. Попытка формализации некоторой области знаний с помощью концептуальной схемы. Обычно такая схема состоит из иерархической структуры данных, содержащей все релевантные классы объектов, их связи и правила, принятые в этой области.

Параметр порядка – величина отличительный признак фазы, в которой находится система, например неравная нулю а ферромагнитной фазе и равная в парамагнитной.

Поисковый механизм (англ. - Search Engine) - основной компонент любой информационно-поисковой системы. Программный модуль, осуществляющий поиск в базе данных по запросу (поисковому предписанию), заданному пользователем.

Полнота, охват (англ. - Recall) - отношение количества релевантных документов в отклике информационно-поисковой системы к общему количеству релевантных документов в исходном массиве.

Полнотекстовая поисковая система (англ. - Full-text search engine) - информационно-поисковая система, которая при составлении индекса охватывает

все слова в тексте документа (иногда за исключением стоп-слов) и учитывает порядок их расположения по отношению друг к другу.

Профайл (от англ. profile – профиль) - совокупность величин определяющих (базовых) параметров некоторого объекта или технологического процесса, описывающих и характеризующих этот объект или технологический процесс.

Ранжирование (от англ. Ranking) - упорядочение результатов поиска – отклика поисковой системы по некоторым критериям, например, по дате публикации документов или по релевантности.

Релевантность (от англ. Relevancy – соответствие) - мера того, насколько точно документ, найденный информационно-поисковой системой, отвечает запросу пользователя. Обычно выражается в числовой форме. Единых взглядов на это понятие нет. Далеко не всегда документ, отмеченный информационно-поисковой системой как наиболее релевантный по формальным признакам, будет таковым по мнению самого пользователя.

Реляционная модель данных - логическая модель данных, описывающая структурный аспект, аспект целостности и аспект обработки данных:

- структурный аспект - данные в базе данных представляют собой набор отношений;
- аспект целостности - отношения отвечают определенным условиям целостности. Реляционная модель поддерживает ограничения целостности уровня типов данных, уровня отношения и уровня базы данных;
- аспект обработки - реляционная модель поддерживает операторы манипулирования отношениями - так называемую реляционную алгебру.

Семантический веб (от англ. Semantic Web) - проект консорциума W3C, в рамках которого предлагается способ сделать информацию в Сети более доступной, что, в свою очередь, позволит создавать интеллектуальное программное обеспечение, которое могло бы искать в WWW необходимые данные, выявляло их семантику, создавало перекрестные ссылки и использовало

эти данные для решения практических задач. Одина из основных концепций Семантического веб – ориентация на формат XML.

Скейлинг – масштабирование**,** в частности, возможность представить функцию двух переменных как функцию одной.

Сниппет (от англ. Snippet – фрагмент, отрывок) - часть текста, отрывки веб-страницы, которая содержит слова поискового запроса, выводящегося поисковой системой в результатах поиска по самому этому запросу.

Спам (SPAM) - непрошенное рекламное сообщение, сетевой мусор, мусорная почта, рассылаемые по электронной почте в личные почтовые ящики или телеконференции. Рассылка спама считается нарушением этикета и правил применения компьютерных сетей.

Стемминг (от англ. Stemming) – выделение основы слова - обеспечивает возможность поиска слова не только в строго заданном виде, но и во всех его морфологических формах. Например, слову "программа", будут соответствовать: "программе", "программный" и т.д.

Стоп-слова (англ. - Stop words) - слова, исключаемые из индекса системы и/или запроса пользователя. Отдельные информационно-поисковые системы для сокращения размеров индекса и увеличения производительности не включают в индекс часто встречаемые на веб-страницах слова. К стоп-словам обычно относятся предлоги, междометия и другие сочетания, которые не несут содержательного смысла.

СУБД (англ. - DBMS) - система управления базами данных - комплекс программных и лингвистических средств общего или специального назначения, реализующий поддержку создания баз данных, централизованного управления и организации доступа к ним различных пользователей в условиях принятой технологии обработки данных.

СУБД реляционная - система управления реляционной базой данных, содержащая:

- командный язык;
- язык программирования с ориентацией на обработку таблиц;
- интерпретирующую и/или компилирующую систему;

- пользовательскую оболочку.

Тег (от англ. tag):

1. Специальная последовательность знаков в размеченном тексте, указывающая на структуру или формат его представления.
2. Команда и знак языка разметки гипертекста. Знаки разметки употребляются парами, обозначая начало и конец области действия тега.

Текстовый корпус (от англ. Text corpus) - массив текстов, собранных в соответствии с определенными принципами, размеченных по определенному стандарту и обеспеченных специализированной поисковой системой. В некоторых случаях текстовым корпусом первого порядка называют произвольное собрание текстов, объединенных каким-то общим признаком. Разработкой, созданием и использованием текстовых (лингвистических) корпусов занимается специальный раздел языкознания – корпусная лингвистика.

Терм (от англ. Term) – в контексте данной книги – слово или устойчивое словосочетание. В математической логике понятие «терм» широко используется в качестве «символьного выражения».

