ФЕДЕРАЛЬНОЕ АГЕНТСТВО НАУЧНЫХ ОРГАНИЗАЦИЙ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 1 Москва 2018

ОБЩИЙ РАЗДЕЛ

УДК 17 : 004

В.А. Канке

Метанаучные и философские основания информационной этики*

Развитие аксиологических теорий остро поставило вопрос об их обогащении этическим началом. Показан путь перевода этики на научные рельсы, состоящий во включении во все аксиологические теории принципа максимизации благоденствия всех людей, затрагиваемых той или иной ситуацией. Информатика, будучи формальной наукой, не содержит этический принцип максимизации благоденствия непосредственно в себе, но в силу вовлеченности в междисциплинарные отношения со всеми аксиологическими науками, она обладает исключительно актуальной этической относительностью. Выяснение метанаучных оснований этический относительности информатики открывает возможности для ее всестороннего изучения с позиций науки, а не плохо понятой метафизики.

Ключевые слова: *метанаука, этика, компьютерная и информационная этика, этическая относительность информатики*[?]

[?] Статья подготовлена при финансовой поддержке РФФИ и Правительства Калужской области (проект 17-13-40004).

ВВЕДЕНИЕ

Информатика как область исследований, удовлетворяющая самым высоким научным стандартам, сложилась в 1930-гг. благодаря, в первую очередь, работам К. Гёделя, А. Чёрча и А.М. Тьюринга [1-3]. Понадобилось несколько десятков лет, прежде чем в поле зрения исследователей попала этическая проблематика. В связи с этим остро встал вопрос о статусе информационной этики или же ее своеобразного коррелята – компьютерной этики. В ныне считающейся классической работе Дж.Х. Мур достаточно отчетливо расставил некоторые исходные акценты, без которых статус компьютерной этики остается недоопределенным. "В некотором смысле, – отмечал он, – я отстаиваю специальный статус компьютерной этики как особой области изучения. Прикладная этика – это не просто применяемая этика. Тем не менее, я также желаю подчеркнуть фундаментальное значение для компьютерной этики общей этики и науки. Этическая теория предоставляет категории и процедуры для определения того, что является этически релевантным. Например, какие вещи хороши? Каковы наши основные права? Что такое беспристрастная точка зрения? Эти соображения имеют важное значение для сопоставления и обоснования политики этического поведения. Соответственно, научная информация имеет решающее значение в этических оценках" [4, с. 267].

Дж.Х. Мур был прав, статус компьютерной этики невозможно определить без тщательного рассмотрения соотносительности науки, этики и информатики. На наш взгляд, как раз в этом отношении и он сам, и другие исследователи преуспели в меньшей степени, чем хотелось бы. Возобладали две точки зрения. Согласно первой из них, традиционная философская этика в слегка измененном виде прилагается к информатике [4–6]. Этой позиции придерживаются также многочисленные авторы кодексов компьютерной этики. Другая популярная точка зрения состоит в том, что информатика вызывает к жизни некоторую всеобъемлющую глобальную этику, которая поглощает все другие, в том числе и традиционную философскую этику [7, 8]. Сторонники второй точки зрения, как правило, выступают от имени информационной этики. Их оппоненты предпочитают использовать термин компьютерная этика. Чтобы обосновать нашу точку зрения, нам придется рассмотреть сначала соотношение науки и этики, затем – науки и информатики, а также информатики и этики.

НАУКА И ЭТИКА

Соотношение науки и этики отмечено печатью довольно неожиданных метаморфоз. Важной вехой на пути развития этики стали три книги Аристотеля, особенно его "Никомахова этика", написанная в IV в. до н.э. [9]. Никогда ранее этика не рассматривалась столь пространно и обстоятельно. Аристотель считал, что моральные добродетели, например, мужество и благоразумие, контролируют страсти по избытку и по недостатку. По сути, он руководствовался известным принципом античности: ничего слишком.

Аристотель творил в эпоху отсутствия развитых аксиологических наук, в частности, общественных. В этих условиях этика не могла приобрести полновесный научный статус. Дело в том, что как учение о правильном поведении она остро нуждается в поддержке аксиологических наук. Неудивительно, что более 2000 лет этика пребывала в зачаточном состоянии. Лишь в XIX в. стали стремительно набирать вес общественные науки, в частности, экономика, политология, социология и юриспруденция. Разумеется, предстояло осмыслить их статус, для чего требовалась развитая философия науки. Она в первоначальном ее варианте пришла в образе позитивизма О. Конта и Дж.С. Милля. Позитивизм имел важнейшее значение для будущего этики, причем далеко не во всем положительное.

Первые позитивисты относились к этике довольно благосклонно. Не случайно двое из них, И. Бентам и Дж.С. Милль стали основателями утилитаризма с его принципом максимального счастья для максимально большого числа людей. И. Бентам настаивал на нормировании "фабрики благоденствия в соответствии с указаниями разума и права" [10, с. 4]. Утилитаристы стремились перевести этику на научные рельсы. Но с позиций развитого позитивизма эта попытка была признана неудачной. Дело в том, что он предполагает отчетливое указание на объективные факты. Ссылки утилитаристов на чувства наслаждения и страдания в качестве фактов неопозитивистами XX в., как правило, не признавалась. Эту позицию наиболее отчетливо сформулировал близкий к позитивизму аналитический философ Л. Витгенштейн, утверждавший, что "невозможны предложения этики" [11 с. 70]. Он полагал, что подлинные науки имеют дело с фактами, а не с ценностями. Л. Витгенштейн и неопозитивисты А.Дж. Айер и Ч.Л. Стивенсон решительно вытеснили этику за пределы науки [11–13]. В действительности нет никаких оснований противопоставлять факты и ценности. Факты – это результаты наблюдений и экспериментов, которые есть в любой как естественной, так и аксиологической теории. Есть, например, физические и экономические факты. Ценности же являются переменными аксиологических теорий. Ценностями являются, например, быстродействие компьютера, цена товара, свобода вероисповеданий.

Отношение позитивистов к этике было настолько негативным, что они вытеснили ее не только за пределы науки, но и философии. Восстановление философского авторитета этики в англоязычном мире состоялась во многом благодаря Р.М. Хэару, который утверждал, что этические предложения приемлемы постольку, поскольку они являются предписаниями, необходимыми для совершения некоторых действий в конкретных ситуациях [14]. Но, на наш взгляд, ему, равно как пропагандистам феноменологической [15], герменевтической [16] и постструктуралистской [17] и многих других вариантов философской этики, не удалось перевести ее на научные рельсы. Это утверждение нуждается в пояснении, которое мы приводим далее.

МЕТАНАУКА И ЭТИКА

Наука – это мир теорий. Каждая теория представляет собой достаточно сложное для понимания образование. Разумеется, без его понимания невозможно определить место этики в мире науки. По определению, осмысление теорий происходит в метатеориях. Отсюда следует, что метанаука есть ключ к пониманию науки как совокупности теорий. Прежде всего, в метатеориях рассматриваются концепты, образующие их состав, и методы управления этими концептами; непременно значительное внимание должно быть уделено также методам управления теориями [18]. Объектами теории являются сущие, например, элементарные частицы в физике, звезды в космологии, отдельные люди и организации в аксиологических науках. Сущие в теориях представлены принципами, законами и переменными. Из этих трех разновидностей концептов наиболее емким содержанием обладают принципы, например, принцип наименьшего действия в физике или принцип максимизации прибыли в экономике. Родственные теории образуют ряды теорий, лигатеории. Лигатеориями являются, например, электродинамика Дирака-Эйнштейна-Максвелла, теория трудовой стоимости Маркса-Риккардо-Смита. Различные лигатеории связаны друг с другом символической связью. Например, биологи часто вынуждены учитывать влияние на живые организмы физических факторов, в частности, излучений. Их интерес сосредоточен не на самих факторах как таковых, а на их значимости для биологических явлений. Для биологов физические явления являются символами биологических процессов. Соответственно экономисты рассматривают политические институты как символы экономических процессов, политологи, напротив, оценивают экономические факторы как символы политических явлений. Приведенная характеристика метанауки достаточна для рассмотрения статуса научной этики.

Во-первых, очевидно, что по своему статусу этические теории относятся не к естественным или формальным, а к аксиологическим концепциям. Во-вторых, нет оснований для зачисления их в разряд автономных аксиологических теорий, столь же самостоятельных как, например, экономические и политические концепции. Можно привести сотни примеров моральных или же аморальных поступков, но все они относятся к разряду вполне определенных явлений, теориями которых не является этика. Экономическая коррупция, подавление политической демократии, установление несправедливых юридических законов по определению относятся соответственно к экономике, политологии и юриспруденции, а не к этике как таковой. Из отмеченного следует, что этика не может рассматриваться как символ некоторой аксиологической теории. В связи с этим остаются лишь две возможности: либо этика причастна непосредственно к устройству аксиологических теорий, либо ее придется по примеру позитивистов перевести в разряд ненаучных концептов. Наша позиция заключается в том, что этическое содержание должно содержаться непосредственно в каждой аксиологической теории. Это обстоятельств нуждается в доказательстве.

Возраст этического проекта, если его отсчитывать от Аристотеля, составляет около 2,5 тыс. лет. Разумеется, в рамках статьи не могут быть рассмотрены все метаморфозы этого проекта. Отметим самое главное. Как правило, содержанием всех авторитетных этических теорий, в частности, тех, авторами которых являются Аристотель, И. Кант, И. Бентам и Дж.С. Милль, Р. Хэар, Дж. Дьюи, Ю. Хабермас, является принцип максимизации благоденствия всех людей, затрагиваемых той или иной ситуацией, т.е. стейкхолдеров. Благоденствие понимается в данном случае как некоторый ценностный конструкт, определяемый содержанием теорией. Им может быть, например, доход частного предпринимателя, улучшение здоровья пациента, повышение производительности труда, равно как и многое другое.

Итак, от имени этики научным аксиологическим теориям может и должен быть представлен принцип максимизации благоденствия стейкхолдеров. Способны ли эти теории принять его в свое лоно или же отвергают его? Нетрудно обнаружить, что все аксиологические теории прекрасно совместимы с принципом максимизации благоденствия. Проиллюстрируем это обстоятельство на примере экономической теории.

Наиболее часто экономисты руководствуются принципом максимизации прибыли. Однако не только им, но и другими принципами, например, принципами максимизации объема продаж или роста занятости населения. Если эти принципы приводят к росту благосостояния исключительно богатых людей, а не бедных, то они в этическом отношении несостоятельны. Если же изначально руководствоваться принципом максимизации благосостояния всех людей, то каждый другой принцип приобретает этическую состоятельность. Нечто аналогичное происходит при включении принципа максимизации благоденствия в любую аксиологическую теорию. При этом показательно, что он в иерархии принципов непременно занимает лидирующую позицию. Если, например, подчинить принцип максимизации благоденствия принципу максимизации прибыли, то его просто не удастся реализовать, не приняв установку на его неукоснительное выполнение.

Таким образом, этический принцип должен включаться непосредственно в состав всех аксиологических теорий, а именно – технических, аграрных, медицинских, психологических, педагогических, экономических, политологических, социологических, юридических и исторических. Если это не делается, то этика обречена на прозябание в области метафизики, не выдерживающей научной критики. Именно это тревожное обстоятельство наблюдается во всех современных основных философских направлениях, в частности, в аналитической философии, феноменологии, герменевтике и постструктурализме. Абсолютное большинство представителей этих теорий выступают от имени не научной, а метафизической этики. Философия может быть причастна к научному делу, но лишь определенным образом.

Допустим, этическое начало реализуется в рамках ядерной энергетики, медицины и политологии. В свя-

зи с этим правомерно рассуждать об этике ядерной энергетики, а также медицинской и политологической этике. В силу разделения научного труда представители каждой из трех указанных отраслей науки остаются в ее пределах. Они не рассматривают схожие черты различных этических теорий, выходящие за пределы их компетенции. В связи с этим открывается поле исследования для философов. Именно им сподручно изучать схожие черты различных этических теорий. Отмеченное означает, что научная философская этика имеет формальный характер. В этом ее качестве она имеет актуальное значение, которое ни в коей мере не умаляет актуальность философии. Философская этика как формальная концепция актуальна, но не в качестве метафизической теории.

К сожалению, от имени философии науке предлагается не формальная, а как раз метафизическая этика. Далее мы рассмотрим пути придания ей научной актуальности. Пока же мы вынуждены рассматривать философскую этику в ее ненаучном облачении.

ИНФОРМАЦИОННАЯ И КОМПЬЮТЕРНАЯ ЭТИКА

Содержание предыдущих разделов позволяет перейти непосредственно к определению статуса компьютерной и информационной этики. В связи с этим необходимо, прежде всего, определиться со статусом информатики. Если она является аксиологической наукой, то вполне правомерно ставить вопрос об актуальности информационной этики. В противном случае само утверждение о наличии информационной этики является необоснованным.

Статус информатики мы рассмотрели в специальной статье [18], здесь достаточно привести некоторые выводы. Информатика стала закономерным результатом развития цепочки формальных наук: лингвистика – логика – математика – информатика. Подобно своим предшественницам она является формальной наукой. Часто информатику считают технической наукой постольку, поскольку ее невозможно представить без вычислительной техники, прежде всего, компьютеров. При этом не учитывается донорский характер вычислительной техники. В науке широко практикуются интердисциплинарные, или, как мы предпочитаем выражаться, интерлигатеоретические отношения. Соотносительность двух лигатеорий всегда выражается в том, что одна из них является акцепторной (главной), а вторая донорской (вспомогательной). Какая из двух лигатеорий признается акцепторной и какая донорской, зависит от постановки исследовательской задачи.

Рассмотрим для примера соотносительность, с одной стороны, экономики, а с другой стороны, вычислительной техники. Для экономиста теория вычислительной техники является донорской концепцией. Для представителя вычислительной техники, вынужденного учитывать при закупке аппаратуры финансовые реалии, донорской теорией является экономика. Нечто аналогичное имеет место в соотношении информатики и вычислительной техники. Для информатики вычислительная техника является донорской. В качестве такой она не определяет статус информатики, в составе которой решающее значение имеет реализация следующего концептуального перехода: образцы алгоритмических вычислений – парадигмы программирования – программы.

Вывод, к которому мы пришли, состоит в том, что информатика подобно любой другой формальной науке не обладает присущим ей этическим началом. Образцы алгоритмических вычислений и парадигмы программирования безразличны к принципу максимизации благоденствия стейкхолдеров столь же определенно, как правила грамматики или аксиомы логики и математики. В силу изложенного понятие информационной этики несостоятельно.

Что же касается теорий вычислительной техники, то они, используя критерии эффективности и результативности, имеют аксиологический характер. В силу этого принцип максимизации благоденствия не чужд этим теориям, а потому вполне правомерно ввести представление об этике вычислительной техники, которая является достоянием не вычислительных машин, а людей, оперирующих ими.

Во Введении мы рассматривали термины "информационная этика" и "компьютерная этика". По нашему мнению, они оба являются неудачными. Сторонники информационной этики неверно истолковывают статус информатики [7, 8]. В их интерпретации информатика является универсальной наукой, более того – глобальной. Но таких наук просто нет. Нет никаких оснований утверждать, что из двух дюжин отраслей науки именно информатика обладает первостепенным значением. Каждая отрасль науки хороша на своем месте. Ни одна из них не способна непосредственно представлять достижения ее двух десятков сестриц.

Сторонники компьютерной этики [4–6] были бы правы, если бы они понимали под компьютерной этикой соответствующее содержание теорий вычислительной техники. Но они сопрягают понятие компьютерной этики не с вычислительной техникой, а с информатикой. Но как уже отмечалось такой ход мысли в метанаучном отношении несостоятелен.

ИНФОРМАТИКА И ЭТИКА

Может показаться, что мы явно принизили значение информатики для этики. Но это впечатление не соответствует действительности. Чтобы развеять его, обратимся к статусу человечества. Нет такого принципа, который бы выражал статус человечества в большей степени, чем принцип максимизации благоденствия всех людей, имеющих отношение к той или иной конкретной ситуации. Люди призваны не угнетать, а поддерживать друг друга. Как они могут добиться в этом деле максимальной эффективности? Во-первых, всемерно развивая научные институты. Во-вторых, развивая этическое начало применительно к каждой аксиологической теории. В-третьих, используя достижения буквально всех наук, не только аксиологических, но и формальных, и естественных. В-четвертых, объединяя достижения всех наук.

Пикантная особенность рассматриваемой ситуации состоит в том, что любая аксиологическая наука сама по себе недостаточна. Ее представители сильны

в этой науке, но не за ее пределами, где они, тем не менее, вынуждены совершать определенные поступки. Этика же предполагает интеграцию всех этических достижений в единое целое. В связи с этим исключительно актуальны и информатика, и формальная философская этика. Сама по себе, т.е. взятая в изоляции от других наук, информатика не обладает этическим содержанием. Но в рамках целого, называемого современной наукой, она, безусловно, имеет исключительно актуальную этическую относительность. Важно правильно понимать ее содержание, не абсолютизируя значимость информатики для современных людей. Приписывание этического начала непосредственно информатике приводит к искажению действительного положения дел и, как следствие, к недопониманию существа многих проблем.

ОТ МЕТАФИЗИЧЕСКОЙ ЭТИКИ К НАУЧНОЙ

Правильный путь развития теории этической относительности информатики был определен выше. Но абсолютное большинство исследователей идет другой дорогой. Они исходят из метафизической этики. Из нее, причем в довольно произвольной манере, извлекаются некоторые концепты, которые затем приспосабливаются к специфике информатики. Такой путь развития этической относительности информатики далеко не бесполезен. Проиллюстрируем его на примере этического кода международной Ассоциации вычислительной техники (ACM). До сих пор члены Ассоциации руководствовались кодом 1992 г. Но в 2018 г. организация полагает принять этический код в новом варианте. Мы рассматриваем второй проект нового этического кода ACM [19].

В преамбуле проекта его авторы отмечают, что они придают первостепенное значение необходимости обеспечения общественного блага. И в связи с этим авторы, руководствуясь принципом ответственности, предлагают принципы, во-первых, общего характера, во-вторых, более профессиональные, в-третьих, для руководителей различных подразделений. Приводятся формулировки 22-х принципов. Акцент на принципе обеспечения общественного блага, видимо, понимаемого как принцип всеобщего блага, вполне правомерен. Заслуживает также одобрения и акцент на принципе ответственности. Но его содержание не раскрывается.

Принцип ответственности фигурирует в абсолютном большинстве этических кодов различных организаций и предприятий. Обычно ответственность понимается как подотчетность. Субъект *A* отчитывается перед субъектом *B*, являющимся представителем инстанции ответственности, например, руководства предприятия. Сведение ответственности к подотчетности может привести к негативным последствиям в случае, если представитель инстанции ответственности придерживается устаревших воззрений. С учетом этого замечания мы полагаем, что принцип ответственности заключается в обязательстве всегда и везде стремиться к реализации самых развитых теорий, в частности, информационных.

Обратимся теперь к общим принципам. Авторы рассматриваемого проекта этического кода относят к ним следующие императивы: не навреди, будь честным и заслуживающим доверия, справедливым, поддерживай творческую работу, не разглашай тайны частной жизни и конфиденциальную информацию, не допускай дискриминацию кого-либо. Все перечисленные ценностные установки являются эпиконцептами (от др.-греч., ἐπι – над, расположение поверх чего-либо). Особенность эпиконцептов состоит в том, что для выяснения их подлинного значения необходимо обратиться к той основе, над которой они возвышаются в качестве надстройки. В рассматриваемом случае следует обратиться к метанауке. Не навредить способен лишь тот, кто не руководствуется устаревшими теориями. Честным является субъект, который не искажает теорию. Доверия заслуживает тот, кто многократно успешно реализовывал потенциал самых развитых теорий. Справедлив тот человек, который в соответствии с определенными теориями действует в интересах всех, а не избранных лиц. Поддерживать творческие процессы способен лишь тот, кто обладает способностью совершенствовать теории.

Все приведенные разъяснения основываются на метанаучном понимании существа научных теорий, в том числе информационных. Без использования концепта теории они не могли бы состояться. Что же касается рассматриваемого проекта, то в нем вообще отсутствует понятие теории. В результате подлинное содержание эпиконцептов, которым присвоен высокий ранг принципов, остается невыясненным. Тут же отметим, что содержание этических эпиконцептов не раскрывается в метафизической философской этике. В этом и состоит ее изъян. В ее границах эпиконцепты остаются закованными в свои метафизические оболочки.

Обратимся теперь к профессиональным принципам. Авторы рассматриваемого проекта относят к ним следующие императивы: обеспечивай высокое качество своей работы, поддерживай высокие стандарты профессиональной компетентности, делай соответствующие обзоры, соблюдай законы профессиональной деятельности и т.д. На этот раз авторы используют набор не эпиконцептов, а понятий обобщенной теории информатики. Есть много информационных теорий, которые обладают схожими чертами, которые находят свое выражение в обобщенной теории информатики. Концепты этой теории как раз и используют авторы рассматриваемого проекта, что является вполне корректным, ибо известны их основания в специфических информационных теориях.