Фазовые переходы – переход системы из одной фазы в другую при изменении внешних условий, например, в физике, переход при повышении температуры железа, кобальта, никеля,… из ферромагнитной фазы в парамагнитную. Фазовый переход происходит при определенной, т.н. критической температуре. Согласно П. Эренфесту различаю фазовые переходы 1-го, 2-го и т.д. родов.

Фрактал (от лат. Fractus – дробленый, состоящий из фрагментов) – бесконечно самоподобный (точно или приближенно) объект (множество), каждая часть которого повторяется при уменьшении масштаба. Более точно – размерность Хаусдорфа-Безиковича такого объекта должна быть нецелой, поэтому фрактал самоподобен, обратное не обязательно. Возможно и такое определение: фрактал - самоподобное множество нецелой размерности.

Энтропия – в физике - мера вероятности осуществления какого-либо макроскопического состояния; в теории информации - мера неопределенности какого-либо опыта.

ARPANET (Advanced Research Projects Agency Network, Сеть Управления Перспективных Исследований) - глобальная исследовательская сеть с коммутацией пакетов, предшественница Интернет. Основана в 1969 году под эгидой Агентства перспективных исследований Министерства обороны США (Defense Department's Advanced Projects Research Agency). В сети ARPANET впервые были реализованы многие из сетевых принципов, которые используются сегодня. Завершила свое существование в 1990 году.

Data Mining (глубинный анализ данных):

1. Data mining - это процесс обнаружения в сырых данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности (G. Piatetsky-Shapiro, GTE Labs)

2. Data mining - это процесс выделения (selecting), исследования и моделирования больших объемов данных для обнаружения неизвестных до этого структур (patterns) с целью достижения преимуществ в бизнесе (SAS Institute).

Deep Web (глубинный, скрытый, невидимый веб) - кроме видимой для поисковых систем части WWW-пространства существует огромное количество страниц, которые ими не охватываются. Как правило, эти веб-страницы доступны в Интернет, однако выйти на них невозможно, если не знать точного адреса. В состав ресурсов Deep Web входят и динамически формируемые веб-страницы, содержание которых хранится в базах данных и доступно лишь по запросам пользователей.

DNS (Domain Name System) - система доменных имен — распределенная система (распределенная база данных), позволяющая преобразовывать символьные имена доменов в IP-адреса в сетях TCP/IP. Кроме того DNS может хранить и обрабатывать обратные запросы определения имени хоста по его IP адресу.

HTML (HyperText Markup Language) - язык гипертекстовой разметки - стандартный язык для описания содержания и структуры гипертекстовых документов. HTML-документы представляют собой текстовые файлы со встроенными специальными командами (разметкой), которые, как правило,

отмечают определенную область текста. HTML состоит из независящих от программного обеспечения и аппаратной платформы команд, описывающих структуру гипертекстовых документов. HTML является прикладной разновидностью языка SGML.

HTTP (HyperText Transport Protocol) - протокол передачи гипертекста – протокол, предназначенный для общения клиента и сервера в WWW. Обеспечивает передачу веб-страниц по Интернет.

MARC - Проект МАРК - проект, начатый в 1966 году 16 библиотеками США для разработки стандарта формата обмена библиографическими записями в электронном виде. В 1972 году модернизированный стандарт МАРК-2 получил международное признание.

OSI (Open Systems Interconnection Reference Model) - абстрактная модель для сетевых коммуникаций и разработки сетевых протоколов. Представляет семиуровневый подход к построению архитектуры сети. Каждый уровень обслуживает свою часть процесса взаимодействия и может взаимодействовать только со своими соседями и выполнять отведенные только ему функции.

OWL (Web Ontology Language) - язык веб-онтологий для Семантического веб на основе стандартов XML/RDF. Язык веб-онтологий OWL предназначен для описания классов веб-документов и приложений, а также отношений между этими классами. В основу языка положена модель данных «объект – свойство».

P2P (Peer-to-peer) – пиринговые сети, основанные на равноправии участников. В таких сетях отсутствуют выделенные серверы, а каждый узел (peer) является как клиентом, так и сервером. В отличие от сетей с архитектурой «клиент-сервер», такая организация сети позволяет сохранять работоспособность сети при произвольном количестве и сочетании узлов.

RDF (Resource Description Framework) - разработанная консорциумом W3C модель для описания ресурсов. В основе этой модели лежит идея об использовании специального вида утверждений, соответствующих ресурсам. Каждое утверждение имеет вид «субъект - предикат - объект», называемое триплетом.

RFC (Request for Comments) - запрос для комментариев - совокупность публикуемых документов, в которых излагаются стандарты, проекты стандартов и принципиально согласованные идеи по деятельности Интернет. Первый RFC вышел в 1969 году. Общее количество RFC на сегодня превышает пять тысяч.

SQL (Structured Query Language) - язык структурированных запросов - язык системы управления базой данных, использующий соответствующие команды и синтаксис для управления процессом взаимодействия и обработки данных в базе данных.

TCP/IP (Transmission Control Protocol/Internet Protocol) – два основных протокола, обеспечивающих (вместе с другими протоколами) функционирование и работу в сетях Интернет в режиме коммутации пакетов. Используются как правило совместно:

- TCP (Transmission Control Protocol) - протокол, определяющий порядок разделения данных на дискретные пакеты и контролирующий передачу и целостность передаваемых данных;
- IP (Internet Protocol) - описывает формат пакета данных, передаваемых в сети, а также порядок присвоения и поддержки адресов абонентов сети.

Text Mining - глубинный анализ текста - это алгоритмическое выявление прежде не известных связей и корреляций в уже имеющихся текстовых данных. Важная компонента технолгоии Text Mining связана с извлечением из текста его характерных элементов или свойств, которые могут использоваться в качестве метаданных, ключевых слов, аннотаций. Другая важная задача состоит в отнесении документа к некоторым категориям из заданной схемы их систематизации. Text Mining также обеспечивает новый уровень семантического поиска документов.

W3C - World Wide Web Consortium - Консорциум W3C - международный индустриальный консорциум, образованный в 1994 г. первоначально в рамках CERN при поддержке DARPA и Европейской комиссии. В настоящее время W3C поддерживается совместно Лабораторией информатики Массачусетского технологического института (США), INRIA (Франция) и университетом Кейо (Япония). Целью создания W3C является разработка общих протоколов,

позволяющих расширить доступность и эффективность ресурсов World Wide Web, а также руководство эволюцией системы протоколов.

WAIS (Wide Area Information Service) - служба поиска распределенной информации:

1. WAIS-протокол Интернет, позволяющий осуществлять поиск информации в Интернет в соответствии с библиографическим стандартом Z39.50.

2. Информационно-поисковая система, построенная в соответствии с WAIS-протоколом.

XML (Extensible Markup Language) – Расширяемый Язык Разметки - стандарт языка разметки, принятый консорциумом W3C в феврале 1998 г. Главные его особенности заключаются в возможности расширения набора тегов, используемых для разметки документов, возможности задания структуры документа, правильность которой верифицируется браузером, в отделении средств разметки по содержанию от разметки, ориентированной на представление документов.

ЛИТЕРАТУРА

[1] Аграновский А.В., Арутюнян Р.Е. Индексация массивов документов // Мир ПК, - № 06. -2003. URL: http://www.osp.ru/pcworld/2003/06/165855/

[2] Алексеев Н.Г. Применение закона Бредфорда при комплектовании фонда научной библиотеки // Тезисы докладов конференции "Библиотечное дело-1996". URL: http://libconfs.narod.ru/1996/4s/4s_p1.html

[3] Антонов А.В. Методы классификации и технология Галактика-Зум // Научно-техническая информация. - Сер. 1. - Вып. 6. - 2004. - С. 20-27.

[4] Аракелян С.М., Духанов А.В., Прокошев В.Г., Рощин С.В. Самоорганизующаяся информационная среда с децентрализованным управлением для взаимодействия образовательных учреждений // В сб. науч. ст. "Интернет-порталы: содержание и технологии". - Вып. 4 / ФГУ ГНИИ ИТТ "Информика". - М.: Просвещение, 2007. - С. 440-464. URL: http://window.edu.ru/window_catalog/ redir?id=45290&file=440-464.pdf

[5] Арнольд В.И. Аналитика и прогнозирование: математический аспект // Научно-техническая информация. - Сер. 1. - Вып. 3. - 2003. - С. 1-10.

[6] Арнольд В.И. Обыкновенные дифференциальные уравнения. - М.: Наука, 1971, - 240 с.

[7] Бак П., Чен К. Самоорганизованная критичность // В мире науки, 1991. - №3. - С. 16-24.

[8] Божокин С.В., Паршин Д.А. Фракталы и мультифракталы. - Ижевск: НИЦ «Регулярная и хаотическая динамика», 2001. - 128 с.

[9] Босс В. Лекции по математике. Том 4. Вероятность, информация, статистика. – М.: КомКнига, 2005. - 216 с.

[10] Брайчевский С.М., Ландэ Д.В. Современные информационные потоки: актуальная проблематика // Научно-техническая информация. - Сер. 1. - Вып 11. -2005. - С. 21-33. URL: http://dwl.visti.net/art/nti05/

[11] Вольтерра В. Математическая теория борьбы за существование. - М.: Наука, 1976.

[12] Гарднер М. Математические досуги. – М.: Мир, 1972.

[13] Гринченко В.Т., Мацыпура В.Т., Снарский А.А. Введение в нелинейную динамику. Хаос и фракталы. Изд. 2. - М: УРСС, 2007. - 263 с.

[14] Гуркин Ю.Н. Семенов Ю.А. Файлообменные сети P2P: основные принципы, протоколы, безопасность // “Сети и системы связи” - № 11. - Стр. 62. - 2006. URL: http://www.ccc.ru/magazine/depot/06_11/read.html?0302.htm

[15] Данилов Ю.А. Лекции по нелинейной динамике. Изд 2-е. - М.: КомКнига, УРСС, 2006. - 208 с.

[16] Джейн А.К. Введение в искусственные нейронные сети // Открытые системы. - № 4. -1997. URL: http://www.osp.ru/text/302/179189/

[17] Дмитриев В.И. Прикладная теория информации: Учеб. для студ. Вузов по спец. «Автоматизированные системы обработки информации и управления». – М.: Высш. шк., 1989. – 320 с.