Проведенный анализ проекта этического кода ACM позволяет сделать определенный вывод. Многочисленные этические коды различных организаций и предприятий строятся по одной и той же схеме. От имени метафизической философской этики предлагается список избранных эпиконцептов, которым присваивается почетный статус принципов. По сути, их подлинные основания в составе не философских, а специфических научных теорий не выясняются. Тем не менее, делаются попытки извлечь эпиконцепты из их интуитивных оболочек. Эпиконцепты дополняют-

ся желаемыми профессиональными обязанностями, которым также присваивается статус принципов. Обилие принципов при отсутствии законов свидетельствует о том, что составленные по описанной схеме этические кодексы не являются полновесными научными теориями, в которых из принципов должны выводиться законы.

Альтернативный путь построения этического кодекса, обоснованный в настоящей статье, состоит в том, что за основу берется не метафизическая философская этика, а вполне конкретные аксиологические теории, например, технические или экономические, дополненные принципом максимизации благоденствия всех стейкхолдеров. В таком случае нет необходимости в метафизической этике с ее неясными концептами. Все принципы оказываются подлинными. Например, применительно к вычислительной технике принцип максимизации благоденствия дополняется принципами эффективности, результативности, быстродействия, безопасности и надежности.

Во избежание недоразумений отметим также, что этические коды соответствуют специфике только тех организаций и предприятий, которые в своей деятельности руководствуются аксиологическими теориями. В противном случае, строго говоря, следует вместо этического кодекса руководствоваться либо кодексом этической относительности, либо кодексом этикета. С этой точки зрения, например, физикам, химикам, математикам и программистам следует руководствоваться кодексами этической относительности.

ЗАКЛЮЧЕНИЕ

Для того чтобы определить этическую значимость информатики, нам пришлось провести многозвенный анализ, рассматривая соотношение науки, метанауки, информатики и этики. Современная наука представляет собой сложное образование, состоящее из многочисленных формальных, естественных и аксиологических теорий. Их статус стал изучаться с необходимой степенью тщательности лишь начиная с появления позитивизма О. Конта и Дж.С. Милля, т.е. с середины XIX в. При этом с позитивистски настроенными исследователями случился казус, сказавшийся на судьбе этики крайне негативно. Этике было отказано в научном статусе. На многие десятилетия она оказалась лишенной значимых контактов с научными теориями. В этих условиях ей не суждено было покинуть пределы метафизики. Однако в силу совершенствования аксиологических наук и роста многочисленных проблемных аспектов жизнедеятельности современного человечества остро встал вопрос об извлечении этики из ее метафизической ссылки и придания ей подлинно научного характера. Но как добиться желаемого? Абсолютное большинство исследователей решило, что можно, не затрудняя себя специальными метанаучными изысканиями, придать метафизической этике прикладное значение. Из ее состава извлекается некий набор концептов, который считается актуальным для той или иной науки. Но опыт проведенных исследований показывает, что такое незамысловатое приобщение этики к сонму научных теорий не приводит к существенному прогрессу. Этика остается метафизической. В определении статуса любой этики нет альтернативы метанауке. В связи с этим мы показали, как именно может быть реализован метанаучный подход при определении этической значимости различных научных теорий. Эта значимость оказывается принципиально различной природы применительно, с одной стороны, к аксиологическим теориям и, с другой стороны, к естественным и формальным теориям.

Все аксиологические теории чрезвычайно восприимчивы к принципу максимизации благоденствия всех стейкхолдеров. Они не только принимают этот этический принцип в свой состав, но и подчиняют ему все остальные принципы. В результате этическое начало получает свое четкое научное оформление. Оно является концептуальной вершиной всех аксиологических теорий.

Существенно по-другому обстоит дело с естественными и формальными теориями. Принципы этих теорий, например, аксиомы математики и парадигмы программирования, никак не реагируют на принцип максимизации благоденствия и не образуют с ним никакого органического единства. Но это не означает их полнейшей этической индифферентности. Дело в том, что этическое содержание жизнедеятельности человечества реализуется не отдельными теориями, а лишь их полным единством. В связи с этим трудно переоценить значение как естественных, так и формальных наук, без которых, как известно не может состояться ни одна аксиологическая наука. Ярко выраженные интертеоретические отношения формальных теорий, в том числе информатики, указывают на их этическую относительность. Сторонники как компьютерной, так и информационной этики отождествляют информатику с аксиологическими теориями. Это ошибочный путь, затрудняющий понимание ее этической относительности.

Выяснение метанаучных оснований этический относительности информатики открывает путь для ее всестороннего изучения с позиций науки, а не плохо понятой метафизики.

СПИСОК ЛИТЕРАТУРЫ

1. Gödel K. [1934] On Undecidable Propositions of Formal Mathematical Systems // Davis M. (Ed.). The Undecidable. – New York: Raven, 1965 – P. 41–74.
2. Church A. An Unsolvable Problem of Elementary Number Theory // American Journal of Mathematics. – 1936. – Vol. 58, № 2. – P. 345–363.
3. Turing A.M. On Computable Numbers, with an Application to the Entscheidungsproblem // Proceedings of the London Mathematical Society. Ser. 2. – 1936-1937. – Vol. 42, № 1. – P. 230–265.
4. Moor J. H. What Is Computer Ethics // Metaphilosophy. – 1985 – Vol. 16, № 4. – P. 266–275.
5. Capurro R. Informationsethik – Eine Standortbestimmung // International Journal of Information Ethics. – 2004. – Vol. 1, № 6. – P. 4–10.
6. Introna L.D. Maintaining the Reversibility of Foldings: Making the Ethics (Politics) of Information Technology Visible // Ethics and Information Technology. – 2007. – Vol. 9, № 1. – P. 11–25.

7. Gorniak K. The Computer Revolution and the Problem of Global Ethics // Science and Engineering Ethics. – 1996. – Vol. 2, № 2. – P. 177–190.
8. Floridi L. Information Ethics: On the Philosophical Foundation of Computer Ethics // Ethics and Information Technology. – 1999. – Vol. 1, №1. – P. 37–56.
9. Aristotle's Nicomachean Ethics / A translation by R.C. Bartlett, S.D. Collins (with an interpretive essay, notes, and glossary). – Chicago: University of Chicago Press, 2011. – 368 p.
10. Bentham J. An Introduction to the Principles of Morals and Legislation. – Oxford: Clarendon Press, 1907. – 338 p.
11. Wittgenstein L. Tractatus Logico-Philosophicus. – New York: Dover Publication, Inc., 1999. – 122 p.
12. Ayer A. J. Language, Truth and Logic. – London: V. Gollancz, ltd., 1936. – 254 p.
13. Stevenson C.L. Ethics and Language. – New Haven: Yale University Press, 1944. – 338 p.
14. Hare R. M. The Language of Morals. – Oxford: Clarendon Press, 1952. – 202 p.
15. Scheler M. Der Formalismus in der Ethik und die materiale Wertethik. – Halle a. d. Saale: Verlag Niemeyer, 1916. – 620 s.
16. Habermas J. Moralbewußtsein und kommunikatives Handeln. – Frankfurt am Main: Suhrkamp, 1983. – 207 s.
17. Critchley S. The Ethics of Deconstruction: Derrida and Levinas, 3rd ed. – Edinburgh: Edinburgh University Press, 2014. – 352 p.
18. Канке В.А. Метанаучные и философские основания определения статуса информатики // Научно-техническая информация. Сер.2. – 2017. – № 6. – С. 1-7; Kanke V.A. Metascientific and Philosophical Reasons to Define the Status of Computer Science // Automatic Documentation and Mathematical Linguistics. – 2017. – Vol. 51, № 3. – P. 101–107.
19. 2018 ACM Code of Ethics and Professional Conduct: Draft 2. – URL: https://ethics.acm.org/2018-code-draft-2/ (дата обращения 01.11.2017).

Материал поступил в редакцию 23.10.17.

Сведения об авторе

КАНКЕ Виктор Андреевич – доктор философских наук, профессор кафедры философии Национального исследовательского ядерного университета МИФИ, Москва
e-mail: kanke@obninsk.ru

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК [004.774 : 001.32(470)] – 047.44

А.А. Печников

Ассортативное смешивание в российском академическом Вебе

Приводятся результаты исследования фрагмента Веба, представляющего собой взаимосвязанные веб-сайты 546 научных организаций России, на наличие (отсутствие) дискретного ассортативного смешивания по видам научной деятельности. Показано, что в общем случае ассортативность практически отсутствует. Однако, в случае удаления примерно 40 сайтов, относящихся к научно-организационной и библиотечной деятельности, ассортативность проявляется достаточно четко. Это означает, что институты, занимающиеся непосредственно научной деятельностью, имеют очевидную тенденцию выставлять на своих сайтах гиперссылки на сайты своих коллег. При этом гиперссылки, связывающие их с научно-организационными учреждениями, имеют гораздо большее значение для структуры веб-пространства, фактически сводя к нулю ассортативность.

Ключевые слова: *веб-пространство, сайт, гиперссылка, веб-граф, ассортативное смешивание*

ВВЕДЕНИЕ

Понятие «ассортативное смешивание» (*assortative mixing*) возникло в социологии при изучении закономерностей формирования супружеских пар [1]. Исследования, проводимые в г. Сан-Франциско для 1958 супружеских пар различных рас, показали четкую тенденцию в образовании пар из представителей одной расы.

Применение этого подхода для исследования сложных сетей (социальных, биологических, технологических) позволяет обнаруживать новые структурные свойства сетей, имеющие большое значение. Наличие ассортативного смешивания (или просто «ассортативности») в сети говорит о стремлении ее к разбиению на отдельные сообщества по некоторым признакам. Например, в случае «Живого журнала» [2] для пользователей, имеющих от нескольких десятков до тысячи связей, смешивание является ассортативным, т.е. общительные блогеры предпочитают дружить с общительными [3].

В работе [4] определены два вида ассортативного смешивания – дискретное и скалярное. Примером использования дискретного ассортативного смешивания является характеризация исследуемого множества объектов по расовому признаку, а скалярного – характеризация по количеству связей (как в «Живом журнале»).

В настоящей статье приводятся результаты исследования фрагмента Веба, представляющего собой множество веб-сайтов 546 организаций, подведомственных Федеральному агентству научных организаций (ФАНО) [5], на наличие (отсутствие) дискретного ассортативного смешивания по такому признаку, как вид научной деятельности.

ОСНОВНЫЕ ИСХОДНЫЕ ПОЗИЦИИ

Данные о подведомственных организациях были взяты с официального сайта ФАНО [5] в декабре 2016 г. и насчитывали 675 организаций. Все официальные сайты организаций, указанные ФАНО, были проверены на работоспособность, а обнаруженные ошибки и неточности исправлялись в основном с использованием собственных баз данных [6], а также данных из Википедии [7]. Некоторые сайты на конец декабря 2016 г. оказались неработающими (по различным причинам) и были исключены из исследования.

В целевое множество были добавлены 4 сайта организаций, не находящихся в ведении ФАНО, – это сайт РАН и сайты Уральского, Сибирского и Дальневосточного региональных отделений РАН. Полученное базовое расширенное целевое множество на начало сканирования содержало 597 сайтов.

Сканирование сайтов проводилось в период с декабря 2016 г. по май 2017 г. с использованием краулера BeeBot [8]. Данные в виде внешних гиперссылок, сделанных с сайтов целевого множества, были занесены для дальнейшей обработки в базу данных [6]. После соответствующей обработки, представляющей собой отбор гиперссылок, связывающих только сайты целевого множества, был построен фраг-

мент Веба, который (по традиции) мы будем в этой статье называть академическим. Он содержит 597 сайтов (593 сайта, подведомственных ФАНО, сайт РАН и три сайта его региональных отделений), связанных между собой более чем тремя тысячами гиперссылок.

Коэффициент ассортативности (определение)

В качестве математического инструментария для проверки на наличие (отсутствие) дискретного ассортативного смешивания применим подход, предложенный в работе [4]. Очевидно, что фрагменту Веба однозначно соответствует веб-граф (ориентированный граф с кратными дугами без петель), вершинами которого являются сайты, а дугами – гиперссылки.

Пусть $L=(l_{ij})$ матрица смежности такого веб-графа, где l_{ij} – количество дуг, у которых начальная вершина i, а конечная j; $i,j=\overline{1,n}$, n – мощность целевого множества.

Пусть целевое множество N разбито на m непересекающихся подмножеств (групп) $N_1,...,N_m$, объединение которых $N_1\cup...\cup N_m=N$.

По матрице L построим матрицу $E=(e_{ij})$, в которой $e_{ij}=\dfrac{\sum\limits_{s\in N_i,t\in N_j} l_{st}}{\sum\limits_{s=1}^{n}\sum\limits_{t=1}^{n} l_{st}}$, т. е. e_{ij} – это доля от общего количества дуг, а именно – отношение количества дуг, соединяющих вершины подмножества N_i с вершинами подмножества N_j к общему количеству дуг веб-графа.

Обозначим суммы элементов по строкам и по столбцам $a_i=\sum\limits_{j=1}^{m} e_{ij}$ и $b_i=\sum\limits_{i=1}^{m} e_{ij}$; понятно, что $\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m} e_{ij}=1$.

Для количественной оценки уровня ассортативного смешивания определим коэффициент ассортативности r в соответствии с [5]:

$$r=\frac{\sum\limits_{i=1}^{m} e_{ii}-\sum\limits_{i=1}^{m} a_i\cdot b_i}{1-\sum\limits_{i=1}^{m} a_i\cdot b_i} \qquad (1)$$

Продемонстрируем изложенное на примере, представляющем собой значительную часть академического фрагмента Веба, содержащего 260 институтов, из которых 135 относятся к отделению сельскохозяйственных наук, 58 – к отделению биологических наук и 67 – к отделению наук о Земле [8]. Связи между сайтами показаны на рис. 1.

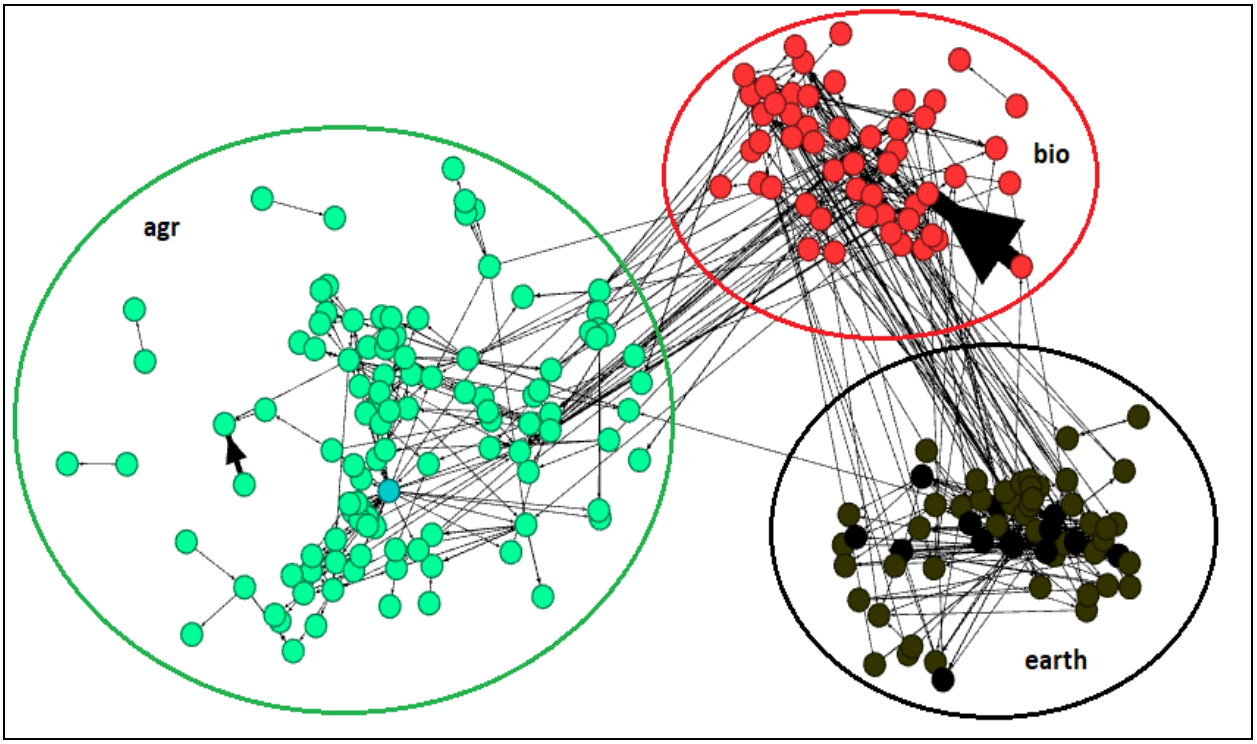


Рис. 1. Часть академического фрагмента Веба

Овалами выделены группы сайтов, относящиеся к научным отделениям, что обозначено соответствующими пометками (*agr*, *bio* и *earth*). Рисунок достаточно очевидно демонстрирует большое количество гиперссылок между сайтами «внутри» овалов, и малое – между сайтами разных научных отделений. Матрица смежности *L* велика по размерам, и мы ее приводить не будем, а приведем матрицу $\overline{L}$, показывающую связи между группами сайтов внутри каждого научного отделения и между сайтами отделений

$$\overline{L} = \begin{array}{c|ccc} & agr & bio & earth \\ agr & 197 & 18 & 0 \\ bio & 14 & 129 & 20 \\ earth & 1 & 28 & 168 \end{array}.$$

Например, здесь $\bar{l}_{agr,bio} = 18$ означает, что с 135 институтов отделения сельскохозяйственных наук сделано 18 гиперссылок на институты отделения биологических наук, а на институты отделения наук о Земле подобных ссылок нет и $\bar{l}_{agr,earth} = 0$.

Матрица *E* в этом случае имеет следующие значения элементов:

$$E = \begin{array}{c|ccc} & agr & bio & earth \\ agr & 0{,}343 & 0{,}031 & 0{,}000 \\ bio & 0{,}024 & 0{,}224 & 0{,}035 \\ earth & 0{,}002 & 0{,}049 & 0{,}292 \end{array}.$$

Далее, по формуле (1) получаем $r = 0{,}788$.

Коэффициент ассортативности (свойства)

При исследовании свойства ассортативности сети достаточно естественным представляется условие, когда каждый ее участник имеет хотя бы одну связь с другим участником, т. е. изолированные участники не рассматриваются. При таком условии из (1) понятно, что наибольшее значение $r = 1$ достигается в том случае, когда в матрице $E = (e_{ij})$ элементы по диагонали e_{ii} больше нуля, а все остальные равны 0. В этом случае мы говорим о совершенном ассортативном смешивании.

Если же элементы e_{ii} равны нулю, то значение

$$r_{\min} = -\frac{\sum_{i=1}^{m} a_i \cdot b_i}{1 - \sum_{i=1}^{m} a_i \cdot b_i}$$

будет отрицательным в диапазоне $-1 \le r < 0$ и сеть является неассортативной (дисассортативной).

Заметим, что это свидетельствует об очень высокой степени ассортативности рассмотренной части фрагмента академического Веба.

АССОРТАТИВНОСТЬ ФРАГМЕНТА АКАДЕМИЧЕСКОГО ВЕБА ПО ВИДАМ НАУЧНОЙ ДЕЯТЕЛЬНОСТИ

Как уже было сказано, в приведенном примере, в качестве характеристики-признака ассортативного смешивания была выбрана сфера научной деятельности. Большинство научных организаций классифицируются по этому признаку благодаря спискам организаций, связанных с соответствующими научными отделениями РАН [8]. Для таких организаций, как собственно РАН, Уральское, Сибирское и Дальневосточное региональные отделения РАН (не входящие в ведение ФАНО), а также региональные научные центры РАН (входящие в ведение ФАНО, например, Карельский научный центр), был введен признак "научно-организационный вид деятельности". В отдельную группу были выделены научные библиотеки (библиотечная деятельность). Кроме того, несколько организаций были исключены в связи с сомнениями в правильности классификации. В результате предпринятых действий фрагмент академического Веба, исследуемый далее, содержит 546 сайтов, связанных 3103 гиперссылками.

Классификатор исследуемых организаций с указанием количества организаций, входящих в каждую группу, приведен в табл. 1.

Полученную матрицу $\overline{L}$, показывающую связи между группами сайтов, приведем в табл. 2.

Граф, построенный по матрице $\overline{L}$, и показывающий связи между группами по видам деятельности, приведен на рис. 2.

При изображении связей сохранены пропорции по количеству дуг, связывающих группы. Петли в графе показывают большое количество гиперссылок "внутри" каждой группы. При этом также очевидно большое количество связей между группой *adm* и всеми остальными группами.

Опустим детали вычислений коэффициента ассортативности *r* в соответствии с формулой (1); в данном случае $r = 0{,}226$, что говорит о незначительной ассортативности построенного фрагмента академического Веба.

Теперь удалим из фрагмента академического Веба все сайты, входящие в группу *adm*. Получаем редуцированный фрагмент Веба, содержащий 516 сайтов и 1885 связывающих их гиперссылок. Заметим, что удаление 30 сайтов привело к ликвидации 1218 гиперссылок, из чего следует очень большая значимость так называемых "административных" связей на связность фрагмента Веба [10]. Полученный граф связей показан на рис. 3.

В этом случае мы имеем значение $r = 0{,}454$. Удаление еще одной группы сайтов организаций, занимающихся информационно-аналитическим сопровождением научной деятельности, а именно, группы сайтов библиотек, дает значение $r = 0{,}537$.