[18] Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений. – М.: Мир, 1973. – 165 с.

[19] Зеленый Д.М. Основы теории информации и оптимального приема. – Ленинград: Изд-во ВМА, 1966.

[20] Иванов С.А. Стохастические фракталы в информатике // Научно-техническая информация. - Сер. 2. - Вып 8. - 2002. - С. 7-18.

[21] Иванов С.А. Ранговые распределения в информатике // Научно-техническая информация. - Сер. 2. - Вып. 12. - 1985. - С. 14-19.

[22] Иванов С.А. Устойчивые закономерности мировой системы научной коммуникации // Научно-техническая информация. - Сер. 2. - Вып. 1. - 2003. - С. 1-7.

[23] Иванов С.А., Круковская Н.В. Статистический анализ документальных информационных потоков // Научно-техническая информация. - Сер. 2. - Вып. 2. - 2004. - С. 11-14.

[24] Иудин Д.И., Гелашвили Д.Б., Розенберг Г.С. Мультифрактальный анализ видовой структуры биотических сообществ // Докл. Акад. Наук. – Т. 389. - № 2. – 2003. – С. 279-282.

[25] Кириченко К.М, Герасимов М.Б. Обзор методов кластеризации текстовых документов // Материалы международной конференции Диалог'2001, URL: http://www.dialog-21.ru/Archive/2001/volume2/2_26.htm

[26] Кларк Д. Закон Мура останется в силе // Ведомости. - 2003. - № 11. URL: http://www.silicontaiga.ru/home.asp?artId=2066

[27] Колмогоров А.Н. Три подхода к определению понятия «количество информации» // Проблемы передачи информации, 1965. - Т. 1. - Вып. 1. - С. 25-38

[28] Кришнан Н. JXTA решения для P2P // Java World. -№ 10, 2001. URL: http://www.javaworld.com/javaworld/ jw-10-2001/jw-1019-jxta.html

[29] Кузьмин И.В., Кедрус В.А. Основы теории информации и кодирования. – К.: Вища школа, 1977. – 280 с.

[30] И. Кураленок, И. Некрестьянов. Оценка систем текстового поиска // Программирование. - № 28(4). - 2002. - С. 226-242.

[31] Ландэ Д.В. Основы интеграции информационных потоков. – К.: Инжиниринг, 2006. - 240 с. URL: http://dwl.visti.net/art/monogr-osnov/ spusk3.pdf

[32] Ландэ Д.В. Поиск знаний в Internet. - М.: Диалектика-Вильямс, 2005. URL: http://poiskbook.kiev.ua

[33] Лившиц Ю. Курс лекций "Алгоритмы для Интернета". URL: http://logic.pdmi.ras.ru/~yura/internet.html

[34] Лоскутов А.Ю., Михайлов А.С. Введение в синергетику. - М.: Наука, 1990.

[35] Нехаев С.А., Кривошеин Н.В., Андреев И.Л., Яскевич Я.С. Словарь прикладной интернетики. – Сетевой холдинг WEB-PLAN Group, 2001. URL: http://www.nbuv.gov.ua/texts/libdoc/01nsaopi.htm

[36] Мандель И.Д. Кластерный анализ. - Г.: Финансы и статистика, 1988. – 176 с.

[37] Мандельброт Б. Фрактальная геометрия природы. - М.: Институт компьютерных исследований, 2002. - 656 с.

[38] Мандельброт Б. Фракталы, случай и финансы. - М.: Регулярная и хаотическая динамика, 2004. - 256 с.

[39] Мирзаджанзаде А.Х., Хасанов М.М. , Бахтизин Р.Н. Моделирование процессов нефтегазодобычи. Нелинейность, неравномерность, неопределенность. - М.-Ижевск: Институт компьютерных исследований, 2005. - 368 с.

[40] Некрестьянов И.С., Добрынин В.Ю., Клюев В.В. Оценка тематического подобия текстовых документов // Труды второй всероссийской научной конференции “Электронные библиотеки”. - Протвино, 2000. - С. 204-210.

[41] Нейман Дж. Теория самовоспроизводящихся автоматов. - М.: Мир, 1971. – 382 с.

[42] Павлов А.Н., Сосновцева О.В., Зиганшин А.Р. Мультифрактальный анализ хаотической динамики взаимодействующих систем // Изв. вузов, Прикладная нелинейная динамика. -Т. 11, - № 2, 2003. - С. 39-54.

[43] Плотинский Ю.М. Модели социальных процессов. – Изд. 2-е. – М.: Логос, 2001. – 296 с.

[44] Попов А. Поиск в Интернете - внутри и снаружи // Intrnet. - 1998. - № 2. URL: http://www.citforum.ru/pp/search_03.shtml

[45] Рычагов М.Н. Нейронные сети: многослойный перцептрон и сети Хопфилда // EXPonenta Pro. Математика в приложениях, 2003. - № 1. URL: http://nature.web.ru/db/msg.html?mid=1193685

[46] Сегалович И.В. Как работают поисковые системы // Мир Internet. – 2002. - № 10. URL: http://www.dialog-21.ru/direction_fulltext.asp?dir_id=15539

[47] Семенов Ю.А. Алгоритмы телекоммуникационных сетей. Часть 1. Алгоритмы и протоколы каналов и сетей передачи данных. –М: Интернет-университет информационных технологий - ИНТУИТ.ру, 2007. – 640 с. URL: http://book.itep.ru/1/intro1.htm

[48] Снарский А.А., Безсуднов И.В., Севрюков В.А. Процессы переноса в макроскопических неупорядоченных средах: От теории среднего поля до перколяции. – М.: УРСС, Изд-во ЛКИ, 2007. -304 с.

[49] Снарский А.А., Ландэ Д.В., Григорьев А.Н., Брайчевский С.М., Дармохвал А.Т. Ранжирование сайтов «по Хиршу» // Доклады международной конференции «MegaLing'2006 Горизонты прикладной лингвистики и

лингвистических технологий». 20-27 сентября 2006, Украина, Крым, Партенит. - С. 248-249.

[50] Солодов А.В. Теория информации и ее применение к задачам автоматического управления и контроля. – М.: Наука, 1967.

[51] Таненбаум Э., ван Стеен М. Распределенные системы: принципы и парадигмы. - Спб.: Питер, 2003. - 876 с.

[52] Тарасевич Ю.Ю., Перколяция: теория, приложения, алгоритмы. - М., УРСС, 2002. – 112 с.

[53] Тоффоли Т., Марголус Н. Машины клеточных автоматов. - М.: Мир, 1991. - 280 с.

[54] Труды четвертого российского семинара РОМИП'2006 (Суздаль, 19 октября 2006 г.) - СПб: НУ ЦСИ, 2006, 274 с. URL: http://romip.narod.ru/ romip2006/index.html

[55] Хан У., Мани И. Системы автоматического реферирования // Открытые системы, 2000. - № 12. URL: http://www.osp.ru/os/2000/12/ 067.htm

[56] Уоссермен Ф. Нейрокомпьютерная техника. - М.: Мир, 1992. - 184 с.

[57] Фаддеев Д.К. К понятию энтропии конечной вероятностной схемы // Успехи матем. наук. - Т. 11. - № 1, 1956. - С. 227-231.

[58] Федер Е. Фракталы. - М.: Мир, 1991. - 254 с.

[59] Хартли Р. Передача информации. Теория информации и ее приложения / под. ред. А.А. Харкевича. – М.: Физматгиз, 1959.

[60] Хехт-Нильсен Р. Нейрокомпьютинг: история, состояние, перспективы // Открытые системы. - № 4. -1998. URL: http://www.osp.ru/text/302/179534/

[61] Шелухин О.И., Тенякшев А.М., Осин А.В. Фрактальные процессы в телекоммуникациях. - М.: Радиотехника, 2003. - 480 с.

[62] Шеннон К. Работы по теории информации и кибернетике. – М.: ИЛ, 1963.

[63] Эфрос А.Л., Физика и геометрия беспорядка. - М., Наука, 1982. – 176 с.

[64] Яглом А.М., Яглом И.М. Вероятность и информация. - М.: Наука, 1973. – 512 с.

[65] Albert R., Jeong H., Barabasi A. Attack and error tolerance of complex networks // Nature. - 2000. - Vol. 406. - pp. 378-382.

[66] APPN/HPR in IP Networks (APPN Implementers' Workshop Closed Pages Document). IBM. URL: http://www.javvin.com/protocol/rfc2353.pdf

[67] Avram H.D., Knapp J.F., RatherL.J. The MARC II Format: A Communications Format for Bibliographic Data, Library of Congress. -Washington, D.C., 1968.

[68] Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval. - ACM Press Series/Addison Wesley, New York, 1999. – 513 p.

[69] Bak P. How nature works: The science of self-organized criticality. Springer-Verlag, New York, Inc., 1996.

[70] Bak P., Tang C., Wiesenfeld K. Self-organized criticality: An explanation of 1/f-noise // Phys. Rev. Lett. 1987. –Vol . 59, - pp. 381-384.

[71] Bak P., Tang C., Wiesenfeld K. Self-organized criticality// Phys. Rev. A., 1988. – Vol. 38. -№ 1. - pp. 364-374.

[72] Bandini S., Mauri G., Serra R. Cellular automata: From a theoretical parallel computational model to its application to complex systems // Parallel Computing. – Vol. 27, Issue 5, April 2001. - pp. 539-553.

[73] Bell A., Fosler-Lussier E., Girand C., Raymond W. Reduction of English function words in Switchboard // Proceedings of ICSLP-98. - Vol 7. – 1998. - pp. 3111-3114.

[74] Berners-Lee T., Hendler J., Lassila O. The Semantic Web. Scientific American, 2001. URL: http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

[75] Berry M.W. Survey of Text Mining. Clustering, Classification, and Retrieval. - Springer-Verlag, 2004. - 244 p.

[76] Bhargava S.C., Kumar A., Mukherjee A. A stochastic cellular automata model of innovation diffusion // Technological forecasting and social change, 1993. - Vol. 44. - № 1. - pp. 87-97.