Таблица 1

Классификатор видов научной (научно-организационной) деятельности

№	Вид деятельности/научное отделение	Признак	Количество организаций
1	Научно-организационная	*adm*	30
2	Сельскохозяйственные науки	*agr*	135
3	Библиотечная деятельность	*bibl*	7
4	Биологические науки	*bio*	58
5	Химия и науки о материалах	*chem*	39
6	Науки о Земле	*earth*	67
7	Энергетика, машиностроение, механика и процессы управления	*energ*	33
8	Глобальные проблемы и международные отношения	*intern*	6
9	Историко-филологические науки	*ist-fil*	33
10	Математические науки	*math*	13
11	Медицинские науки	*med*	33
12	Нанотехнологии и информационные технологии	*nano*	20
13	Физические науки	*phys*	38
14	Физиологические науки	*physiol*	9
15	Общественные науки	*soc*	25
	ВСЕГО		546

Таблица 2

$\bar{L} =$

	adm	*agr*	*bibl*	*bio*	*chem*	*earth*	*energ*	*intern*	*ist – phil*	*math*	*med*	*nano*	*phys*	*physiol*	*soc*
adm	98	9	3	99	68	114	54	4	55	19	21	26	75	13	37
agr	40	197	29	18	2	0	0	0	4	0	2	2	1	0	2
bibl	17	7	20	17	9	23	7	0	13	4	0	4	17	3	4
bio	64	14	25	129	13	20	2	1	10	3	9	6	13	8	10
chem	53	0	15	6	93	8	7	0	1	1	0	3	22	0	2
earth	106	1	30	28	12	168	12	1	5	6	0	4	23	2	3
energ	45	1	17	3	12	9	36	2	2	10	0	5	18	0	6
intern	6	1	2	0	0	2	2	8	5	0	0	0	0	0	6
ist – phil	40	1	13	12	7	12	0	4	137	3	0	3	6	3	10
math	24	0	5	4	5	8	3	0	5	24	0	6	10	0	4
med	25	0	5	5	0	0	0	0	0	0	7	0	0	0	0
nano	20	0	2	2	3	3	5	0	2	8	2	4	7	1	0
phys	56	0	18	11	19	14	10	0	9	9	1	6	76	1	6
physiol	4	0	1	2	0	2	0	0	1	0	1	2	2	2	1
soc	23	0	0	3	1	8	5	4	11	4	0	1	2	0	59

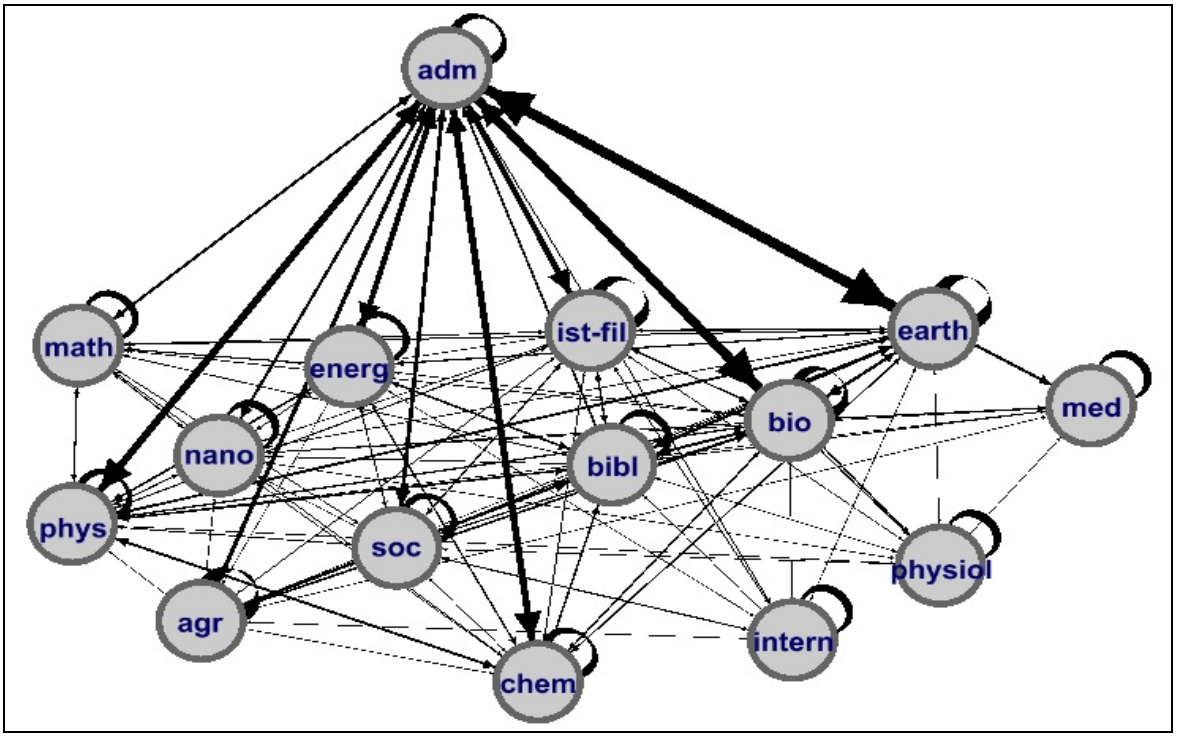


Рис. 2. Граф связей между группами сайтов академического фрагмента Веба.

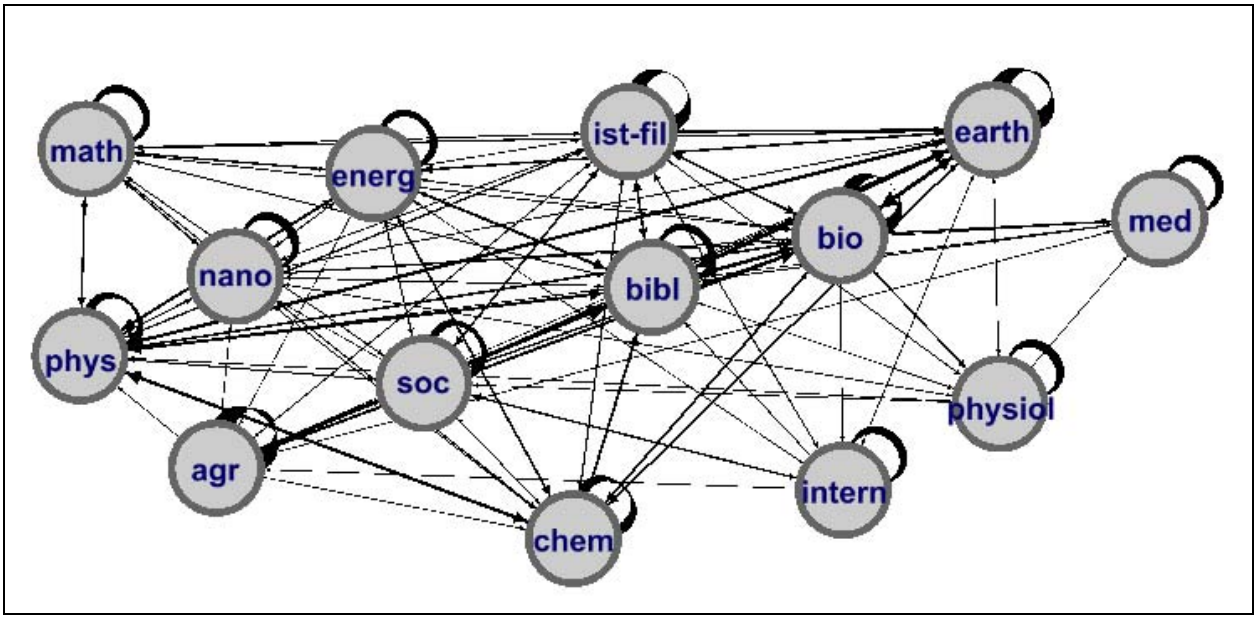


Рис. 3. Граф связей между группами сайтов академического фрагмента Веба после удаления группы *adm*.

СОДЕРЖАТЕЛЬНАЯ ИНТЕРПРЕТАЦИЯ И ВЫВОДЫ

Итак, исследуемый фрагмент академического Веба, содержащий 546 сайтов, классифицированных по 15 видам научной деятельности (включая научно-организационную и библиотечную), имеет невысокий коэффициент ассортативности ($r = 0{,}226$). Казалось бы, это говорит о том, что в данном фрагменте Веба отсутствует ярко выраженная тенденция создания гиперссылок между веб-сайтами организаций, относящихся к одному и тому же виду деятельности.

Однако этот вывод был бы преждевременным. Удаление 30 веб-сайтов, относящихся к научно-организационной деятельности, и 7 сайтов, относящихся к библиотечной деятельности, приводит к тому, что для редуцированного фрагмента Веба значение коэффициента ассортативности увеличивается более чем вдвое ($r = 0{,}537$). Институты, занимающиеся непосредственно научной деятельностью,

имеют очевидную тенденцию выставлять на своих сайтах гиперссылки на сайты своих коллег по виду деятельности.

Но при этом гиперссылки, связывающие их с научно-организационными учреждениями, имеют гораздо большее значение для структуры веб-пространства, фактически сводя ассортативность практически к нулю. Этот результат подтверждает сделанный ранее вывод о том, что административный каркас играет системообразующую роль в организации академического Веба [10], где под административным каркасом фрагмента Веба понимаются гиперссылки между научными организациями, соответствующие их иерархической подчиненности.

Поэтому, говоря о значимости видов научной деятельности как факторов, влияющих на формирование структуры академического веб-пространства посредством гиперссылок, можно перечислить их в порядке убывания приоритетов:

- научно-организационная деятельность и библиотечно-информационное обслуживание;
- научная деятельность одного вида,
- междисциплинарная научная деятельность.

СПИСОК ЛИТЕРАТУРЫ

1. Catania J.A., Coates T.J., Kegels S., Fullilove M.T. The population-based AMEN (AIDS in Multi-Ethnic Neighborhoods) study // American Journal of Public Health. – 1992. – № 82. – P. 284–287.
2. Главное – Живой Журнал. – URL: http://www.livejournal.com (дата обращения: 11.08.2017).
3. Митин Н.А., Подлазов А.В., Щетинина Д.П. Исследование сетевых свойств Живого журнала // Препринты ИПМ им. М.В.Келдыша. – 2012. – №78. – 16 с. – URL: http://library.keldysh.ru/preprint.asp?id=2012-78.
4. Newman M.E.J. Mixing patterns in networks // Physical Review. E 67. 2003. 026126.
5. Подведомственные организации | ФАНО России. – URL: http://fano.gov.ru/ru/about/sub_organizations (дата обращения: 11.08.2017).
6. Головин А.С., Печников А.А. База данных внешних гиперссылок для исследования фрагментов Веба // Информационная среда вуза XXI века: материалы VII Всероссийской научно-практической конференции (23-27 сентября 2013 г.). – Петрозаводск, 2013. – С. 55-57.
7. Институты РАН – Википедия. – URL: https://ru.wikipedia.org/wiki/Институты_РАН (дата обращения: 11.08.2017).
8. Печников А.А., Чернобровкин Д.И. Адаптивный краулер для поиска и сбора внешних гиперссылок // Управление большими системами. Вып. 36. – М.: ИПУ РАН, 2012. – С.301-315.
9. Список отделений РАН по направлениям науки. – URL: http://www.ras.ru/sciencestructure/departments.aspx (дата обращения: 16.08.2017).
10. Печников А.А. О связности веб-сайтов Российской академии наук на административном каркасе // Труды междунар. научно-практической конференции «Теория активных систем» (14-16 ноября 2011 г., Москва). Том III. – М., 2011. – С. 95-98.

Материал поступил в редакцию 02.10.17.

Сведения об авторе

ПЕЧНИКОВ Андрей Анатольевич – доктор технических наук, доцент, главный научный сотрудник, руководитель лаборатории Телекоммуникационных систем Института прикладных математических исследований Карельского научного центра Российской академии наук, г. Петрозаводск
e-mail: pechnikov@krc.karelia.ru

УДК [004.774 : 002.6] – 047.44

Ю.М. Брумштейн, Е.Ю. Васьковский

Исследование функциональности и вебометрических показателей специализированных сайтов, связанных с научной деятельностью

Рассмотрены назначение, функциональность и вебометрические показатели (ВМП) сайтов основных международных организаций-агрегаторов специализированной научной информации, включая системы учета цитирований и оценки наукометрических показателей; сайтов, предназначенных для поддержки публикаций на национальных языках и/или посвященных национальной тематике. Выполнено сравнение ВМП зарубежных сайтов и функционально аналогичных им российских интернет-ресурсов.

Представлена структура зарубежных и российских сайтов, на которых размещена информация о диссертациях и авторефератах, а также их полнотекстовые версии. Охарактеризованы условия доступа пользователей к таким материалам. Сравнены величины ВМП для этих сайтов.

Проанализирована функциональность и ВМП интернет-ресурсов следующих категорий: основных зарубежных и российских сайтов, с которых обеспечивается доступ к автономным информационно-аналитическим и информационным системам по научным журналам; сайтов отдельных журналов, связанных с наукометрическими исследованиями, информационным менеджментом (сервисом) в сфере науки.

Ключевые слова*: научная информация, организации-агрегаторы, интернет-сайты, тематическая специализация, диссертации, авторефераты, функциональная специализация, вебометрические показатели, информационно-телекоммуникационные технологии, web-аналитика, поведение Интернет-пользователей*

ВВЕДЕНИЕ

В предыдущей работе авторов [1] было выполнено категорирование сайтов – агрегаторов научной информации (НИ), представлена методика и проанализированы совокупности вебометрических показателей (ВМП) основных сайтов политематического характера – зарубежных и российских. Здесь рассматриваются научные интернет-ресурсы, специализированные по тематике, по странам публикаций, по функциональному назначению и пр. Для ограничения объема статьи в ней рассмотрены только сайты категорий «К4…К11» – по классификации, введенной в [1]. Наряду с политематическими научными ресурсами (в России это, прежде всего, elibrary.ru [2–4], «КиберЛенинка», сайт ВИНИТИ [5], сайт ГПНТБ) специализированные сайты занимают важное место в научно-информационном пространстве как международном, так и отдельных стран.

Специализированные научные сайты используются, в частности, для решения следующих задач. (1) Агрегация НИ определенной тематики [6] или определенного места происхождения/создания. (2) Обеспечение свободного распространения работ, находящихся в открытом доступе [7–9]. (3) Поддержка доступности через Интернет работ, распространяемых на платных условиях. (4) Обеспечение удобства доступа к информации не текстового характера [10]. (5) Создание особых условий доступности публикаций для некоторых групп академических пользователей [11]. Пункты 1–5 могут рассматриваться как направленные на решение задач *управления знаниями* [12], в том числе и для лиц, профессионально занимающихся научной деятельностью. (6) Исключение повторного проведения ранее уже выполненных исследований. (7) Объективная разовая оценка (или мониторинг) роли сайтов отдельных организаций [13, 14], включая НИИ и вузы; сравнение вебометрических показателей этих сайтов. (8) Оценка публикационной активности [15] и наукометрических показателей отдельных исследователей и организаций. (9) Формирование научного имиджа [16] организаций в информационном пространстве. (10) Изучение тенденций развития научных исследований и разработок в определенных предметных областях [17]. Отметим, что несвоевременное или неточное выявление НИ, а также существующих тенденций развития исследований может рассматриваться и как фактор информационной безопасности в отношении управления научной деятельностью. (11) Информационно-аналитическая поддержка принятия решений по управлению «тематическими направлениями научной деятельности» [18], распределению между ними финансовых, трудовых, иных ресурсов и пр. При этом могут ставиться/решаться нелинейные задачи опти-

мизации при четких и нечетких условиях (критериях оптимальности и ограничениях). В нечетких условиях получаемые оптимальные решения также будут носить нечеткий характер, нуждаться в анализе их устойчивости по входным данным.

Необходимо учитывать, что по мере научно-технического прогресса и, в частности развития информационно-телекоммуникационных технологий, увеличиваются объемы научной и инженерно-технической информации; меняется структура этой информации за счет появления новых направлений исследований (областей деятельности, разработок), изменения относительной важности существующих направлений [19, 20]. Результаты исследований по этим направлениям в форме научных публикаций отражаются не только на политематических, но и на специализированных интернет-сайтах – научного, инженерно-технического и производственного характера, в том числе ориентированных на массовых пользователей. В конечном счете, появление и развитие новых направлений исследований и разработок приводит и к изменениям соотношений видов научной и инженерно-технической информации, которую потребляют пользователи, а в более общем плане – к изменению информационной инфраструктуры общества [19]. Однако сопоставительный анализ функциональности и ВМП специализированных научных сайтов в существующих публикациях практически отсутствует. Такой анализ может рассматриваться не только как часть науковедческих исследований, но и как метод информетрического [21] анализа. В свою очередь, полнота и объективность результатов изучения ВМП специализированных научных сайтов должны опираться на адекватные методы анализа [1, 22]. Кроме того, и сами эти методы должны совершенствоваться – с учетом изменения с течением времени используемого программного обеспечения (в том числе на серверах); подходов к оптимизации структуры используемых сайтов и их информационному наполнению, применению на них гиперссылок (внутренних и исходящих) [20], изменению моделей поведения пользователей при поиске информации [23], включая и мобильных пользователей.

Цель настоящей статьи – комплексное рассмотрение проблематики ВМП специализированных сайтов научного характера с точки зрения оценки их известности в информационном пространстве, доступности, востребованности информации, размещенной на этих сайтах. Как и ранее [1], мы делаем акцент на сайты, содержащие исключительно (или преимущественно) научную информацию, а не инженерно-техническую или производственную. Подчеркнем, что часть рассматриваемых сайтов может быть отнесена одновременно к разным категориям.

МАТЕРИАЛ, МЕТОДИКА ИССЛЕДОВАНИЯ, ОСОБЕННОСТИ ПРЕДСТАВЛЕНИЯ РЕЗУЛЬТАТОВ

Аналогично [1] объектами анализа были сайты Интернета, прежде всего, их вебометрические показатели. На этих сайтах размещена информация «открытого доступа», т.е. не содержатся сведения об исследованиях и разработках «закрытого» характера.

Сведения о функциональном назначении этих сайтов, о количестве источников, с которых агрегируется информация на этих сайтах, брались с их стартовых страниц, частично из статей по этим сайтам в Википедии, иногда и из других источников. Отметим сразу же, что количественные показатели из разных источников в отношении агрегируемых на сайтах научных изданий, статей и других материалов иногда серьезно различались.

Информация по рассматриваемым нами сайтам рассеяна по многим интернет-ресурсам. Коллекции гиперссылок (обычно лишь на наиболее известные сайты-агрегаторы научной информации) есть на сайтах многих российских НИИ и вузов, в том числе ориентированных на информационную поддержку деятельности молодых ученых. Однако лишь в ряде случаев содержание материалов тех сайтов, на которые указывают гиперссылки, комментируется. В основном российские исследователи этими ссылками и пользуются.

Перечни тематически специализированных сайтов встречаются в ряде научных статей, доступных в Интернете (включая русскоязычные), в материалах Википедии на русском и английском языках (в том числе с некоторыми краткими комментариями). Согласно [24], крупнейшими каталогами (реестрами) репозиториев/архивов по материалам, содержащим НИ, являются: (1) Реестр открытых архивов: http://www.openarchives.org/Register/BrowseSites; (2) Реестр репозиториев открытого доступа: http://roar.eprints.org; (3) Каталог открытых репозиториев *OpenDOAR* / *Directory of Open Access Repositories* – http://www.opendoar.org; (4) Европейский реестр репозиториев открытого доступа: http://www.openarchives.eu/ home/home.aspx. Российским исследователям, впрочем, эти ресурсы малоизвестны.

Гиперссылки с одних тематически специализированных сайтов-агрегаторов НИ на функционально аналогичные им сайты встречаются редко.

При отборе сайтов для анализа ВМП нами учитывались следующие факторы: степень известности различных сайтов для российских пользователей; потенциальная перспективность использования некоторых малоизвестных сайтов научного характера; объемы информации на сайтах; тематическое содержание информации, представленной на сайтах (авторы старались сбалансировано охватить различные тематические направления исследований). Понятно, что при отсутствии четких количественных критериев отбора сайтов их выбор в определенной степени носил субъективный характер.

Причинами снижения фактической доступности информации, размещенной на рассматриваемых сайтах, могут быть: чрезмерно высокая насыщенность страниц сайтов графическими объектами – это снижает скорость работы с сайтами, прежде всего, для мобильных пользователей; отсутствие автоматической адаптации вида страниц сайтов к размерам и разрешению экранов, применяемых пользователями мобильных устройств; невысокое качество, а иногда и неполнота перевода страниц сайтов интернет-переводчиками; высокий уровень оплаты за доступ к информации (если такой доступ не предусмотрен по соглашениям организаций, в которых работают исследователи, с владельцами сайтов-агрегаторов).

Используемое нами категорирование сайтов соответствует введенному в [1]. Подчеркнем, что часть сайтов могут быть одновременно отнесены к двум и более категориям, при этом применяемую классификацию не следует рассматривать как единственно возможную.

Анализ ВМП специализированных научных сайтов-агрегаторов призван оценивать не только объемы накопленной информации, но и ее востребованность у пользователей [1, 22]. ВМП сайтов определялись, в основном, по методике, подробно описанной в [1]. Приводимые нами количественные значения ВМП относятся к периоду с 05.08.2017 по 20.08.2017. Используемые в таблицах настоящей статьи обозначения ВМП в основном носят стандартный характер. Нестандартные обозначения подробно расшифрованы в [1].