[77] Bjorneborn, L., Ingwersen, P. Toward a basic framework for webometrics. Journal of the American Society for Information Science and Technology, 55(14): 1216-1227. – 2004.

[78] Boyle A. Net not as interconnected as you think. URL: http://news.zdnet.com/2100-9595_22-502388.html

[79] Bradford, S.C. "Sources of Information on Specific Subjects". Engineering: An Illustrated Weekly Journal (London), 137, 1934 (26 January), - pp. 85-86.

[80] Brin S., Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. WWW7, - 1998.

[81] Broadbent S.R., Hammersley J.M. Percolation processes // I. Crystals and mazes, Proc Cambridge Philos. Soc. – pp. 629-641. – 1957.

[82] Broder A. Identifying and Filtering Near-Duplicate Documents, COM'00 // Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching. – 2000. – pp. 1-10.

[83] Broder A., Kumar R., Maghoul F. etc. Graph structure in the Web. // Proceedings of the 9th international World Wide Web conference on Computer networks: the international journal of computer and telecommunications networking. - Amsterdam, 2000. - pp. 309-320. URL: http://www.almaden.ibm.com/cs/k53/www9.final/

[84] CJC Burges. A Tutorial on Support Vector Machines for Pattern Recognition. URL: http://www.music.mcgill.ca/_rfergu/adamTex/references/Burges98.pdf

[85] Cohen R., Erez K., ben-Avraham D., Havlin S. Resilience of the Internet to. Random Breakdown // Phys.Rev.Lett. 85, 4626 (2000).

[86] Donetti L., Hurtado P.I., Munoz M.A. Entangled Networks, Synchronization, and Optimal Network Topology // Physical Review Letters. - Vol. 95, - № 18, 2005.

[87] Dorogovtsev S.N., Mendes J.F.F. Evolution of Networks: from biological networks to the Internet and WWW, Oxford University Press, 2003.

[88] Erdős, P., Rényi A. On Random Graphs. I. // Publicationes Mathematicae 6, - pp. 290-297. -1959.

[89] Erdős P., Rényi A. On the evolution of random graphs, Publ. Math. Inst. Hungar. Acad. Sci. 5. - pp. 17-61. -1960.

[90] Fano U. Ionization yield of radiations. II. The fluctuations of the number of ions. Phys. Rev., 72. – pp. 26-29. -1947.

[91] Fox G.C. From Computational Science to Internetics: Integration of Science with Computer Science, Mathematics and Computers in Simulation, Elsevier, 54 (2000) 295-306 . URL: http://www.npac.syr.edu/users/gcf/internetics2/

[92] Fox G.C. Internetics: Technologies, Applications and Academic Fields // Invited Chapter in Book: Feynman and Computation, edited by A.J.G. Hey, Perseus Books (1999). Technical Report SCCS-813, Syracuse University, NPAC, Syracuse, NY, February 1998. URL: http://www.new-npac.org/users/fox/documents/internetics/

[93] Furnas G.W., Deerwester S., Dumais S.T., etc. Information retrieval using a Singular Value Decomposition Model of Latent Semantic Structure. - ACM SIGIR, 1988.

[94] Del Corso G.M., Gullí A., Romani F. Ranking a stream of news. International World Wide Web Conference // Proceedings of the 14th international conference on World Wide Web. Chiba, Japan, 2005. - pp. 97 - 106.

[95] Graham P. A Plan for Spam. - 2002. URL: http://paulgraham.com/spam.html.

[96] Grootjen F.A., Van Leijenhorst D. C., van der Weide T.P. A formal derivation of Heaps' Law // Inf. Sci. – Vol. 170(2-4). - pp. 263-272. - 2005. URL: http://citeseer.ist.psu.edu/660402.html

[97] Hei X., Liang Ch., Liu Y., Ross K.W. Insight into PPLive: A Measurement Study of a Large-Scale P2P IPTV System. URL: http://photon.poly.edu/~jliang/pplive.pdf

[98] Heaps H.S. Information Retrieval - Computational and Theoretical Aspects. Academic Press, 1978.

[99] Hinrichsen H. Nonequilibrium Critical Phenomena and Phase Transitions into Absorbing States Adv. in Phys. 49, 815 (2000). URL: http://arxiv.org/abs/cond-mat/0001070v2

[100] Hirsch J.E. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the USA, 102(46), 16569-16572. - 2005.

[101] Hofmann T. Probabilistic latent semantic indexing. In Proc. of the SIGIR'99. - 1999. - pp. 50-57.

[102] Hurst H. E. Long-term storage capacity of reservoirs. // Trans. Amer. Soc. Civil Engineers 116. – pp. 770-799. -1951.

[103] Ilyinsky S., Kuzmin M., Melkov A., Segalovich I. An efficient method to detect duplicates of Web documents with the use of inverted index // WWW2002, 2002.

[104] Kalogeraki V., Gunopulos D., Zeinalipour-Yazti D. A Local Search Mechanism for Peer-to-Peer Networks. // Proc. of CIKM'02, McLean VA, USA, 2002.

[105] Kleinberg J.M. Authoritative sources in a hyperlink environment. // In Processing of ACM-SIAM Symposium on Discrete Algorithms, 1998, 46(5):604-632.

[106] Landauer, T.K., Foltz, P.W., Laham, D. An introduction to latent semantic analysis. - Discourse Processes. - Vol. 25. - 1998. - pp. 259-284.

[107] Lande D. Model of information diffusion // Preprint Arxiv (0806.0283), 2008. – 5 p. URL: http://arxiv.org/abs/0806.0283

[108] Lempel R. and Moran S. The stochastic approach for link-structure analysis (SALSA) and the TKC effect // In Proceedings of the 9th International World Wide Web Conference, Amsterdam, The Netherlands, 2000. - pp. 387–401.

[109] Lu Q., Cao P., Cohen E., Li K., Shenker S. Search and replication in unstructured peer-to-peer networks. // Proc. of ICS02, New York, USA, June 2002.

[110] Manber U. Finding similar files in a large file system. Proceedings of the 1994 USENIX Conference, pp. 1-10, January 1994.

[111] Manning C.D., Schütze H. Foundations of Statistical Natural Language Processing - Cambridge, Massachusetts: The MIT Press, 1999.

[112] Maymounkov P., Mazi`eres D. Kademlia: A Peer-to-peer Information System Based on the XOR Metric. URL: http://kademlia.scs.cs.nyu.edu

[113] Milgram S. The small world problem, Psychology Today, 1967, - Vol. 2. - pp. 60-67.

[114] Miller E., Swick R., Brickley D., McBride B., Hendler J., Schreiber G., Connolly D. Semantic Web. W3C (MIT, ERCIM, Keio) - 2001. URL: http://www.w3.org/2001/sw/

[115] Mockapetris P. Domain Names - Concepts and Facilities // Request for Comments: 1035, - 1987. – 55 p.

[116] Newman M.E.J. The structure and function of complex networks // SIAM Review. - 2003. - Vol. 45. - pp. 167-256.

[117] Newman M.E.J., Watts D.J. Scaling and percolation in the small-world network model, Phys.Rev.E,7332,1999.

[118] Onnela J.-P., Saramaki J., Hyvonen J., Szabo G., Lazer D., Kaski K., Kertesz J., Barabasi, A.-L. Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences. May 1, 2007, vol. 104. no. 18, 7332-7336.

[119] Page S. E. Computational models from a to z // Complexity. –Vol. 5, Issue 1, 1999. – pp. 35 – 41.

[120] Papka, R. On-line News Event Detection, Clustering, and Tracking. Ph. D. Thesis, University of Massachusetts at Amherst, September 1999.

[121] Peng C.-K., Havlin S., Stanley H.E., Goldberger A.L. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series // Chaos. - Vol. 5. - 1995. - pp. 82.

[122] Piatetsky-Shapiro G., Fayyad U., Smith P. - Advances in Knowledge Discovery and Data Mining. - Cambridge, Mass: AAA/MIT Press. – pp. 1-35. -1996.

[123] Platt J. Sequential Minimal Optimization. URL: http://research.microsoft.com/ users/jplatt/smo.html

[124] Powell A.L., French J.C., CallanJ., Connell M., Viles C.L. The Impact of Database Selection on Distributed Searching // Proc. of ACM SIGIR'00, pages 232{239, Athens, Greece, 2000.

[125] Program to evaluate TREC results using SMART evaluation procedures. URL: http://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/ trec_eval/ README

[126] Redner S., Directed and diode percolation. Phys. Rev. B, 25, 3242,1982.

[127] RFC1625 - WAIS over 39.50-1988. Network Working Group. Request for Comments: 1625. M.St. Pierre, J. Fullton, K. Gamiel, J. Goldman, B. Kahle, J. Kunze, H. Morris, F. Schiettecatte, 1994. URL: http://www.faqs.org/rfcs/ rfc1625.html

[128] Riedi R.H., Vehel J.L. Multifractal Properties of TCP traffic: a numerical study. // Technical Report № 3128 INRIA Rocquencourt. - March 1997.

[129] Rocchio, J. Relevance feedback in information retrieval // In G. Salton ed., The SMART Retrieval System: Experiments in Automatic Document Processing, Englewood Cliffs, New Jersey, Prentice-Hall, pp. 313-323, 1971.

[130] Salton G., Fox E., Wu H. Extended Boolean information retrieval. Communications of the ACM. -2001. -Vol. 26. -№ 4. - pp. 35-43.

[131] Salton G, Wong A, Yang C. A Vector Space Model for Automatic Indexing. // Communications of the ACM, 18(11):613-620, 1975.

[132] Sarshar N., Boykin P.O., Roychowdhury V.P. Scalable Percolation Search in Power Law Networks. Preprint. - 2004. URL: http://arxiv.org/abs/cond-mat/0406152

[133] Scime A. Web mining: application and techniques. - Idea Group Publishing, 2005. - 427 p.

[134] Sebastiani F. Machine Learning in Automated Text Categorization. URL: http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf

[135] Simon H. A. Biometrika 42, 425 (1955).