С учетом выявленных в [1] сложностей при оценке некоторых ВМП сайтов большого объема мы дополнительно использовали оценки количества страниц сайтов с помощью информационно-поисковой системы Google. Полученные таким образом значения представлены в приводимых таблицах в угловых скобках (т.е. « < > »), при необходимости в виде знаменателя в соответствующих клетках таблиц. Однако следует понимать, что разные поисковые системы Интернета определяют различное количество страниц на одних и тех же сайтах (см. табл. 1 в [1]), между объемами сайтов в мегабайтах (определяемых по предложенной авторами методике) и количеством страниц сайтов, определяемых ИПС Google, нет прямо пропорциональной зависимости.

Отметим, что для подсчета AVD в применяемых авторами программных средствах в качестве временны*х* меток используется время (дата) открытия каждой страницы, просматриваемой посетителем. Иными словами AVD – это временной промежуток между крайними метками (первой и последней открытыми страницами на сайте). Время, проведенное посетителем на последней странице, очень сложно правильно рассчитать, так как нельзя сохранить соответствующую временную метку. Причина – достоверно неизвестно, какое именно действие совершил посетитель при работе с этой страницей (например, набрал адрес другого сайта в адресной строке, либо просто закрыл браузер), т.е. открытие последней страницы можно охарактеризовать как ее посещение без взаимодействия с контентом. Поэтому принято вообще не учитывать это время в расчетах AVD. На практике это может существенно занижать AVD сайта – например, в случае если, посетители после открытия pdf-файла с нужной статьей длительное время с ней работают (последней учитываемой временно*й* меткой является момент перехода к pdf-файлу).

Числовые значения вебометрических показателей, которые соответствуют непосредственно рассматриваемым Интернет-ресурсам (указанным в левой колонке), в таблицах приводятся без фигурных скобок. Значения, соответствующие «сайтам в целом», на которых размещены исследуемые ресурсы, даются в фигурных скобках, т.е. в «{ }». Комбинация фигурных и угловых скобок (в виде {< >}) означает указание количества страниц для «сайта в целом», полученное с помощью ИПС Google.

МЕЖДУНАРОДНЫЕ ТЕМАТИЧЕСКИ СПЕЦИАЛИЗИРОВАННЫЕ СИСТЕМЫ УЧЕТА ЦИТИРОВАНИЙ И ОЦЕНКИ НАУКОМЕТРИЧЕСКИХ ПОКАЗАТЕЛЕЙ, ПРИЗНАВАЕМЫЕ ВАК РОССИИ

В соответствии со справочным списком от 30.06.2017 г. Высшей аттестационной Комиссии (ВАК) России при защите диссертаций «признаются» публикации в научных журналах, включенных лишь в ведущие международные системы учета цитирований и расчета наукометрических показателей.

Количество российских изданий (включая переводные и переводные составные), которые входят в список ВАК, соответствующий указанным международным системам (список «С1» от 25.09.2017), было равно 1028. В то же время в основном списке ВАК российских изданий (список «С2» от 04.12.2017) находилось 2202 издания (Таким образом количество журналов в списке «С1» составляет 46,7%. от списка «С2»). Основная часть списка «С2» представлена печатными изданиями, однако есть и небольшое количество электронных журналов.

Многие российские научные журналы из указанных в перечне 1028 изданий, входят сразу в несколько международных систем учета цитирований и оценки наукометрических показателей – как политематических [1], так и специализированных (последние соответствуют категории «К4» по [1]). При этом действующие правила ВАК для российских журналов предусматривают возможность публикации статей не более чем по трем отраслям науки и не более чем по пяти группам специальностей ВАК. Используемое ВАК категорирование групп специальностей носит внутрироссийский характер и в зарубежной практике издания научных журналов не применяется.

Отметим, что большинство зарубежных периодических научных изданий являются тематически специализированными в пределах определенных предметных областей – это отражается и в их названиях. В то же время среди российских журналов из списка «С2» многие имеют названия типа «Вестник …», «Бюллетень …», «Известия …» и пр., – т.е. отражают вид издания, а не его тематическую специализацию [20]. Справедливости ради отметим, что часть таких периодических изданий имеет разбивку по тематическим сериям с собственными названиями. Однако, если в Российском индексе научного цитирования (РИНЦ) издания зарегистрированы «в целом» (т.е. не по отдельным сериям), то и расчет наукометрических показателей осуществляется для «изданий в целом», а не для серий.

Для многих российских изданий из списка «С2», в том числе тематически специализированных, стоит задача войти в список «С1». Проблемы такого вхождения для многочисленных университетских журналов России имеют свою специфику [25], связанную с тематически расплывчатыми названиями (см. предыдущий абзац); с объединением в одном издании нескольких тематически различных серий; со значительной политематичностью содержания публикуемых статей.

Публикации в российских и, особенно, зарубежных журналах, входящих в списки международных систем учета цитирований и расчета наукометриче-

ских показателей, играют важную роль при защитах докторских и кандидатских диссертаций; отдельно учитываются при решении вопросов о выделении научных грантов; важны для обеспечения научного имиджа [16] вузов и НИИ. Эти публикации обычно отдельно отражаются в отчетности организаций о научной деятельности.

Переходим к рассмотрению ресурсов, соответствующих категории «К4» по [1].

1. Zentralblatt Math (zbMath) – ресурс, специализированный на математике. Он размещен на сайте https://zbmath.org. Издатель – Springer. Внутренняя ИПС ресурса обеспечивает возможность поиска не только по ключевым словам и авторам, но и по программному обеспечению, по формулам. Судя по имеющимся в Интернете данным, на ресурсе индексируется более 2300 журналов и иных периодических изданий разных стран по математике и некоторым смежным направлениям.

2. MathSciNet (сайт http://www.ams.org/ mathscinet/) – реферативная база данных (БД) в основном по математике и естественным наукам. По сведениям, приведенным на http://info.mipt.ru/index/materials/n_5fn7l7.html, на этом ресурсе индексируется более 1800 журналов, есть также записи о 85 тыс. монографий и 300 тыс. докладов на научных конференциях. Всего в БД ресурса более 3 млн. записей, причем 2,2 млн из них снабжены рецензиями/рефератами. Однако в некоторых других источниках для этого ресурса указаны более скромные показатели – по крайней мере, по количеству журналов.

Поскольку ресурс расположен на сайте Американского математического общества (сайт http://www.ams.org), то это затрудняет анализ ВМП непосредственно для групп страниц сайта, соответствующих MathSciNet.

3. GeoRef – ресурс, специализированный по направлению «науки о Земле», находится по адресу по адресу https://www.americangeosciences.org/georef/georef-information-services. Отсутствие отдельного ресурса для этого сайта-агрегатора также затрудняет анализ его ВМП.

4. PubMed – это англоязычный ресурс, специализированный на медицинской информации. Расположен по адресу http://www.pubmed.org. Имеет собственную базу индексируемых научных журналов, в которую на 01.07.2017 г. входило и 66 российских журналов из указанного выше перечня ВАК «С1». При этом значительная часть из этих 66 журналов входила и в списки других международных систем учета цитирований и расчета наукометрических показателей, признаваемых ВАК России.

5. ADS (Astrophysics Data System) – этот ресурс содержит более 8 млн статей по астрономии и физике. Размещается по адресу http://adswww.harvard.edu. Это наиболее известный (но не единственный – см. например [26]) ресурс по данной проблематике.

6. Chemical Abstracts (CA) [27] – крупнейший ресурс, специализированный на химической информации. Размещается на сайте https://www.cas.org. Отметим, что в списке журналов ВАКа «С1» (от 30.06.2017 г.) упоминаются две разновидности этого ресурса: *CA core* и *CA* (*pt*).

7. AGRIS – ресурс, специализированный на информации о сельскохозяйственной науке и технологиях; размещается на сайте agris.fao.org. Однако наличие этого ресурса в списке ВАК «С1» подвергается критике за то, что он работает не с журналами в целом, а лишь с отдельными статьями из профильных изданий. Как следствие, другие статьи из тех же журналов (российских и зарубежных) автоматически приобретают статус, соответствующий списку ВАК «С1».

По решению ВАК (см. [1]) с 01.01.2018 г. ресурсы AGRIS и ADS не будут ею «признаваться». Таким образом для категории «К4» с 01.01.2018 г. ВАК будут признаваться два ресурса по математике и по одному ресурсу по биомедицине, химии, наукам о Земле. Следовательно в группе сайтов «К4» вообще нет (и не будет) специализированных ресурсов по физике, информационным технологиям, техническим, экономическим и гуманитарным наукам.

Сводка вебометрических показателей для сайтов указанных нами ресурсов представлена в табл. 1 и 2.

Таблица 1

Показатели для сайтов категория «К4» (первая часть)*

№	Название ресурса	AT, сек.	Count, URLs			Size, Mb	Scholar Google (SG)
			Text/html	Image	Application		
1	Zentralblatt Math (zbMath)	0,34	1	0	0	0,036	0
2	MathSciNet	0,36	417	27	2	4,8	0 {65500}
3	GeoRef	0,78	1	20	2	0,039	0 {103}
4	PubMed	-	0<0>	0	0	0	0
5	ADS (Astrophysics Data System)	0,62	2	6	0	0,039	0
6	Chemical abstracts (CA)	4,22	1975	958	311	73,4	23
7	AGRIS	0,37	543	1046	0	130	504000

* Нумерация категорий здесь и далее в таблицах дается по [1]

Таблица 2

Показатели для сайтов категории «К4» (вторая часть)

№	Название ресурса	Абсолютное к-во ссылок				КУнПос. за месяц	КПросм. (за месяц)	AVD, чч:мм:сс
		Входящие	Вн	Исх				
				Ω	Ψ			
1	Zentralblatt Math (zbMath)	2777206	0	7	7	16626	66510	00:05:01
2	MathSciNet	{14626691}	1457	145	876	{139953}	{559800}	{00:02:27}
3	GeoRef	15	0	60	60	{19772}	{79080}	{00:02:05}
4	PubMed	633642	0	0	0	0	0	00:00:00
5	ADS (Astrophysics Data System)	259294	1	53	81	1269037	5076150	00:01:06
6	Chemical abstracts (CA)	3181843	73503	474	11603	203279	813120	00:08:36
7	AGRIS	737924	9751	65	15996	468278	1873110	00:01:11

Отметим, что применяемые нами методики не позволяют определять некоторые из показателей сайтов (это относится, прежде всего, к данным табл. 1). Основные причины: использование на сайтах парольного доступа для входа на них; размещение информации в БД, расположенных вне сайтов и пр.

Как и следовало ожидать, востребованность ресурсов, данные по которым представлены в табл. 1 и 2, ниже, чем для сайтов-агрегаторов политематической научной информации, которые были рассмотрены в [1]. Обращают на себя внимание низкие AVD для ресурсов ADS и AGRIS, что может свидетельствовать и о «неудовлетворенности» посетителей информацией, представленной на этих ресурсах. С другой стороны, для Chemical Abstracts величина AVD высокая.

МЕЖДУНАРОДНЫЕ ТЕМАТИЧЕСКИ СПЕЦИАЛИЗИРОВАННЫЕ ХРАНИЛИЩА НАУЧНОЙ ИНФОРМАЦИИ, НЕ ПРИЗНАВАЕМЫЕ ВАК

Публикации в изданиях, отражаемых на интернет-ресурсах категории «К5», не признаются ВАК России при защитах кандидатских и докторских диссертаций как имеющие статус, соответствующий изданиям, включенным в списки ВАК «С1» или «С2». Однако часть таких ресурсов достаточно известна в России, что и определяет целесообразность анализа их ВМП. Другие ресурсы категории «К5» рассматриваются в силу их перспективности в будущем. Отметим, что на части рассматриваемых ресурсов рассчитываются некоторые наукометрические показатели; большинство ресурсов категории «К5» размещают материалы только на английском языке.

1. Arxiv.org (http://xxx.lanl.gov) – содержит архив статей по физике, математике, нелинейной динамике, информационным технологиям, количественным методам в биологии и финансах. На ресурсе допускается размещение материалов по инициативе авторов, в том числе оригинальных работ, ранее не опубликованных в периодических изданиях. Однако предоставляемые для размещения материалы предварительно модерируются (проверяются) специалистами в соответствующих предметных областях с целью поддержки качества массива размещаемых работ. Сайт поддерживается Корнельским университетом в США.

2. International Society of Universal Research in Sciences (EyeSource Indexed Open Access Journals – http://isurs.org/master_list.php?topic_id=3). Несмотря на его «универсальное» название, на этом ресурсе отражены лишь публикации из относительно небольшого количества журналов по менеджменту, бизнесу, социальным наукам.

3. CiteSeerX (http://citeseerx.ist.psu.edu) – система была разработана для индексирования научной литературы и автоматического подсчёта индекса цитирования с целью количественного определения значимости отдельных публикаций. В настоящее время в некоторых интернет-источниках эта система рассматривается и как открытая электронная библиотека научных статей, главным образом, в области информатики, прикладной математики и технических наук. Сайт ресурса администрируется Pensylvania State University в США.

4. Citations in Economic (http://citec.repec.org) – специализированный ресурс по экономической тематике, обеспечивающий библиометрический анализ документов из электронной библиотеки RePEC. Позволяет оценивать частоту цитирований документов.

5. EBSCOhost (http://www.ebsco.com) – служба, которая предоставляет доступ к БД англоязычных научных журналов (частично – к полным текстам, частично – только к аннотациям). Бо̀льшая часть ресурсов для EBSCO, перечисляемых далее, специализированные. Поэтому EBSCO рассматривается в настоящей статье, а не в работе по политематическим интернет-ресурсам [1].

Наиболее полезными (интересными) БД EBSCOhost считаются: (**а**) Academic Search Premier (http://www.ebscohost.com/academic/academic-search-premir) – универсальная БД, отращающая по совокупности более чем 3600 журналов (полные тексты есть из 2700 изданий); (**б**) Bisiness Source Premier (http://www.ebscohost.com/academic/business-source-premir) – это БД по примерно 2800 журналам, относящимся к бизнесу и экономике; (**в**) MasterFile

(http://www.ebscohost.com/academic/masterfile-premer) – БД универсального характера, обеспечивающая доступ к библиографическим ссылкам, рефератам, полным текстам работ, более чем 10 тыс. биографий ученых, коллекциям карт, флагов и пр.; (**г**) Business Abstracts with Full Text (http://www.ebscohost.com/academic/business-abstracts-with-full-text) – специализированный ресурс по бизнес-тематике, относящейся к различным направлениям деятельности, в том числе инженерно-технической; (**д**) AppliedScience & Technology Index (http://www.ebscohost.com/academic/applied-science-technology-index) – эта БД обеспечивает индексацию работ в научно-технических журналах – в том числе научных, инженерно-технических и пр. статей. Считается, что эта БД хорошо отражает междисциплинарные исследования и разработки, находящиеся на стыке отраслей; (**е**) EconLiT – отражает литературу по бизнесу/экономике. Объем – 1,8 млн записей, охват – 74 страны. Создается American Economic Association (https://www.ebsco.com/products/research-databases/econlit); (**ж**) American Doctoral dissertations (https://www.ebsco.com/products/research-databases/american-doctoral-dissertations) – содержит материалы диссертаций, защищенных в США.

6. Free Medical Journals (http://www.freemedical-journals.com) – через этот сайт обеспечивается доступ к медицинским журналам, находящимся в свободном доступе.

7. ABC Chemistry – белорусский ресурс на русском языке. Обеспечивает возможности бесплатного доступа к полнотекстовым версиям журналов по химии (сайт http://www.abc.chemistry.bsu.by/free-journals/).

8. World History of Science Online (http://www.dhst-whso.org) – был создан в рамках международного библиографического проекта и используется для классификации и индексации интернет-ресурсов по истории науки и технологий.

9. Nature (https://www.nature.com) – на этом сайте размещены материалы из научных журналов по естественным наукам. Имеется также специальная страничка world library of science (A Global Community for Science Education). Располагается по адресу https://www.nature.com/wls.

10. CAB Direct (www.cabdirect.org) от CABI – БД содержит около 12 млн записей, соответствующих тематике «прикладные науки о жизни».

11. Europe PMC (https://europepmc.org) – БД биомедицинской тематики с доступом к полным текстам статей. Используются инструменты text-mining, а также связи с внешними тематически специализированными БД.

12. International Directory of Philosophy – охватывает соответствующую информацию по 130 странам, касающуюся отделений вузов, исследовательских центров, журналов, а также исследователей (https://www.pdcnet.org/idphil/International-Directory-of-Philosophy).

13. Lingbuzz (https://ling.auf.net) – открытый архив статей по лингвистике. Его ведет (администрирует) университет Тромсё (Норвегия).

14. RePEc: Research Papers in Economics – децентрализованная библиографическая база по экономике. Использует 1900 архивов из 92 стран. Таким образом этот ресурс может рассматриваться как носящий «надстроечный» характер (www.repec.org).

15. Semantic Scholar (https://www.semanticscholar.org/) – ресурс по публикациям в области компьютерных наук. Предназначен для обнаружения наиболее важных статей и связей между ними. С этой целью используются средства интеллектуального анализа данных. Ресурс ведется Allen Institute for Artificial Intelligence.

16. Social Science Research Network (SSRN – http://www.ssrn.com) – крупный репозиторий научных статей и препринтов по ключевым направлениям управленческой и экономической науки. Допускается инициативное размещение работ авторами.

Таким образом, для категории «К5» охват тематических направлений исследований в целом значительно шире, чем для «К4». Вебометрические показатели по ресурсам категории «К5» сведены в табл. 3 и 4.

Таблица 3

Показатели для сайтов категории «К5» (первая часть)

№	Название ресурса	AT, сек.	Count, URLs			Size, MB	Scholar Google
			Text/html	image	application		
1	Arxiv.org	3,49	169193	12	26325	18906	0
2	International Society of Universal Research in Sciences	1,02	1	5	0	0,019	0 {0}
3	CiteSeerX	0,58	1389	9	8	47,6	361000
4	Citations in Economic	0,73	>1000000 <162000>	-	-	-	629
5	**EBSCO**	3,99	1749	439	16	154	1
6	Free Medical Journals	0,27	5315	9	0	9,8	2
7	ABC Chemistry	0,58	45	0	0	0,387	0 {1}
8	World History of Science Online	1,75	754	29	1	11,6	0
9	Nature	0,55	4335	89	25	446	157000
10	CAB Direct	1,84	6	77	12	4,4	1180000
11	Europe PMC	0,49	11653	245	2483	329888	2340000
12	International Directory of Philosophy	1,99	1	6	0	0,013	0 { 133 000 }
13	Lingbuzz	13,34	58346	0	1389	984	807
14	RePEc	1,95	1	11	0	0,06	19
15	Semantic Scholar	1,13	82594	1	0	20621	0

Таблица 4

Показатели для сайтов категории «К5» (вторая часть)

№	Название ресурса	Абсолютное количество ссылок				К_Ун_Пос. за месяц	К_Просм. за месяц	AVD, чч:мм:сс
		Входящие	Вн	Исх				
				Ω	Ψ			
1	Arxiv.org	1907148	1181926	356829	2149722	72280	289110	00:03:19
2	International Society of Universal Research in Sciences	176	0	63	63	{3246}	{12990}	{00:02:32}
3	CiteSeerX	29380675	3743	97	4891	1019115	4076460	00:01:30
4	Citations in Economic	1307096	-	-	-	162640	650550	00:01:35
5	**EBSCO**	3471323	35380	1374	34639	77022	289140	00:02:25
6	Free Medical Journals	1153883	26323	6370	12393	12397	49590	00:01:48
7	ABC Chemistry	{47120}	1175	765	1082	{97808}	{391230}	{00:01:56}
8	World History of Science Online	2632	31358	457	1223	4100	12000	00:00:18
9	Nature	233941992	10944	2508	5530	1379427	5517720	00:02:50
10	CAB Direct	848687	18	16	58	72857	291420	00:01:27
11	Europe PMC	7546570	112908	268	70331	150669	602670	00:01:20
12	International Directory of Philosophy	10	0	78	78	{26045}	{104190}	{00:01:51}
13	Lingbuzz	17525	109413	17	1455	7639	30570	00:02:15
14	RePEc	30834	0	54	54	162640	650550	00:01:45
15	Semantic Scholar	4705352	158911	5035	12401	457939	1831770	00:01:14

Можно сделать вывод, что используемые авторами настоящей статьи программные средства определяют некоторые «объемные» характеристики сайтов, включенных в табл. 1 и 2. Однако из-за различий владельцев сайтов в подходах к их построению и размещению на них информации сравнивать целесообразно главным образом динамику этих показателей для отдельных сайтов во времени, но не вебометрические показатели разных сайтов.

Посещаемость ряда сайтов этой группы весьма высока и сопоставима с теми политематическими сайтами, которые были рассмотрены в [1]. Однако обращают на себя внимание достаточно низкие AVD для всех сайтов группы «К5».

САЙТЫ, ПРЕДНАЗНАЧЕННЫЕ ДЛЯ ИНФОРМАЦИОННОЙ ПОДДЕРЖКИ ПУБЛИКАЦИЙ НА НАЦИОНАЛЬНЫХ ЯЗЫКАХ И/ИЛИ В ЖУРНАЛАХ ОПРЕДЕЛЕННЫХ СТРАН, ГРУПП СТРАН

Основная цель создания и ведения группы ресурсов категории «К6» по [1] – агрегация и повышение известности в информационном пространстве научных работ, опубликованных на национальных языках и отражающих национальную тематику. «Национальные ресурсы» существуют лишь в некоторых крупных и развитых в научном отношении странах мира. Для России соответствующий национальный ресурс это elibrary.ru (он был рассмотрен в [1]). Сайты категории «К6» по терминологии из [28] могут рассматриваться как важный компонент «национального информационного ресурса» соответствующих стран. Мы также включаем в эту группу ресурсы, отражающие публикации в изданиях, выходящих на определенных континентах.

«Национальные ресурсы» используются для продвижения в научно-информационном пространстве публикаций в национальных журналах, в том числе и выходящих на языках соответствующих стран. Наряду с этим для повышения известности практикуется и размещение материалов некоторых из журналов на платформе Web of Science (для России это проект «Russian Science Citation Index»).

1. "***Українiка наукова***" – реферативная база данных, национальный ресурс Украины (http://nbuv.gov.ua/node/512) [29], который ведет (администрирует) Национальная библиотека Украины им. В.И. Вернадского. Одновременно ставятся/решаются и вопросы «интеграции» публикаций украинских исследователей с «европейскими наукометрическими системами» [30]. Издается и реферативный журнал «Джерело», отражающий публикации на украинском языке.

2. Chinese Social Sciences Citation Index (***CSSC***) – индекс цитирования китайских общественно-научных статей из более чем 500 китайских журналов соответствующего направления (http://cssci.nju.edu.cn). Ресурс поддерживается Нанкинским университетом и Гонконгским научно-техническим университетом. В качестве приложения к БД ресурса публикуется отчет со статистическим анализом состояния общественных наук в Китае [31, 32].

Интересно, что на Тайване «Национальной академией образовательных исследований» создан отдельный индекс цитирования для журналов гуманитарного направления, выпускаемых на острове, – «Taiwan Humanities Citation Index» (http://terms.naer.edu.tw).

3. China citation database (***CCD***) – сайт http://ccd.cqvip.com (англоязычная страничка – http://en.cqvip.com/ccd.html**).**

4. China science and technology journal database (***CSTJ***) – сайт http://lib.cqvip.com (англоязычная страница – http://en.cqvip.com/cstj.html).

В научной периодике (например, [31, 32]), а также в Википедии, имеются ссылки еще на два китайских сайта научной информации, относящихся к категории «К6»: (**А**) Chinese Science Citation Database (CSCD) – разработка Центра информации и документации Китайской Академии Наук (http://sciencechina.cn/index.jsp). На сайте имеется переключатель между китайским и английским языками. (Б) China Scientific and Technical Papers and Citations (CSTPC) – разработка Китайского Института научной и технической информации, некоторого аналога российского ВИНИТИ РАН (сайт на английском языке – http://www.istic.ac.cn/English/, а на китайском – http://www.istic.ac.cn). На стартовой странице англоязычной версии сайта: нет средства переключения на китайский язык; переходы по гиперссылкам, соответствующие аббревиатуре CSTPC, не просматриваются.

Все русскоязычные публикации, доступные в Интернете по этим двум ресурсам (CSCD и CSTPC), относительно давние – заканчиваются 2009 г. Отметим, что по [32] на основе CSTPC формируется и публикуется аналитический отчет «Chinese S&T Journal Citations Report», а на основе CSCD – «Chinese Scientometric Indicator». Последний, судя по сведениям из [32], содержит более 190 различных наукометрических характеристик.

5. Corea Science Citation Index Service– национальный индекс научного цитирования республики Корея (http://ksci.kisti.re.kr).

В Японии есть несколько [33, 34] «национально ориентированных» баз данных научной информации, но мы отразим только две из них.

6. CiNii Articles – CiNii Articles Incorporated Databases (https://support.nii.ac.jp/ja/cia/ cinii_db, англоязычная страница – https://support.nii.ac.jp/en/cia/cinii_db). Ресурс предназначен для информационной поддержки публикаций на японском языке (более 15 млн статей из 3600 журналов). Помимо статей обеспечивается доступ к книгам и диссертациям.

7. J-STAGE (Japan Science and Technology Information Aggregator, Electronic) – система обеспечения доступа к электронным журналам Японии (https://www.jstage.jst.go.jp).

8. SciELO – Scientific Electronic Library OnLine [35] – ресурс ориентирован на журналы, издаваемые в развивающихся странах, хотя первоначально на нем отражались в основном статьи из изданий Бразилии (http://www.scielo.br).

9. ERIHPLUS (European Reference Index for the Humanities and Social Sciences – https://dbh.nsd.uib.no/publiseringskanaler/erihplus/) – ресурс работает только с одобренными им научными журналами, т.е. осуществляет их отбор как для категорий «К1» и «К4».

10. African Journals OnLine (AJOL) – отражает публикации в журналах, издаваемых на Африканском континете (https://www.ajol.info).

11. Indian Citation Index (www.indiancitationindex.com) – национальный ресурс для поддержки индекса цитирования в Индии.

12. I.S.A.C. – Iranian Scientific Association Commission (https://isacmsrt.ir) [36] – содержит сведения по научным изданиям в Исламской Республике Иран.

Вебометрические показатели для сайтов этой группы организаций представлены в табл. 5 и 6.

Таблица 5

Показатели для сайтов категории «К6» (первая часть)

№	Название ресурса	AT, сек.	Count, URLs			Size, MB	Scholar Google
			Text/html	image	application		
1	Українiка наукова	6,87	1	4	6	0,05	0 {145000}
2	CSSC - Chinese Social Sciences Citation Index	2,32	4	18	0	0,531	0
3	CCD - China citation database	1,45	232	145	0	1,1	0
4	CSTJ - China science and technology journal database	1,49	66439	173	0	2668	0
5	Corea Science Citation Index Service	1,93	166203	134	8	4991	0
6	CiNii Articles (на японском)	2,31	34	0	0	0,97	0
7	CiNii Articles (на английском)	2,31	33	0	0	0,92	0
8	J-STAGE	1,56	27800	1409	11925	3513	3350000
9	SciELO	0,90	1533916	11656	191164	189805	365000
10	ERIHPLUS (European Reference Index for then Humanities)	0,67	464	0	0	4,2	0
11	African Journals OnLine	1.07	344233	576	7066	10900	96500
12	Indian Citation Index	1.52	39	65	0	9,2	0
13	I.S.A.C.	1.58	4266	302	159	262	0

Таблица 6

Показатели для сайтов категории «К6» (вторая часть)

№	Название ресурса	Абсолютное количество ссылок				К_Ун_Пос. за месяц	К_Просм. за месяц	AVD, чч:мм:сс
		Входящие	Вн	Исх				
				Ω	Ψ			
1	Українiка наукова	93	0	107	107	{26994}	{107970}	{00:01:16}
2	CSSC – Chinese Social Sciences Citation Index	616300	3	0	0	305898	1223580	00:02:24
3	CCD – China citation database	139	903	28	98	73470	293880	00:00:14
4	CSTJ – China science and technology journal database	577685	5528630	6269	607443	146941	587760	00:02:20
5	Corea Science Citation Index Service	117551	1764579	31885	229308	52358	209430	00:00:49
6	CiNii Articles (на японском)	{21642}	1058	90	1033	{410703}	{1642800}	{00:00:23}
7	CiNii Articles (на английском)	{21642}	1025	95	1037	{410703}	{1642800}	{00:00:23}
8	J-STAGE	6304620	389341	2148	3643	583593	2334360	00:02:28
9	SciELO	13801480	4368059	242583	797578	472483	1889940	00:02:06
10	ERIHPLUS (European Reference Index for then Humanities)	{1064400}	3660	91	1138	{167666}	{670650}	{00:02:39}
11	African Journals OnLine	1138944	2832048	59768	158257	99722	398880	00:02:35
12	Indian Citation Index	105998	610	0	0	8330	33330	00:04:32
13	I.S.A.C.	249839	108892	641	10338	10023	40080	00:01:30

Для этой группы сайтов можно сделать вывод, что характеристики посещаемости страниц ресурсов на национальных языках и на английском языке серьезно различаются. Отметим также, что Scholar Google позволил проанализировать «количество страниц» лишь примерно для половины рассматриваемых ресурсов.

Среди ресурсов этой группы наибольшее значение AVD имеет Indian Citation Index. Низкое значение для Corea Science Citation Index Service может, вероятно, быть объяснено «заходами» на сайт случайных посетителей, которые не предполагают на нем работать. В то же время, судя по интерфейсу сайта, на ресурсе используются современные подходы к агрегированию и представлению в наглядной форме сводных показателей по научной информации.

ИНТЕРНЕТ-РЕСУРСЫ С КАТАЛОГАМИ ДИССЕРТАЦИЙ И АВТОРЕФЕРАТОВ, ИХ ТЕКСТАМИ

Диссертации играют важнейшую роль в фиксировании уровня подготовки научных кадров [37], в агрегировании информации, ее распространении. Наряду с монографиями, диссертации включают тематически специализированные литературные обзоры, фрагменты материалов из научных статей авторов, обобщения опубликованных ими материалов, сведения о разработанных авторами программах для ЭВМ, БД, полученных патентах. Анализ номенклатуры защищенных диссертаций позволяет оценивать актуальные направления научных исследований, процессы формирования/развития научных школ и пр. На практике для пользователей из России наибольшую важность представляют русскоязычные диссертации, на тексты которых (судя по многочисленным коммерческим предложениям в Интернете) существует устойчивый платежеспособный спрос. При этом контроль защищаемых (защищенных) диссертаций на наличие не оформленных надлежащим образом заимствований (не оригинального текста, плагиата) осуществляется как в советах по защите диссертаций, так и ресурсом www.dissernet.org.

1. Международная полнотекстовая база по магистерским и докторским диссертациям содержит примерно 3,5 млн диссертаций из 88 стран от 2700 организаций (http://search.proquest.com [8]). Ежегодное добавление – 100 тыс. новых диссертаций. Однако доступ к ресурсу требует ввода имени учетной записи и пароля. Фактически доступ к ресурсу имеют лишь работники некоторых организаций, список которых перечислен на самом ресурсе. При входе на ресурс российских пользователей автоматически открывается страница с русскоязычным интерфейсом. Отметим, что ресурс www.proquest.com мы уже рассматривали в [1] в качестве политематического.

2. OCLC WorldCat (http://www.worldcat.org) – этот ресурс помимо диссертаций содержит ссылки на книги и статьи.

3. DART-Europe E-theses portal – электронный ресурс Стэнфордского университета (США) (https://searchworks.stanford.edu/view/10436087).

4. NDLTD (Thesis Resources) – ресурс содержит, в частности, удобную ИПС по диссертациям, защищенным в разных странах мира (http://www.ndltd.org/resources/find-etds), однако России в этом списке нет. Помимо сайтов отдельных стран ИПС обеспечивает поиск и на международных ресурсах – при этом используются и возможности Google Scholar.

Информация по диссертациям есть также на рассмотренных ранее ресурсах категории «К6» – CiNii Articles, American Doctoral dissertations.

Многие вузы США обеспечивают доступность материалов по защищенным в них диссертациям без ограничений по срокам.

На сайтах тех российских организаций, где публикуется информация о защите диссертаций (или они были защищены ранее) открыт доступ к их текстам, а также к «шлейфам» сопроводительных документов, включая отзывы оппонентов и ведущих организаций. Фактическая доступность этой информации для исследователей снижается из-за таких факторов: (**а**) информация разбросана по сайтам нескольких сотен организаций – в основном вузов и НИИ; (**б**) структуры их сайтов и места расположения на них информации по диссертациям существенно различаются – это затрудняет обнаружение нужных сведений, а также использование «агентных технологий» для мониторига появления новой информации в соответствующих разделах сайтов; (**в**) по правилам ВАК продолжительность обязательного размещения информации (текстов диссертаций и материалов шлейфа документов) достаточно ограничена. На практике многие организации не заинтересованы в информационном продвижении сведений о защищаемых/защищенных в них диссертациях. Основная причина – чем менее эта информация известна, тем меньше вероятность появления претензий к качеству защищенных (или предполагаемых к защите) диссертационных работ; к процедурам защиты; к оформлению документов, связанных с защитами; к попаданию в поле зрения средств массовой информации, а также ресурса Диссернет (www.dissernet.org) и пр. Получение отзывов на авторефераты диссертаций обычно обеспечивается не за счет их размещения на сайтах, а путем «задействования» налаженных личных связей между учеными или организациями; (**г**) для зарубежных исследователей, не знающих русского языка, русскоязычные диссертации являются малодоступными.

Далее приведен выборочный перечень российских ресурсов по диссертациям.

5. На сайте Российской государственной библиотеки – РГБ (http://diss.rsl.ru) предоставляется доступ к текстам защищенных в России диссертаций и авторефератов либо в читальных залах библиотеки, либо с некоторых ПЭВМ тех организаций, которые заключили соответствующие договора с РГБ. На апрель 2017 г. в базе ресурса имелось около 900 тыс. диссертаций и авторефератов.

6. Ресурс www.dissercat.com предоставляет тексты русскоязычных диссертаций и авторефератов за плату всем желающим, хотя часть материалов выложена в открытый доступ. На www.dissercat.org имеется достаточно удобная ИПС.

7. На www.dslib.net, судя по информации на его стартовой странице, предлагаются тексты русскоязычных (около 800 тыс.) и англоязычных (около 1,2 млн) диссертаций и авторефератов. Часть материалов представлена в открытом доступе.

8. На сайте https://new-disser.ru часть материалов также находится в открытом доступе, хотя получение основной массы текстов диссертаций носит платный характер.

9. Сайт http://www.dissforall.com, судя по информации, размещенной на его стартовой странице, это электронная библиотека диссертаций (ЭБД), защищенных в России за последние 20 лет. При этом можно бесплатно скачать одну диссертацию (по акции).

В Интернете есть достаточно много и других предложений от коммерческих фирм по доставке за плату полнотекстовых версий диссертаций и авторефератов (причем не только из России, но и из таких стран ближнего зарубежья, как Белоруссия и Украина).

Помимо приведенных специализированных российских сайтов, информация по диссертациям и авторефератам отражена и на www.elibrary.ru. На сайте ГПНТБ России открыт доступ к авторефератам защищенных в России диссертаций, в том числе и зарубежных аспирантов, обучавшихся в российских вузах.

Сведения по сайтам категории «К7» отражены в табл. 7 и 8.

Таблица 7

Показатели для сайтов категории «К7» (первая часть)

№	Название ресурса	AT, сек.	Count, URLs			Size, Mb	Scholar Google
			Text/html	image	application		
Зарубежные ресурсы							
1	http://search.proquest.com	0,61	188	84	47	7,1	2180000
2	OCLC WordCat	0,82	28868	81	22	2509	52
3	https://searchworks.stanford.edu	1,70	5	0	0	0,132	0
4	Ресурс NDLTD (Thesis Resources)	0,95	1	0	0	0,051	0
Российские ресурсы							
5	http://diss.rsl.ru	0,64	581	480	31	52,7	0
6	www.dissercat.com	1,51	320383	192283	0	29441	54
7	www.dslib.net	0,81	697276	1153858	758	62622	160
8	https://new-disser.ru	0,24	196803	109631	112	18710	20
9	http://www.dissforall.com	0,93	164591	42	8	3792	0

Таблица 8

Показатели для сайтов категории «К7» (вторая часть)

№	Название ресурса	Абсолютное количество ссылок				К_Ун_Пос. за месяц	К_Просм. за месяц	AVD, чч:мм:сс
		Входящие	Вн	Исх				
				Ω	Ψ			
Зарубежные ресурсы								
1	http://search.proquest.com	9040768	206	88	2747	485468	1941870	00:03:34
2	OCLC WordCat	59563645	-	-	-	517359	2069430	00:02:58
3	https://searchworks.stanford.edu	{301942}	4	56	196	{1478592}	{5914380}	{00:00:54}
4	Ресурс NDLTD (Thesis Resources)	5107	0	100	100	{23953}	{95820}	{00:04:57}
Российские ресурсы								
5	http://diss.rsl.ru	820193	76733	522	3102	110940	443760	00:01:27
6	www.dissercat.com	483108	5282664	133	795767	172189	441180	00:01:33
7	www.dslib.net	123601	-	1076	-	62206	248820	00:01:06
8	https://new-disser.ru	13037	6138885	8	172655	9744	18605	00:01:15
9	http://www.dissforall.com	5012	2940303	164590	529183	11460	45840	00:02:28

Можно сделать вывод, что востребованность ресурсов, отраженных в табл. 7 и 8 в целом достаточно высокая. Высокая посещаемость зарубежных (англоязычных) ресурсов является ожидаемой, так как англоязычных аспирантов и исследователей многократно больше, чем русскоязычных.

Относительно низкие AVD для российских ресурсов группы «К7» можно, очевидно, объяснить тем, что во многих случаях поиск материала идет только целенаправленно, а изучение списков диссертаций по определенной тематике выполняется редко.

ИНФОРМАЦИОННО-АНАЛИТИЧЕСКИЕ И ИНФОРМАЦИОННЫЕ СИСТЕМЫ ПО НАУЧНЫМ ЖУРНАЛАМ

В качестве средств для поиска и анализа публикаций в журналах (а также самих журналов) может рассматриваться большинство политематических ресурсов, уже проанализированных в [1] – включая www.sciencedirect.com, elibrary.ru и др. Поэтому здесь мы рассматриваем лишь некоторые дополнительные сайты. Часть их может считаться и системами поиска научной информации в целом.

1. Scimago Journal and Country Rank (SJR) – на ресурсе рейтингуются отдельные научные издания и страны, в которых издаются журналы (http://www.scimagojr.com/).

2. CAS Source Index (CASSI) Search Tool (http://cassi.cas.org) – бесплатный ресурс для получения библиографической информации о химических журналах (аббревиатура и полное название издания, ISSN и др.), индексированных с 1907 г. Является подразделением American Chemical Society.

3. Social Science Research Network (SSRN – https://www.ssrn.com/en/) – фактически – полифункциональный, тематически специализированный сайт.

4. Реестр электронных научных изданий Роскомнадзора (Информрегистр –http://catalog.inforeg.ru).

Данные по вебометрическим показателям этих ресурсов отражены в табл. 9 и 10.

Обращает на себя внимание высокая посещаемость зарубежных ресурсов этой группы по сравнению с сайтом Информрегистра. При этом AVD для приведенных трех зарубежных сайтов значительно различаются. Российским исследователям наиболее известен SJR (Scientific Journal Ranking или Scientific Journal and Country Ranking – http://www.scimagojr.com). К сожалению, российские журналы в нем представлены слабо.

САЙТЫ ОТДЕЛЬНЫХ ЗАРУБЕЖНЫХ НАУЧНЫХ ЖУРНАЛОВ

Отметим, прежде всего, что имеются публикации как по вебометрическим показателям научной периодики зарубежных стран (например, Украины [38]), так и по сравнению особенностей российских и зарубежных научных журналов [39]. Количество издаваемых журналов в развитых странах очень велико. Как следствие, направление исследований «Вебометрические показатели сайтов журналов» – весьма обширное. Поэтому для сравнительного анализа мы выбрали лишь отдельные зарубежные и российские (см. следующий раздел) журналы, имеющие непосредственное отношение к теме настоящей статьи.

1. «Scientometrics» (An International Journal for All Quantitative Aspects of the Science of Science, Communication in Science and Science Policy). Издается совместно Akadémiai Kiadó и Springer Science+Business Media . В открытом доступе на 13.07.2017 было лишь 158 статей из общего количества опубликованных 5128. Импакт-фактор (взят с сайта журнала) равен 2,147. Стартовая страница – http://www.springer.com/computer/database+management+%26+information+retrieval/journal/11192. Каталог архива журнала – https://link.springer.com/ journal/volumesAndIssues/ 11192.

2. «Journal of Infometrics» издательства Elsevier (https://www.journals.elsevier.com/journal-of-informetrics/). Импакт фактор – 2,920. Публикуются статьи по «количественным аспектам информационной науки».

3. «Journal Citation Reports» (JCR – https://clarivate.com/products/journal-citation-reports/).

Считается библиометрическим справочником статистических данных, отражающих продуктивность и степень использования научных журналов.

4. COLLNET – «Journal of Scientometrics and Information Management», Индия (архив издания находится по адресу http://www.tandfonline.com/loi/tsim20).

5. «Information Resources Management Journal» (https://www.igi-global.com/journal/information-resources-management-journal-irmj/1073).

6. «Journal of Data and Information Science» (http://www.jdis.org/EN/2096-157X/home.shtml)

7. «Science» (American Association for the Advancement of Science) – этот авторитетный журнал выходит еженедельно (http://www.sciencemag.org).

Судя по оценкам на сайте Википедии, суммарное количество читателей бумажной и электронной версий – порядка одного миллиона человек. Вебометрические показатели по этому журналу мы приводим для сравнения. Показатели для сайтов категории «К9» даны в табл. 11, 12.

Обращает на себя внимание значительный разброс во времени открытия стартовых страниц сайтов группы «К9». Частично его можно, вероятно, объяснить разной насыщенностью этих страниц графическими объектами.

Количества просмотров для зарубежных изданий достаточно велики, но они относятся к сайтам в целом. Величины AVD можно считать находящимися на среднем уровне.