[136] Snarskii A. FreeBSD Stack Integrity Patch. 1997. URL: ftp://ftp.lucky.net/pub/unix/local/libc-letter

[137] Soumen C. Mining the web. Discovery knowledge from hypertext data. - Publisher: Morgan Kaufmann, 2002. - 344 p.

[138] Stanley H.E. , Amaral L.A.N., Goldberger A.L., Havlin S., Ivanov P.Ch., Peng C.-K. Statistical physics and physiology: monofractal and multifractal approaches // Physica A. 1999. - Vol. 270, - pp. 309.

[139] Stauffer D., Aharony A., Introduction to percolation theory. – Taylor & Francis, London, Washington DC, 1992. – 182 p.

[140] Stanley H.E., Amaral L.A.N, Goldberger A.L., Havlin S., Ivanov P.Ch., Peng C.-K. Statistical physics and physiology: monofractal and multifractal approaches. // Physica A. 1999. - Vol. 270. - pp. 309.

[141] The Deep Web: Surfacing Hidden Value, 2000 BrightPlanet.com LLC, 35 p. URL: http://www.dad.be/library/pdf/BrightPlanet.pdf

[142] The Twelfth Text Retrieval Conference (TREC 2003). Appendix 1. Common Evaluation Measures. URL: http://trec.nist.gov/pubs/trec12/

[143] Ukkonen E. On-line construction of suffix trees URL: http://www.cs.helsinki.fi/u/ukkonen/SuffixT1withFigs.pdf

[144] Understanding the Impact of P2P: Architecture and Protocols URL: http://www.cachelogic.com/home/pages/understanding/architecture.php

[145] Van Raan A.F.J. Fractal geometry of Information Space as Represented by Cocitation Clustering // Scientometrics. -1991. - Vol. 20, -№ 3. - pp. 439-449.

[146] Vapnik V.N. Statistical Learning Theory. NY: John Wiley, 1998. – 760 p.

[147] Watts D.J., Strogatz S.H. Collective dynamics of “small-world” networks. // Nature. - 1998. - Vol. 393. pp. 440-442.

[148] Wikipedia, Support Vector machine. URL: http://en.wikipedia.org/wiki/Support_vector_machine

[149] Wolfram S. A New Kind of Science. - Champaign, IL: Wolfram Media Inc., 2002. – 1197 pp.

[150] Wolfram S. ed. Theory and Applications of Cellular Automats. - Singapore: World Scientific. 1986.

[151] Yang B., Garcia-Molina H. Comparing hybrid peer-to-peer systems. // Proc. of VLDB'01, Rome, Italy, 2001.

[152] Yang B., Garcia-Molina H. Efficient Search in Peer-to-Peer Networks. // Proc. of ICDCS'02, Vienna, Austria, 2002.

[153] Yeager N., McCrath R. WebServer Technology. - Morgan Kaufmann, San Francisco, California, 1996.

[154] Zamir O.E. Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results. PhD Thesis, University of Washington, 1999.

[155] Zeinalipour-Yazti D. Information Retrieval in Peer-to-Peer Systems. // M.Sc Thesis, Dept. of Computer Science, University of California - Riverside, June 2003.

[156] Zeinalipour-Yazti D., Kalogeraki V., Gunopulos D. Information Retrieval in Peer-to-Peer Networks // IEEE CiSE Magazine, Special Issue on Web Engineering, 2004. – pp. 1-13. URL: www.cs.ucr.edu/~csyiazti/papers/cise2003/cise2003.pdf

[157] Zhou S., Mondragon R.J. Topological Discrepancies Among Internet Measurements Using Different Sampling Methodologies, Lecture Notes in Computer Science (LNCS), Springer-Verlag, - № 3391, - pp. 207-217, Feb. 2005.

Об авторах

Дмитрий Владимирович Ландэ

Доктор технических наук, заместитель директора Информационного центра «ЭЛВИСТИ», профессор кафедры Института специальной связи и защиты информации Национального технического университета Украины «Киевский политехнический институт». Научные интересы: теория информационного поиска, компьютерная лингвистика, методы детерминированного хаоса в информационных потоках, исследование сложных сетей. Автор монографий «Моделирование информационно-электоральных процессов» (Академия правовых наук Украины, 2007), «Поиск знаний в Internet» (Диалектика-Вильямс, 2005).

Андрей Александрович Снарский

Доктор физико-математических наук, профессор кафедры общей и теоретической физики физико-математического факультета Национального технического университета Украины «Киевский политехнический институт». Научные интересы: термоэлектрические явления в анизотропных и неоднородных средах, теория протекания, методы детерминированного хаоса в информационных потоках, магнитная дефектоскопия. Автор монографий «Введение в нелинейную динамику. Хаос и фракталы» (2-е изд. URSS, 2007), «Процессы переноса в макроскопических неупорядоченных средах» (URSS, 2007).

Игорь Васильевич Безсуднов

Заместитель директора Научно-производственного предприятия "Наука-Сервис". Научные интересы: явления в средах с перколяцией, самоорганизованная критичность, математическое и компьютерное моделирование систем с протеканием, методы определения примесей драгметаллов и ртути в объектах окружающей среды. Автор изобретений, монографии «Процессы переноса в макроскопических неупорядоченных средах» (URSS, 2007).