Таблица 9

Показатели для сайтов категории «К8» (первая часть)

№	Название ресурса	AT, сек.	Count, URLs			Size, MB	Scholar Google
			Text/html	image	application		
1	SJR	0,24	>1000000 <570000>	-	-	-	0
2	CASSI	0,35	5	7	0	0,183	0
3	SSRN	2,60	128	29	0	9,29	0 {3}
4	Информрегистр	14,55	14384	11	0	321	0

Таблица 10

Показатели для сайтов категории «К8» (вторая часть)

№	Название ресурса	Абсолютное количество ссылок				К_Ун_Пос. за месяц	К_Просм. за месяц	AVD, чч:мм:сс
		Входящие	Вн	Исх				
				Ω	Ψ			
1	SJR	6983364	-	-	-	195996	783990	00:04:30
2	CASSI	265280	12	7	13	203279	813120	00:09:18
3	SSRN	{3767033}	14923	3755	9372	201997	807990	00:01:56
4	Информрегистр	51430	141118	887	400906	6103	24420	00:00:50

Таблица 11

Показатели для сайтов категории «К9» (первая часть)

№	Название ресурса	AT, сек.	Count, URLs			Size, MB	Scholar Google
			Text/html	image	application		
1	Scientometrics	2,25	4	0	0	0,456	0
2	Journal of Infometrics	0,98	13	0	0	0,889	0
3	JCR	2,68	3	0	0	0,390	0
4	COLLNET Journal of Scientometrics and Information Management	0,88	2	0	0	0,0001	0 {3210000}
5	Information Resources Management Journal	6,99	7	0	0	0,824	0 {41400}
6	Journal of Data and Information Science	1,21	1	0	0	0,0001	0
7	Science	0,78	34483	33033	723	6220	0

Таблица 12

Показатели для сайтов категории «К9» (вторая часть)

№	Название ресурса	Абсолютное количество ссылок				К_Ун_Пос. за месяц	К_Просм. за месяц	AVD, чч:мм:сс
		Входящие	Вн	Исх				
				Ω	Ψ			
1	Scientometrics	128	13	113	404	{1835237}	{7340940}	{00:02:30}
2	Journal of Infometrics	24	18	661	1582	{1122953}	{4491810}	{00:02:10}
3	JCR	1462878	4	219	627	{52851}	{211410}	{00:02:11}
4	COLLNET Journal of Scientometrics and Information Management	17	2	1	1	{1025011}	{4100040}	{00:02:36}
5	Information Resources Management Journal	{2482538}	36	201	1141	{101648}	{406590}	{00:02:23}
6	Journal of Data and Information Science	{235}	0	0	0	{4000}	{12000}	{00:02:42}
7	Science	44251919	1418973	46573	968595	598903	2395620	00:01:41

САЙТЫ НЕКОТОРЫХ РОССИЙСКИХ НАУЧНЫХ ЖУРНАЛОВ, ИМЕЮЩИХ ОТНОШЕНИЕ К ИНФОРМАЦИОННЫМ РЕСУРСАМ

При отборе изданий для этого раздела (категория «К10» по [1]) учитывалось наличие у рассматриваемых журналов отдельных сайтов. Значения ВМП изданий отражены в табл. 13, 14. Отметим, что по крайней мере в российской практике статьи по науковедческой тематике, информационному менеджменту в сфере научных исследований и т.п. публикуют и «библиотековедческие» журналы [40].

Как и следовало ожидать, показатели посещаемости сайтов для русскоязычных журналов оказались на порядки меньше, чем для англоязычных. При этом значения AVD примерно аналогичны тем, которые были определены для зарубежных изданий. Отметим также, что на показатели посещаемости влияет количество статей, ежегодно публикуемых в изданиях.

САЙТЫ ДЛЯ ПРИСВОЕНИЯ DOI СТАТЬЯМ И ORCID АВТОРАМ

Причина рассмотрения нами этих двух функционально специализированных сайтов (категория «К11» по [1]) – полученная на этих ресурсах информация может использоваться при оформлении в статьях библиографических списков и сведений об авторах. Для сайтов этой группы ВМП представлены в табл. 15 и 16.

CrossRef (www.crossref.org) – этот сайт дает возможность присвоения статьям DOI для обеспечения удобного доступа к ним по гиперссылкам, включаемым в библиографические описания источников в списках литературы к статьям. В настоящее время получением DOI занимаются в основном редакции журналов, публикующих статьи. Встречаются также и предложения некоторых коммерческих фирм о присвоении DOI ранее уже опубликованным работам. Для большинства российских авторов преимущества использования DOI в библиографических списках пока не вполне очевидны.

На сайте orcid.org ведется международная БД по исследователям, зарегистрировавшимся на ресурсе в инициативном порядке. Оценить долю русскоязычных авторов (в том числе и не имеющих публикаций), которые зарегистрированы на ресурсе orcid.org, не представляется возможным.

Отметим, что на ресурсе elibray.ru всем зарегистрированным в Science Index авторам также присваивается уникальные AuthorID, а также SPIN-коды (однако они используются, в основном, только в России).

Здесь мы не рассматриваем многочисленные российские сайты, содержащие рубрики УДК. Причины: эти индексы носят чисто внутрироссийский характер; по крайней мере в отношении информационных технологий рубрики УДК носят не совсем удачный характер. Кроме того, многие авторы нередко указывают индекс УДК по аналогии с ранее уже опубликованными работами.

При необходимости указания в статьях категорий специальностей научных работников (это требуется лишь в немногих российских журналах) авторы обычно пользуются сайтом ВАК, на котором размещена актуальная версия соответствующего рубрикатора.

Учитывая весьма специфическое функциональное назначение сайтов этой группы показатели посещаемости для них следует считать достаточно высокими. При этом, судя по AVD, Orcid.org пока, видимо, еще не используется достаточно широко для получения «справок» об отдельных ученых. Кроме того и количество зарегистрированных на нем ученых пока сравнительно невелико (по отношению к общему числу исследователей в мире). Для сравнения: по состоянию на 13.10.2017 количество зарегистрированных в Science Index (на elibrary.ru) российских исследователей (468 тыс.) меньше по сравнению с Orcid.org (3927 тыс.) всего примерно в 8,4 раза. Это

значительно меньше, чем соотношение между зарубежными и российскими исследователями. Такая ситуация может свидетельствовать о более полном «охвате» регистрацией российских (точнее – русскоязычных) исследователей на национальном сайте elibrary.ru. Отметим также, что регистрация российских исследователей на Orcid.org сейчас активно стимулируется – в сведения об авторах редакции многих журналов стали требовать включать Orcid-коды авторов.

Таблица 13

Показатели для сайтов ОА категории «К10» (первая часть)

№	Название ресурса	AT, сек.	Count, URLs			Size, MB	Scholar Google
			Text/html	image	Application		
1	Информационные ресурсы России http://rosenergo.gov.ru/information_and_analytical_support/informatsionnie_resursi_rossii	1,68	95	0	0	3,32	0
2	Научная периодика: проблемы и решения (www.nppir.ru)	3	49	0	0	2,02	0
3	Научно-техническая информация ч.1 и ч.2 http://www2.viniti.ru/products/zhurnaly-viniti-ran-v-perechne-vak	24,99	23	0	0	1,13	0 {8}
4	Науковедение http://naukovedenie.ru	2,39	210	276	4367	2726	425
5	Электронное издание «Научная Россия» https://scientificrussia.ru/	0,79	83144	133418	22	10199	0
6	Информационные процессы http://www.jip.ru	0,36	540	12	514	447	383
7	Вестник компьютерных и информационных технологий http://www.vkit.ru	0,81	1627	1153	6	72,9	0

Таблица 14

Показатели для сайтов ОА категории «К10» (вторая часть)

№	Название ресурса	Абсолютное количество ссылок				К_Ун_Пос. за месяц	К_Просм. за месяц	AVD, чч:мм:сс
		Входящие	Вн	Исх				
				Ω	Ψ			
1	Информационные ресурсы России http://rosenergo.gov.ru/information_and_analytical_support/informatsionnie_resursi_rossii	{2054386}	368	194	3589	{6148}	{24600}	{00:05:15}
2	Научная периодика: проблемы и решения (www.nppir.ru)	8255	404	389	3031	1051	4200	00:00:12
3	Научно-техническая информация ч.1 и ч.2 http://www2.viniti.ru/products/zhurnaly-viniti-ran-v-perechne-vak	9	239	212	1031	{7487}	{29940}	{00:01:55}
4	Науковедение http://naukovedenie.ru	10912	2403	82	2337	14700	58800	00:01:11
5	Электронное издание «Научная Россия» https://scientificrussia.ru	1419477	4054457	64941	1273371	18748	75000	00:01:42
6	Информационные процессы http://www.jip.ru	5853	613	1	2	2193	8760	00:00:51
7	Вестник компьютерных и информационных технологий http://www.vkit.ru	530	23117	597	11340	1250	5010	00:02:47

Таблица 15

Показатели для сайтов категории «К11» (первая часть)

№	Название ресурса	AT, сек.	Count, URLs			Size, MB	Scholar Google
			Text/html	image	application		
1	CrossRef	1.08	1073	370	44	191	1
2	Orcid.org	29.99	20769	529	42	2150	0

Таблица 16

Показатели для сайтов категории «К11» (вторая часть)

№	Название ресурса	Абсолютное количество ссылок				К_Ун_Пос. за месяц	К_Просм. за месяц	AVD, чч:мм:сс
		Входящие	Вн	Исх				
				Ω	Ψ			
1	CrossRef	31317717	84319	1619	10354	88840	355350	00:02:50
2	Orcid.org	11633212	1383687	2831	394699	218510	874050	00:00:10

ВЫВОДЫ

1. На сайтах тематически специализированных международных систем учета цитирований и оценки наукометрических показателей, признаваемых ВАК России, содержатся достаточно большие объемы информации, сопоставимые по величине с политематическими ресурсами, рассмотренными нами ранее в [1]. Характеристики посещаемости этих сайтов и среднего времени нахождения на них пользователей достаточно высокие.

2. Международные тематически специализированные хранилища научной информации, не признаваемые ВАК России (категория «К5»), более разнообразны по содержанию чем аналогичные хранилища, признаваемые ВАК (категория «К4»). Однако сайты категории «К5» в среднем имеют меньшие по сравнению с категорией «К4» вебометрические показатели.

3. Для сайтов с национальными информационными ресурсами зарубежных стран (категория «К8» – информационные и информационно-аналитические системы по научным журналам) вебометрические показатели также достаточно высокие. Отметим наличие тематически специализированных национальных БД научной информации.

4. Специализированные российские интернет-ресурсы с каталогами диссертаций и авторефератов достаточно востребованы, несмотря на то, что получение полнотекстовых версий материалов на большинстве таких сайтов предлагается за плату. Вероятные причины: возможности бесплатного доступа к полнотекстовым версиям диссертаций со служебных ПЭВМ вузов и НИИ, судя по всему, значительной долей пользователей по разным причинам не реализуются; сама по себе номенклатура защищенных диссертаций, фамилии их авторов, сроки и места защит важны для многих исследователей.

5. Большинство российских исследователей (включая и аспирантов) не работают с каталогами зарубежных диссертаций, хотя на этих ресурсах накоплено много интересных материалов. Их востребованность у отечественных пользователей может повыситься, если от преподавателей по крайней мере ведущих российских вузов будут требовать присуждения ученых степеней престижными зарубежными вузами.

6. Специализированные сайты информационных и информационно-аналитических систем по научным журналам значительно различаются по востребованности. Российские пользователи могут также получать соответствующую информацию на политематических ресурсах, включая elibrary.ru.

7. Интернет-ресурсы зарубежных научных журналов значительно различаются по посещаемости. Большая часть таких ресурсов носит не автономный характер, а сгруппирована на сайтах издательств, медиахолдингов.

8. Посещаемость сайтов российских научных журналов по рассматриваемым в настоящей статье направлениям исследований, относительно невысокая. По сравнению с тематически аналогичными зарубежными журналами эти показатели ниже. Основные причины такой ситуации: меньшее количество русскоязычных исследователей, работающих в соответствующих научных направлениях, по сравнению с англоязычным; сведения по опубликованным материалам доступны на сайтах-агрегаторах elibrary.ru, «Киберленинка» и др.

9. Сайты, обеспечивающие присвоение DOI статьям и ORCID авторам достаточно востребованы. Однако русскоязычные исследователи сайт www.orcid.org используют, судя по всему, лишь для получения ORCID-кодов, например, для регистрации на SCIENCE.INDEX (сайт www.elibrary.ru). Кроме того, такие коды в ряде изданий стали указываться в составе сведений об авторах. В то же время ресурс orcid.org потенциально обеспечивает возможность внесения авторами в их профили достаточно подробной информации – в том числе и для использования как международной справочной БД по ученым (исследователям).

СПИСОК ЛИТЕРАТУРЫ

1. Брумштейн Ю.М., Васьковский Е.Ю. Анализ вебометрических показателей основных сайтов, агрегирующих политематическую научную информацию // Научно-техническая информация. Сер. 2. – 2017. – № 11. – С 16-32.
2. Осипова В.А. НЭБ: История, устройство и новый этап развития // Университетская книга. – 2013. – № 10. – С. 64-66.
3. Соколова М.Е. Российский региональный индекс научного цитирования: новации и проблемы // Интеллектуальный капитал. – 2016. – № 3. – С. 2-6.
4. Шабанова С.М. Научная электронная библиотека: меняются правила, суть остается // Интеллектуальный капитал. – 2016. – № 2 (4). – С. 29-33.
5. Арский Ю.М., Быков В.А. Деятельность ВИНИТИ РАН – базовой организации государств – участников СНГ по межгосударственному обмену научно-технической информацией // Научно-техническая информация. Сер. 1. – 2013. – .№ 5. – С. 23-30; Arskii Yu.M., Bykov V.A. The Activities of VINITI RAS as a Key Organization of the Commonwealth of Independent States for the Exchange of Scientific and Technological Information // Scientific and Technical Information Processing. – 2013. – Vol. 40, № 2. – P. 101-108.
6. Биктимиров М.Р., Глебский В.Л., Долгов Б.В., Поликарпов С.А. Использование информационных технологий и инфраструктур для агрегации научной информации. Опыт Канады, Нидерландов, США / /Моделирование и анализ информационных систем. – 2015. – Т. 22, № 1. – С. 114-126.
7. Гуров А.Н., Гончарова Ю.Г., Бубякин Г.Б. Открытый доступ к научным знаниям: состояние, проблемы, перспективы развития // Научно-техническая информация. Сер. 1. – 2016. – № 4. – С. 10-16; Gurov A.N., Goncharova Yu.G., Bubyakin G.B. Open Access to Scientific Knowledge: Its State, Problems, and Prospects of Development // Scientific and Technical Information Processing. – 2016. – Vol. 43, № 2. – P. 88-94.
8. Фурнисс К., Трифонова А.В. Варианты доступа к ресурсам для исследователей: где искать ресурсы по техническим наукам? // Динамика систем, механизмов и машин. –2014. – № 5. – С. 87-89.
9. Björk, Bo-Christer Open Access to Scientific Publications – An Analysis of the Barriers to Change? // Information Research. – 2004. – Vol.9, № 2. – P. 170. – URL: http://hdl.handle.net/10227/647 (date of access 04.07.2017).
10. Зацман И.М. Электронные библиотеки научных документов в Интернет: структуризация, формальное описание и поиск невербальной информации // Научно-техническая информация. Сер. 2. – 1998. – № 11. – С. 12-18.
11. Ларук О., Гаранович М.В. Оценка доступа к научной информации для академических пользователей в Интернете // Глобальный научный потенциал. – 2014. – № 9 (42). – С. 135-138.
12. Голицына О.Л., Куприянов В.М., Максимов Н.В. Информационные и технологические решения в задачах управления знаниями // Научно-техническая информация. Сер. 1. – 2015. – № 8. – С. 1-12; Golitsina O.L., Kupriyanov V.M., Maksimov N.V. Information and Technological Solutions Applied for Knowledge-Management Tasks // Scientific and Technical Information Processing. – 2015. – Vol. 42, № 3. – P. 150-161.
13. Поляк Ю.Е. Оценивание и ранжирование вебсайтов. Вебометрические рейтинги // Научный редактор и издатель. – 2017. – Т. 2, № 1. – С. 19-29.
14. Лебеденко М.С. Webometrics rank as index of effectiveness of web-site of enterprise // Економічний вісник Національного технічного університету України "Київський політехнічний інститут". – 2014. – № 11. – С. 401-408.
15. Гиляревский Р.С. Публикационная активность как оценка научных достижений // Научно-техническая информация. Сер. 1. – 2014. – № 8. – С. 1-9.
16. Брумштейн Ю.М., Кузьмина А.Б., Яковлева Л.В. Публикационная политика регионального вуза в контексте управления его научным имиджем // Прикаспийский журнал: управление и высокие технологии. – 2013. – № 2 (22). – С. 099-109.
17. Архипов Д.Б., Буляница А.Л., Щербаков А.П. Вебометрический анализ и его использование для изучения тенденций развития аналитического приборостроения // Научное приборостроение. – 2014. – Т. 24, № 2. – С. 52-60.
18. Сюнтюренко О.В., Гиляревский Р.С. Использование методов наукометрии и сопоставительного анализа данных для управления научными исследованиями по тематическим направлениям // Научно-техническая информация. Сер. 2. – 2016. – № 12. – С. 1-12.
19. Лопатина Н.В. Информационная инфраструктура общества: проблемы изучения и управления // Научно-техническая информация. Сер. 1. – 2016. – № 5 – С.1-4; Lopatina N.V. The Information Infrastructure of Society: Problems of Research and Management // Scientific and Technical Information Processing. – 2016. – Vol. 43, № 2. – P. 95-98.
20. Брумштейн Ю.М. Анализ влияния информационно-коммуникационных технологий на структуру создаваемой в России научно-технической информации // Научно-техническая информация. Сер. 1. – 2016. – № 12. – С. 7-17.
21. Галявиева М.С. О становлении понятия «информетрия» (обзор) //Научно-техническая информация. Сер. 1. – 2013. – № 6. – С. 1-10.
22. Печников А.А., Таракановский Н.А. Сравнение вебометрического ранжирования институтов РАН и их ссылочной популярности в русской Википедии // В сб.: Научно-образовательная информационная среда XXI века. Материалы IX Всероссийской научно-практической

конференции / отв. ред. Н.С. Рузанова . – г. Петрозаводск, 2015. – С. 148-150.
23. Брумштейн Ю.М., Васьковский Е.Ю., Куаншкалиев Т.Х. Поиск информации в Интернете: анализ влияющих факторов и моделей поведения пользователей // Известия Волгоградского государственного технического университета. – 2017. – № 1 (196). – С. 50-55.
24. Филозова И.А. Открытые архивы научной информации // Электронный журнал «Системный анализ в науке и образовании». – 2010. – Вып. №1. – С.1-6.
25. Матушанский Г.У., Витоль Е.В. Университетский научный журнал: проблемы вхождения в зарубежные реферативные базы данных // Вестник Казанского государственного энергетического университета. – 2013. – № 3 (18). – С. 170-173.
26. Prabakaean R., Lihitkar Shalini. Websites of Astronomy and Astrophisics Libraries in India and USA: A Webometric Study // A Journal of Library and Information Science. – 2015. – Vol.9, Issue 3. – P.199-205. DOI 10.5958/0975-6922.2015.00028.5
27. Ефременкова В.М., Круковская Н.В. 100-летний юбилей Chemical Abstracts Service: факты и цифры // Научно-техническая информация. Сер. 1. – 2007. – № 12. – С. 24-29.
28. Шалаева Т.З. Национальный информационный ресурс: иерархическая модель и структура // Вестник Полоцкого государственного университета. Серия D: Экономические и юридические науки. – 2014. – № 6. – С. 119-122.
29. Petrov V.V., Onyshchenko O.S., Kryuchyn A.A., Lobuzina K.V., Minina N.M., Zaichenko N.Y. The development of national referencing system // Вісник Національної академії наук України. – 2015. – № 10. – С. 71-74.
30. Сазонець О.М.В., Пінчук О.Л.Д. Долучення україни до європейських наукометричних систем // Вісник Житомирського державного технологічного університету. Серія: Економічні науки. – 2015. – № 3 (73). – С. 128-133.
31. Ли Сюй. Развитие и совершенствование китайской базы данных индекса цитирования общественно-научных статей // Власть. – 2017. – №2. – С.176-181.
32. Писляков В.В. Зачем создавать национальные индексы цитирования // Научные и технические библиотеки. – 2007. – №2. – С. 9 . - Библиотека Государственного университета Высшей школы экономики (ГУ ВШЭ) – URL: https://library.hse.ru/mirror/pubs/share/200171529 (дата обращения 09.07.2017).
33. Сухоручкина И.Н. Базы данных в национальной системе научно-технической информации Японии // Научно-техническая информация. Сер. 1. – 2016. – № 6. – С. 27-35.
34. Negishi M., Sun Y., Shigi K. Citation Database for Japanese Papers: A new bibliometric tool for Japanese academic society // Scientometrics. – 2004. – Vol.60, Iss.3. – P. 333-351.
35. Хачко О.А. Электронная библиотека полных текстов SciELO (scientific electronic library online): функции и поисковые возможности // В сб.: Информационное обеспечение науки. Новые технологии Сборник научных трудов / ред. Н.Е. Каленов. – М., 2011. – С. 273-280.
36. Jahangiri Saeideh, Asnafi Amir Reza , Amir Hussein Rajabzadeh Assarha & Maryam Pakdaman Naeini. Iranian scientific associations' websites in the field of Humanities: A Webometric study // COLLNET Journal of Scientometrics and Information Management. – 2014. –Vol. 8, Issue 2. – P. 273-279. – URL: http://dx.doi.org/10.1080/09737766.2014.1015310.
37. Брумштейн Ю.М., Зайцев В.Ф., Сокольский А.Ф. Анализ диссертационных работ и сопутствующих им материалов в свете законодательства об авторском праве // Интеллектуальная собственность. Авторское право и смежные права. – 2005. – №12. – С.2-13
38. Копанева Е.А. Вебометрические показатели научной периодики Украины // Научные и технические библиотеки. – 2013. – № 5. – С. 75-82.
39. Истомин И.А., Байков А.А. Сравнительные особенности отечественных и зарубежных научных журналов // Международные процессы. – 2015. – Т. 13, № 41. – С. 114-140.
40. Третьяков А.Л., Король А.Н. Использование методов библио- и вебометрии при изучении микропотока библиотековедческих журналов // Библиосфера. – 2015. – № 3. – С. 69-74.

Материал поступил в редакцию 13.10.2017.

Сведения об авторах

БРУМШТЕЙН Юрий Моисеевич – кандидат технических наук, доцент Астраханского государственного университета, доцент
e-mail: brum2003@mail.ru
ORCID http://orcid.org/0000-0002-0016-7295

ВАСЬКОВСКИЙ Евгений Юрьевич – аспирант кафедры информационных технологий Астраханского государственного университета, ведущий программист отдела Internet-технологий Астраханского государственного университета
e-mail: vaskovskiy_evgeniy@mail.ru
ORCID http://orcid.org/0000-0002-4937-3305

ИНФОРМАЦИОННЫЕ СИСТЕМЫ

УДК 004.6.021.032.24

П.В. Голубцов

Понятие информации в контексте задач обработки больших данных

Исследуется подход к трансформации существующих алгоритмов в системах Больших Данных так, чтобы отдельные фрагменты данных обрабатывались независимо и параллельно. Рассматриваются особенности необходимой для этого хорошо организованной промежуточной компактной формы информации, ее естественные алгебраические свойства и приводится иллюстрирующий пример.

Ключевые слова: *системы больших данных, формы представления информации, параллельная обработка, алгебра информации, информационное пространство*

ВВЕДЕНИЕ

В последнее время наблюдается резкий всплеск исследований, связанных с Большими Данными (*Big Data*). Было обнаружено, что большие объемы данных могут содержать ценную информацию, возможность извлечения которой из такого рода данных ранее даже и не предполагалась. Много интересных примеров можно найти в [1]. Можно сказать, что в задачах Больших Данных, как правило, речь идет об извлечении спрятанной информации и представлении ее в форме, пригодной для интерпретации или принятия решений. Такого рода процессы обычно проходят через несколько стадий, в которых информация извлекается из исходных данных, преобразуется, передается, накапливается и, в конце концов, трансформируется к удобному для интерпретации виду.

Отметим, что использование термина «информация», в последнее время заметно возросло, особенно, в контексте анализа данных. Обычно он понимается слишком широко и неформально. Однако, по мнению автора, такая возросшая частота употребления этого термина свидетельствует о возрастающей потребности в более точном и формальном понимании феномена информации. Может ли проблематика Больших Данных приблизить нас к такому пониманию?

Исследования, связанные с системами Больших Данных, нацелены на проблемы обработки больших объемов распределенных данных и имеют, как правило, ярко выраженную практическую и техническую направленность. В то же время, основная масса исследований по теории информации проводится в контексте теории вероятностной и математической статистики и представляет преимущественно теоретический интерес.

Пожалуй, наиболее прикладная часть теории информации, берущая начало в работах Шеннона, связана с передачей сообщений при наличии помех [2, 3]. При этом речь идет не столько о «смысле» информации, сколько о ее количестве. Особое место в математической статистике занимает информация Фишера, описываемая матрицами [4, 5]. Она обеспечивает более детальное отражение понятия «информация» и, в частности обладает важной аддитивной структурой, в рамках которой объединению независимых статистик отвечает сумма их информационных матриц. Несмотря на многочисленные исследования по теории информации, проблема формализации понятия «информация», отражающего именно смысл информации, содержащейся в данных, представляется еще далекой от удовлетворительного решения. В связи с этим, упомянем работы [6–8], в которых вместо определения информации, содержащейся в данных, исследуется информативность систем преобразующих данные. В рамках такого подхода естественным образом возникает алгебраическая структура источников информации и частичный порядок, позволяющий сравнивать их информативность.

На данный момент сферы интересов Больших Данных и различных подходов к понятию информации слабо пересекаются. Однако, как уже было отмечено, проблематика Больших Данных требует более четкого, формального описания самого понятия информации и информационных процессов. Это необходимо для построения эффективных инструментов преобразования информации, опирающихся на математические (например, алгебраические) свойства информации. В связи с этим, по мнению автора, Большие Данные станут в ближайшее время основным двигателем и потребителем (бенефициаром) общей теории информации. В настоящей работе мы попытаемся показать как некоторая формализация понятия информации и ее алгебраические свойства могут следовать просто из рассмотрения задачи в контексте Больших Данных.

Чем же выделяются задачи «Больших Данных» на фоне задач анализа данных? Данные в таких задачах,

как правило, имеют огромный объем, распределены между многочисленными сайтами и постоянно пополняются. В результате даже самый простой анализ Больших Данных сталкивается с серьезными трудностями. Действительно, традиционные подходы к обработке информации предполагают, что данные, предназначенные для обработки, собираются в одном месте, организуются в виде удобных структур (например, матриц), и только тогда соответствующий алгоритм обрабатывает эти структуры и выдает результат анализа. В случае Больших Данных невозможно собрать все данные, необходимые для исследовательского проекта на одном компьютере. Более того, это было бы непрактично, поскольку один компьютер не сможет обработать их в разумные сроки. В результате возникает необходимость в трансформации существующих алгоритмов, приводящих к их «распараллеливанию», или даже разработке новых подходов к обработке данных, которые по самой формулировке проблемы смогли бы обрабатывать отдельные фрагменты данных независимо и параллельно. Соответствующий алгоритм анализа данных должен, параллельно работая на многих компьютерах, извлекать из каждого набора исходных данных некоторую промежуточную компактную «информацию», постепенно объединять и обновлять ее и, наконец, использовать накопленную информацию для получения результата. По прибытии новых фрагментов данных он должен иметь возможность добавлять их к накопленной информации и, в конечном итоге, обновлять результат.

Мы обсудим особенности такой хорошо организованной промежуточной формы информации, выявим ее естественные алгебраические свойства и рассмотрим иллюстративный пример. Мы также увидим, что такая промежуточная форма представления информации в некотором смысле отражает саму суть информации, содержащейся в данных. Это приводит нас к совершенно новому подходу к самому понятию информации.

ОСОБЕННОСТИ ОБРАБОТКИ ИНФОРМАЦИИ В СИСТЕМАХ БОЛЬШИХ ДАННЫХ

Выделим следующие особенности задач обработки информации в системах Больших Данных:

1) как правило, речь идет об огромных объемах данных;

2) такие данные обычно не собраны воедино, а распределены по многочисленным, возможно, достаточно удаленным компьютерам;

3) постоянно могут возникать новые данные, которые необходимо оперативно включать в обработку.

Традиционные методы обработки обычно не учитывают такую специфику и требуют серьезного пересмотра при необходимости их применения в задачах Больших Данных.

Рассмотрим бегло (и, конечно же, предельно упрощенно) стандартный подход к обработке данных. К задачам такого рода относятся задачи оценивания, принятия решений, обучения, классификации… Обычно в задачах с малым фиксированным набором данных (рис. 1) обработка заключается в применении некоторого алгоритма (метода), определяющего обработку, к этому набору данных и получение результата обработки (например, оценки некоторой величины).

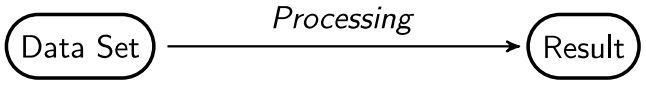


Рис. 1. Стандартный подход к обработке данных

Важным условием здесь является то, что все данные находятся в одном месте и готовы к применению к ним отображения обработки, например, представлены в виде подходящих структур, скажем, матриц. Если же данные распределены по многим различным локациям, для применения обработки их требуется сначала собрать в одном месте, организовать комбинированные данные в виде подходящих структур, и применить к ним алгоритм обработки (рис. 2). Ключевым моментом здесь является необходимость собрать все данные в одном месте. Пунктирными стрелками здесь и далее обозначается передача данных в исходном или частично обработанном виде.

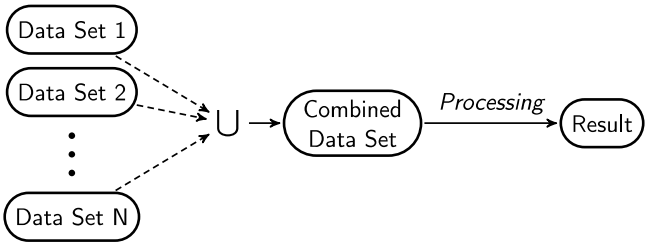


Рис. 2. Стандартный подход к обработке распределенных данных

Недостатки такого подхода к обработке распределенных данных достаточно очевидны:

- передача больших объемов исходных данных создаст чрезмерный трафик;
- хранение полного набора данных в одном месте потребует огромных объемов памяти;
- обработка всех данных на одном компьютере потребует чрезмерных вычислительных и временных ресурсов;
- по мере поступления новых данных, комбинированный их набор будет расти и, как следствие, требовать постоянно возрастающих (потенциально бесконечных) ресурсов для хранения;
- при поступлении новых данных, алгоритм обработки будет необходимо заново применять к постоянно увеличивающемуся объему данных.

ВЫДЕЛЕНИЕ ПРОМЕЖУТОЧНОЙ ИНФОРМАЦИИ В ПРОЦЕССЕ ОБРАБОТКИ

Рассмотрим следующую модификацию процесса обработки, которая позволит преодолеть обозначенные выше недостатки. Предположим, что полный

алгоритм обработки P допускает разбиение на две фазы $P = P_2 \circ P_1$ (рис. 3):

1) P_1 – выделение из исходных данных некоторой промежуточной информации;

2) P_2 – вычисление результата на основании выделенной промежуточной информации.

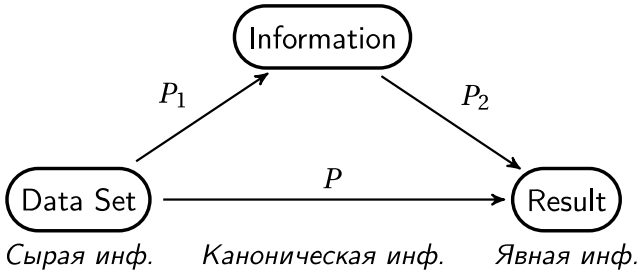


Рис. 3. Разбиение процесса обработки данных на две фазы

Выбор подходящей промежуточной формы представления информации определяется рассматриваемой задачей обработки данных. Будем называть некоторую выбранную форму представления промежуточной информации канонической формой информации или короче – **канонической информацией**.

В определенном смысле узлы диаграммы на рис. 3 отражают представления информации в разных формах:

Data Set – информация в сырой (исходной) форме.

Result – информация в явной (удобной для интерпретации) форме.

Information – информация в промежуточной (удобной для обработки) канонической форме.

Далее более подробно обсудим желательные свойства канонической информации. Отметим, что такая форма представления информации должна быть достаточно полной, т. е. содержать всю необходимую для вычисления результата информацию (в этом и состоит коммутативность диаграммы на рис. 3) и компактной, т. е. иметь минимально возможный размер, в идеале, не зависящий от объема представленных данных.

Рассмотрим, как может быть трансформирована схема обработки данных, если полная обработка может быть разбита на две, указанные фазы (рис. 4).

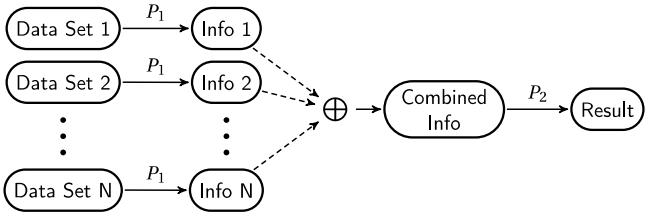


Рис. 4. Модифицированная схема обработки распределенных данных

Такая схема позволяет преодолеть все отмеченные недостатки стандартной схемы обработки распределенных данных:

1) передаются лишь компактные фрагменты выделенной промежуточной информации;

2) хранение комбинированной информации потребует небольших объемов памяти, возможно, таких же, как и объемы, требуемые для хранения отдельных частей промежуточной информации;

3) промежуточная информация выделяется параллельно из каждого отдельного набора данных (фаза P_1). Если основная часть обработки сосредоточена в первой фазе, то вторая фаза P_2, состоящая в построении результата по компактной накопленной информации, не потребует серьезных вычислительных и временных ресурсов;

4) по мере поступления новых данных, потребуется лишь выделять из них промежуточную информацию и «добавлять» ее к накопленной;

5) алгоритм обработки будет необходимо снова применять к компактной информации фиксированного объема.

Отметим, что в приведённых выше рассуждениях мы предполагаем существование операции композиции (сложения) отдельных фрагментов канонической информации. Фактически, мы предполагаем, что на множестве всех фрагментов канонической информации определена операция композиции. При этом, объединению двух наборов данных отвечает композиция соответствующих фрагментов канонической информации (рис. 5).

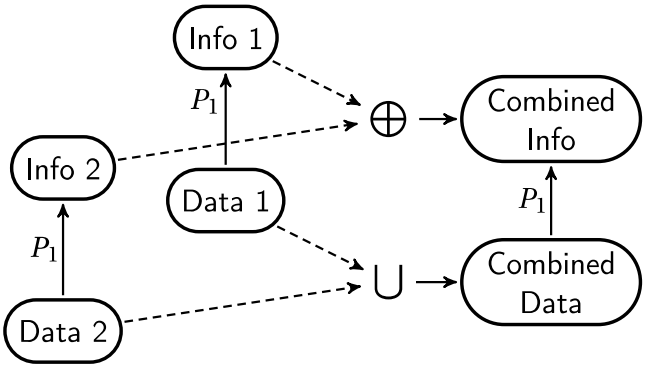


Рис. 5. Соответствие композиции фрагментов канонической информации и объединения наборов исходных данных

Это можно записать как $P_1(D_1) \oplus P_1(D_2) = P_1(D_1 \cup D_2)$, где под $D_1 \cup D_2$ понимается объединение двух наборов данных в один.

Заметим, наконец, что схема обработки распределенных данных, представленная на рис. 4, идеально «вписывается» в архитектуру систем распределенного хранения и анализа данных, таких как, например, Hadoop [9].

ПРИМЕР ФАКТОРИЗАЦИИ АЛГОРИТМА ПУТЕМ ВЫДЕЛЕНИЯ ПРОМЕЖУТОЧНОЙ ИНФОРМАЦИИ

Рассмотрим задачу, которая часто встречается в статистических приложениях и подчеркнем, что мы используем для иллюстрации довольно простую за-

дачу. При этом будем считать, что объемы наборов данных в этой задаче и количество таких наборов крайне велики.

Пусть $(x_1, x_2, ..., x_n)$ – последовательность m-мерных векторов:

$$x_i = \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^m \end{pmatrix}, \ i = 1, ..., n$$

В статистике часто приходится вычислять вектор выборочного среднего:

$$X = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

и выборочную ковариационную матрицу:

$$V = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - X)(x_i - X)^T \ . \qquad (2)$$

Таким образом, исходным набором данных (рис. 6) является последовательность векторов $(x_1, x_2, ..., x_n)$, а требуемым результатом обработки P является пара (X, V), определяемая выражениями (1) и (2), т.е., $P(x_1, ..., x_n) = (X, V)$.

$$(x_1, \ldots, x_n) \xrightarrow{\quad P \quad} (X, V)$$

Рис. 6. Стандартный алгоритм P

Если же исходные данные содержатся в N наборах $(x_1, ..., x_{n_1}), \ldots, (z_1, ..., z_{n_N})$, размещенных на различных компьютерах, то для их обработки с помощью этого алгоритма придется собрать их в одном месте и применить преобразование P (рис. 7).

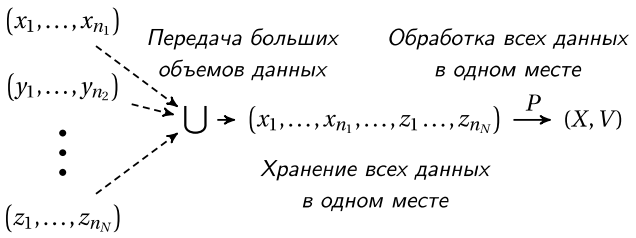


Рис. 7. Объединение исходных данных для обработки

При использовании такой схемы потребуется передавать большие объемы исходных данных, хранить и обрабатывать полный набор $(x_1, ..., x_{n_1}, \ldots, z_1, ..., z_{n_N})$ на одном компьютере. При поступлении нового набора данных придется добавить его к уже имеющемуся полному набору и пересчитать результат (X, V).

Заметим, однако, что вычисление X и V в (1) и (2) можно разбить на два этапа.

Пусть

$$S = \sum_{i=1}^{n} x_i \ , \ T = \sum_{i=1}^{n} x_i x_i^T \qquad (3)$$

соответственно, m-мерный вектор и матрица $m \times m$. Несложно убедиться, что

$$X = \frac{S}{n}, \ V = \frac{nT - SS^T}{n(n-1)} \qquad (4)$$

Второе выражение следует из цепочки равенств:

$$\sum_{i=1}^{n}(x_i - X)(x_i - X)^T = \sum_{i=1}^{n} x_i x_i^T - X\sum_{i=1}^{n} x_i^T -$$

$$-\left(\sum_{i=1}^{n} x_i\right)X^T + nXX^T = T - \frac{1}{n}SS^T.$$

Таким образом, вся информация, достаточная для вычисления X и V, может быть представлена тройкой (n, S, T) и процесс обработки P может быть разбит на две стадии $P = P_2 \circ P_1$ (рис. 8).

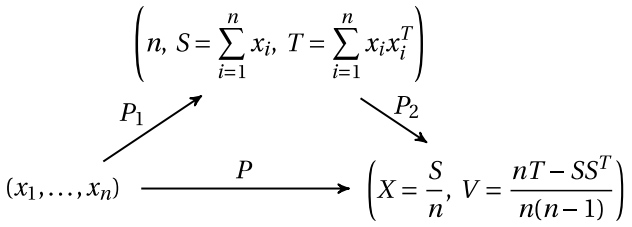


Рис. 8. Разбиение процесса обработки на две стадии выделением канонической информации

Тройка (n, S, T) представляет собой удобную промежуточную форму представления информации об исходных данных в рассматриваемой задаче – каноническую информацию. Отметим, что T является симметричной матрицей $m \times m$ и полностью задается $\frac{m(m+1)}{2}$ числами. Таким образом, тройка (n, S, T) задается $\frac{(m+1)(m+2)}{2}$ числами.

В результате разбиения алгоритма P на две фазы и введения канонической информации, схема обработки распределенных данных, представленная на рис. 7 может быть трансформирована следующим образом. Из каждого отдельного фрагмента данных выделяется каноническая информация (n_j, S_j, T_j), которая впоследствии объединяется и используется для вычисления результата (рис. 9).

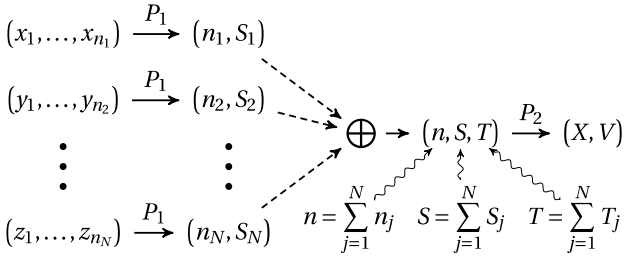


Рис 9. Модифицированная схема обработки распределенных данных

Отметим основные особенности такой модифицированной схемы.

1) Выделение канонической информации $\left(n_j,S_j,T_j\right)$ из *j*-го набора данных (преобразование P_1) может проводиться «на местах» параллельно и независимо. В результате, распределенность исходных данных способствует повышению эффективности обработки за счет распараллеливания.

2) Передаются лишь компактные фрагменты выделенной канонической информации одинакового объема ($\frac{(m+1)(m+2)}{2}$ чисел), не зависящего от объема исходного набора данных.

3) Сложение частей канонической информации максимально упрощено и определяется покомпонентным сложением троек вида $\left(n_j,S_j,T_j\right)$:

$$\begin{aligned}&\left(n_1,S_1,T_1\right)\oplus\left(n_2,S_2,T_2\right)=\\&=\left(n_1+n_2,\,S_1+S_2,\;T_1+T_2\right).\end{aligned}\qquad(5)$$

4) Хранение всей комбинированной канонической информации также требует такого же небольшого объема памяти ($\frac{(m+1)(m+2)}{2}$ чисел).

5) Поскольку основная часть обработки сосредоточена в первой фазе, то вторая фаза P_2, состоящая в построении результата по компактной накопленной информации (4), не зависит от объема исходных данных и не требует серьезных вычислительных и временных ресурсов.

6) По мере поступления новых данных, потребуется лишь выделить из них каноническую информацию и «добавить» ее к накопленной.

7) Алгоритм обработки будет необходимо снова применять к компактной информации фиксированного объема.

Заметим, что тройки вида (*n, S, T*) можно рассматривать как элементы некоторого множества, наделенного дополнительной структурой – канонического **информационного пространства** $\Im$. В данном примере $\Im=\mathbb{N}\times\mathbb{R}^m\times\mathbb{S}_+^m$, где $\mathbb{N}=\{0,1,...\}$ – множество натуральных чисел, $\mathbb{R}^m$ – *m*-мерное пространство столбцов, а $\mathbb{S}_+^m$ – конус симметричных положительно полуопределенных матриц $m\times m$. При этом согласно (5), на пространстве $\Im$ задана операция композиции $\oplus$, определяемая покомпонентно.

ОСНОВНЫЕ СВОЙСТВА ХОРОШО ОРГАНИЗОВАННОЙ ПРОМЕЖУТОЧНОЙ ИНФОРМАЦИИ

В приведенном примере выбор удобной формы представления промежуточной информации позволяет существенно повысить эффективность обработки распределенных данных. Этот пример демонстрирует следующие желательные свойства канонической информации.

Существование для любого исходного набора данных. Любой исходный набор данных должен допускать представление информации в каноническом виде. В частности, минимальный «атомарный» набор данных или даже «пустой» набор должны быть представимы в канонической форме. В нашем примере это условие выполнено. В частности, атомарному набору (*x*), состоящему из единственного столбца *x*, отвечает каноническая информация $P_1\left((x)\right)=\left(1,x,xx^T\right)$, а пустому набору (), отвечает «нулевая» каноническая информация $0=P_1(())=(0,0,0)$.

Заметим, что вычисление окончательного результата может оказаться невозможным для некоторых наборов исходных данных. В частности, согласно (2), для вычисления выборочной ковариационной матрицы необходимо, чтобы исходные данные содержали, как минимум, два столбца. Строго говоря, отображение *P* является не всюду определенным. В то же время, мы требуем, чтобы P_1 было всюду определено.

Полнота (или **достаточность**). Каноническая форма должна содержать всю информацию, имеющуюся в исходных данных, а именно: она должна приводить к тому же результату, что и исходные данные, из которых она получена. Формально это означает, что $P(D) = P_2(P_1(D))$ для всех данных *D* из области определения преобразования *P*.

Единственность представления данных в каноническом виде. Фактически, это свойство означает отсутствие избыточности в канонической информации. Отсюда, в частности, следует, что каноническая информация не должна зависеть от порядка данных в исходном наборе.

Операция композиции $\oplus$. Для обеспечения возможности «объединять» информацию, отвечающую отдельным наборам данных, на каноническом информационном пространстве должна быть определена операция композиции $\oplus$, обладающая следующими свойствами:

a) **a** $\oplus$ **b = b** $\oplus$ **a** – коммутативность. Комбинированная каноническая информация не должна зависеть от порядка поступления данных.

b) **(a** $\oplus$ **b)** $\oplus$ **c = a** $\oplus$ **(b** $\oplus$ **c)** – ассоциативность. Каноническая информация не должна зависеть от порядка комбинирования данных.

c) **a** $\oplus$ **0 = a** – свойство нейтрального элемента. Добавление к некоторой информация пустой информации не меняет эту информацию.

Таким образом, каноническое информационное пространство является коммутативным моноидом.

В приведенном выше примере справедливость этих свойств сразу же следует из покомпонентного определения операции композиции, а именно, композиции двух элементов отвечает сумма их компонент (5).

Компактность. Информация, представленная в канонической форме, должна занимать небольшой (желательно минимальный) объем, по возможности, не зависящий от объема представленных данных.

Эффективность. Представление промежуточной информации в канонической форме должно обеспечивать эффективное выполнение всех стадий обработки данных:

- извлечение канонической информации из исходных данных;
- комбинирование и накопление канонической информации;
- вычисление результата из накопленной канонической информации.

Отметим, что свойства компактности и эффективности носят скорее технический характер, связанный с особенностями реализации соответствующих алгоритмов.

ЗАКЛЮЧЕНИЕ

Отметим, что чисто техническая попытка «распараллелить алгоритм», фактически привела нас к необходимости нахождения специального вида представления информации, обладающего удобными алгебраическими свойствами. В некотором смысле, такое представление отражает саму суть информации, содержащейся в данных. Можно сказать, что сама потребность эффективно манипулировать огромными распределенными массивами данных выдвигает новые требования к осмыслению и формализации понятия информации.

В рассмотренном выше примере выбор канонической формы информации довольно очевиден. В общем случае выбор компактной промежуточной информации может быть не очевиден или даже не возможен. В связи с этим, представляется важным выявление класса задач, в которых возможно выделение достаточно компактной промежуточной информации и нахождение эффективных методов построения подходящих информационных пространств.

В настоящей статье мы старались минимизировать формализм, чтобы акцентировать внимание на содержательной стороне проблемы. Мы наметили основные требования к хорошо организованной промежуточной информации. Это, в свою очередь, поднимает вопрос о выборе в некотором смысле оптимального (или идеального) вида промежуточной информации. Подобная проблематика потребует дальнейшей формализации и исследований.

СПИСОК ЛИТЕРАТУРЫ

1. Mayer-Schönberger V., Cukier K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. – New York: Houghton Mifflin Harcourt, 2013. – 242 p.
2. Яглом А.М, Яглом И.М. Вероятность и информация. – М.: Наука, 1973. – 512 с.
3. Стратонович Р.Л. Теория информации. – М.: Сов. радио, 1975. – 424 с.
4. Барра Ж.-Р. Основные понятия математической статистики. – М.: Мир, 1974. – 280 с.
5. Боровков А.А. Математическая статистика. Оценка параметров, проверка гипотез. – М.: Наука, 1984. – 472 с.
6. Golubtsov P.V. Measurement Systems: Algebraic Properties and Informativity // Pattern Recognition and Image Analysis, – 1991. – Vol. 1, №1. – P. 77–86.
7. Голубцов П.В. Аксиоматическое описание категорий преобразователей информации // Проблемы передачи информации. – 1999. – Т. 35, № 3. – С. 109-127.
8. Golubtsov P.V. Monoidal Kleisli Category as a Background for Information Transformers Theory // Information Processes. – 2002. – Vol. 2, №1. – P. 62–84.
9. White T. Hadoop: The Definitive Guide. – O'Reilly, 2015. – 754 p.

Материал поступил в редакцию 21.10.17.

Сведения об авторе

ГОЛУБЦОВ Петр Викторович – доктор физико-математических наук, профессор Московского государственного университета им. М.В. Ломоносова; профессор Национального исследовательского университета «Высшая школа экономики», Москва.
e-mail: golubtsov@physics.msu.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 004.021 : 81’322.2

Д.О. Жуков, С.А. Головин, Е.Г. Андрианова, В.К. Раев, Б.М. Позднеев

Алгоритм кластеризации текстов на основе разделения терминов на области с заданным критерием соответствия*

Описаны модель и алгоритм кластеризации текстовых документов, основанные на том, что объем смыслового кластера документов и положение его центра (центроид) не должны изменяться в процессе добавления в него новых векторов. Критерием соотнесения является заданная постоянная метрика точности, на которую должно отличаться расстояние между вектором и центроидом оболочки кластера для вхождения вектора в кластер. При построении модели предлагается не находить сразу сходство векторов между собой, а разделить всё информационное пространство размерности R^M на отдельные плотно упакованные области, которые не изменяются в процессе кластеризации. Положение центра такой области определяет смысловое значение всех векторов, попадающих в заданную от него окрестность, и не должно изменяться в процессе кластеризации.

Ключевые слова: *кластеризация текстов, алгоритм кластеризации, информационное пространство, критерий соответствия, пространство терминов*

ВВЕДЕНИЕ

Построение* моделей кластеризации естественно-языковых текстовых документов для решения задач информационного поиска должно основываться на методологии отбора их существенных признаков. Не существует модели кластеризации, которая одновременно имела бы очень небольшое время выполнения, обладала бы абсолютной точностью, и обрабатывала бы большие объемы неструктурированных или слабоструктурированных данных. Кроме того, следует отметить, что набор признаков, необходимых для анализа скрытых закономерностей в больших объемах текстовых данных, неизвестен, поэтому при построении модели кластеризации и информационного поиска необходимо основываться на минимальном числе допущений при максимальном сохранении разнообразных признаков, используемых для кластеризации. Однако следует учесть, что время обработки текстов при использовании применяемой модели кластеризации должно быть минимальным, а реализация необходимых вычислений не требовать значительных аппаратных ресурсов.

ОБЗОР СОВРЕМЕННЫХ МОДЕЛЕЙ И АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ ТЕКСТОВ

Нет необходимости подробно останавливаться на рассмотрении ставших уже традиционными моделях кластеризации текстовых документов [1–3]. Тем не менее, рассмотрим ряд работ, появившихся только за последний год в этой области, основывающихся на использовании новых алгоритмов. Например, в [4] для кластеризации текстов рассматривается основанный на методе стохастической оптимизации алгоритм «искусственной пчелиной колонии» (роевой алгоритм). Авторами [4] предлагается гибридный алгоритм, основанный на модифицированных «роевом» и *K-Means* алгоритмах. Решения по кластеризации текстов, создаваемые модифицированными алгоритмом «искусственной пчелиной колонии», рассматриваются как начальное приближение для алгоритма *K-Means*. Производительность предлагаемого алгоритма оценивалась на шести стандартных наборах данных и сравнивалась с алгоритмами *K-Means* и исходным не модифицированным «роевым» алгоритмом. Экспе-

* Работа выполнена за счет финансирования Министерством образования и науки Российской Федерации конкурсной части государственных заданий. Проект № 28.2635.2017/ПЧ «Разработка моделей стохастической самоорганизации слабоструктурированной информации и реализации памяти при прогнозировании новостных событий на основе массивов естественно-языковых текстов».

риментальные результаты подтверждают превосходство по ряду показателей (например, *F*-мера и ряд других) модифицированного алгоритма «искусственной пчелиной колонии» для кластеризации текстовых данных. Кроме того, авторы [4] показывают, что предложенный алгоритм способен избегать локальных оптимумов и находить лучшие значения целевых функций с гораздо более низким стандартным отклонением по сравнению с другими двумя алгоритмами.

В работе [5] предлагаются три алгоритма кластеризации текстовых документов: с выбором функций, с функцией весовой схемы и с функцией динамического уменьшения размеров. В процессе кластеризации текстовые документы разделяются на несколько когерентных кластеров в соответствии с тщательно подобранными информативными особенностями, и используется правильная функция оценки, с зависящей от времени частотой. Информативные функции, которые определяют набор признаков для кластеризации, в каждом документе выбираются с использованием различных методов их выбора. Генетический алгоритм (*GA*), алгоритм гармонического поиска (*HS*) и алгоритм оптимизации (*PSO*) являются наиболее успешными методами выбора признаков, установленными с использованием новой схемы взвешивания, а именно: весовой коэффициент длины (*LFW*), который зависит от частоты появления в других документах признаков, по которым осуществляется кластеризация. Также в статье предлагается новый метод уменьшения динамических размеров (*DDR*) для уменьшения числа функций, используемых в кластеризации, и, таким образом, для повышения эффективности алгоритмов. В [5] оценивались семь тестовых наборов текстовых данных разного размера и сложности. Анализ результатов показывает, что оптимизация с весом длины и динамическим уменьшением дает оптимальные результаты почти для всех тестируемых наборов данных.

В [6] представлен высокомасштабируемый быстрый и эффективный алгоритм расширенной кластеризации, основанный на использовании *n*-грамм для уменьшения высокой размерности и получения высококачественных кластеров тестовых документов. Также в статье приведён сравнительный анализ, показывающий, что для образцов текстовых наборов данных с удалением стоп-слов, предлагаемый алгоритм работает лучше, чем без удаления стоп-слов.

В работе [7] подчеркивается важность и эффективность совместной кластеризации, особенно при рассмотрении разреженных высокоразмерных данных, а также представлена новая генеративная модель *Sparse Poisson Latent Block Model* (*SPLBM*), основанная на распределении Пуассона для матриц документов-термов. Авторы утверждают, что модель *SPLBM* имеет два больших преимущества: во-первых, это строгая статистическая модель; во-вторых, алгоритм был разработан с нуля, чтобы справиться с проблемами разреженности данных. В результате, помимо поиска как однородных блоков, так и других доступных алгоритмов отфильтровываются однородные, но «шумные» из-за разреженности данные. Эксперименты по различным наборам данных различного размера и структуры показывают, что алгоритм, основанный на *SPLBM*, явно превосходит современные алгоритмы. В частности, представленный на основе *SPLBM* алгоритм преуспевает в получении естественной кластерной структуры сложных несбалансированных наборов данных, которые другие известные алгоритмы эффективно обрабатывать не могут.

РАЗРАБОТКА МОДЕЛИ КЛАСТЕРИЗАЦИИ

Существующие модели кластеризации текстов, вне зависимости от того, являются они лексическими или семантическими, основываются на том, что, исходя из задач информационного поиска для решения которых они предназначены, производится отбор значимых признаков. Далее всё множество документов векторизуется (с использованием терминов или ассоциативно-семантических классов) и в соответствии со значимыми признаками и заданной метрикой точности (или например, заданным числом возможных кластеров и т.д.) кластеризуется по смысловым группам. Для каждого кластера определяются его центр и вектор, задающий положение центра (центроид). При появлении новых документов, их вектора сравниваются по заданной метрике с центроидом, и по результатам сравнения добавляются в тот или иной кластер, или создается новый кластер. При таком подходе добавление каждого нового вектора изменяет положение центра кластера и каждого из ранее вошедших в него векторов (по отношению к нему). В конечном счете это может привести к размыванию первоначального смыслового значения и необходимости перекластеризации всего множества векторов, или использованию посткластеризации при обработке блоков документов. В существующих моделях кластеры могут динамически изменяться не только по числу векторов, но и по объему.

Предлагаемая нами модель кластеризации основана на других принципах, в соответствии с которыми объем смыслового кластера и положение его центра (центроид) не должны изменяться в процессе добавления в него новых векторов. Критерием соотнесения является заданная постоянная метрика точности, на которую для вхождения вектора в кластер должно отличаться расстояние между вектором и центроидом оболочки кластера.

Главной задачей при создании такой модели кластеризации является разработка методики разделения всего векторного пространства текстовых документов на оболочки смысловых кластеров, по возможности, с минимальными ограничениями на признаки отбора (чтобы не потерять скрытые закономерности, поведение которых имеет характер шума). Это необходимо для создания эффективной модели анализа неструктурированных или слабоструктурированных текстовых данных.

Иными словами, при построении модели необходимо разделить всё информационное пространство размерности R^M на отдельные плотно упакованные области, которые не изменяются в процессе кластеризации. Положение центра определяет смысловое значение всех векторов, попадающих в данную об-

ласть, и в процессе кластеризации не должно изменяться (даже незначительно). Разделение должно удовлетворять следующим требованиям: если длина вектора, задающего положение центра какой-либо области, равна некоторой величине Y, то все вектора текстовых документов, положение которых отличается от положения центра не более чем на величину ξY (где ξ заданная точность кластеризации, или смыслового соответствия), принадлежат данной смысловой области.

При построении модели необходимо решить две основные задачи – разработать: 1) методику разделения информационного пространства на смысловые области; 2) методику отнесения вектора к конкретной области.

Перечислим основные требования к модели кластеризации:

1. Исходные данные модели: словарь значимых терминов (содержащий некоторое число объектов), коллекция документов (содержащая некоторое произвольное число объектов, каждый из которых содержит некоторое произвольное число значимых терминов), заданная точность отнесения документа к определенной смысловой группе. Метрикой точности является заданная величина отклонения данного вектора от центра кластера.

2. Положение центра и объем каждого кластера не должны изменяться при добавлении новых векторов. Изменяться может только число векторов внутри кластера, что не приводит к изменению первоначально заданного смыслового значения, а, значит, появлению или исчезновению анализируемой скрытой закономерности.

3. Построение модели не должно требовать сверхбольших аппаратных ресурсов и иметь, в зависимости от объема данных, приемлемое время их обработки.

ПРЕДЛАГАЕМАЯ МОДЕЛЬ КЛАСТЕРИЗАЦИИ

Разделение векторов по классам

Для кластеризации коллекции текстовых документов по смысловым группам (определение однородных классов в произвольной проблемной области) после создания матрицы термин-документ, мы предлагаем следующую модель. Воспользовавшись классическим определением расстояния между двумя точками в пространстве любой размерности, определяем длины всех N-векторов коллекции текстовых документов.

Далее выбираем вектор с максимальной длиной $l_k^{(\max)}$, и вектор с минимальной длиной $l_p^{(\min)}$. Величина $l_p^{(\min)}$ определяет радиус гиперсферы в пространстве $\mathbf{R^M}$, внутри которой не будет находиться ни один вектор, так как их длины будут больше $l_p^{(\min)}$. Все вектора будут лежать внутри гипершарового слоя r, для которого будет выполняться условие $l_p^{(\min)} \le r \le l_k^{(\max)}$.

Когда длины векторов имеют разные значения, для кластеризации документов можно сформулировать следующую задачу: на сколько областей можно разделить гипершаровой слой r таким образом, чтобы выполнялись условия:

- документы, попавшие в один кластер, имели бы примерно одинаковое смысловое значение;
- при очень большом числе документов (например, N→∞) все кластеры имели различимую границу, т.е. не пересекались бы между собой и не образовывали бы единый кластер.

В качестве критерия «близости» можно выбрать условие, при котором расстояние от центроида кластера до любого другого вектора (точки соответствующей его концу), входящего в данный кластер (в том числе и вектора l_p), не превышала бы некоторую заданную величину $\boldsymbol{\xi}$ от величины вектора l_c. Подобный подход позволяет предположить, что при большом и плотном заполнении кластера векторами, его форма должна будет стремиться к гипершару радиуса ξl_c в пространстве размерности $\mathbf{R^M}$.

Используя описанный подход, разделим гипершаровой слой на «субслои» 1,2,3,4,…..k, от $l_p^{(\min)}$ до $l_k^{(\max)}$. При этом каждый следующий субслой будет несколько больше по ширине, чем предыдущий. В общем случае получаем:

$$Y_n^2\left(1-\xi^2\right)-\left(1-\xi^2\right)Y_{n-1}^2-4\xi^2Y_{n-1}Y_n-$$
$$-4\xi Y_{n-1}\sqrt{Y_{n-1}Y_n\left(1-\xi^2\right)}=0,$$

где n=2,3,4,…..k, а для любого Y_n должно выполняться условие $Y_n \le l_k^{\max}/(1+\xi)$.

Если $Y_{n+1} > l_k^{\max}/(1+\xi)$, а $Y_n \le l_k^{\max}/(1+\xi)$, то ширина последнего слоя (с номером n+1) определяется условием: $\{l_k^{\max} - Y_n(1+\xi)\}$.

Следует отметить, что упаковка оболочек кластеров может содержать разреженные области между слоями, может и не являться самой плотной. Наличие разреженных областей снижает общую плотность упаковки оболочек кластеров в гипершаре, но, вместе с тем, не снижает точности соотнесения векторов между субслоями, поскольку в разреженных областях выполняется условие $|BC| > \xi Y_1 + \xi Y_2 = \xi(Y_1 + Y_2)$ и, кроме того, не сказывается на дальнейшем распределении векторов между оболочками кластеров внутри своего субслоя. Таким образом, плотное расположение дает нижнюю границу обеспечения точности распределения текстов по смысловым группам.

Описанная методика позволяет дать определение класса, согласно которому понятие класса совпадает с понятием субслоя.

Покрытие субслоя оболочками кластеров

После распределения векторов по отдельным субслоям (классам), рассмотрим задачу отнесения вектора к конкретному кластеру внутри субслоя. Желательно, чтобы покрытие слоя оболочками кластеров было равномерным. Учитывая, что часть векторов слоя может попадать в пространство вне оболочек, то

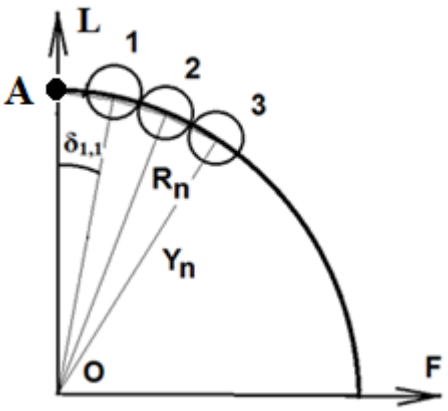
для того, чтобы число таких не соотнесенных векторов было минимальным, необходимо стремиться, чтобы упаковка оболочек внутри слоя была максимально плотной. Далее необходимо определить координаты узлов сети (центров оболочек кластеров). Затем для каждого вектора слоя необходимо попарно вычислить метрики его расстояний от узлов сети. Величина каждой метрики должна быть меньше или равна величине ξY_n. Векторы, удовлетворяющие этой метрике, относятся к выбранному кластеру.

Для полностью регулярной упаковки кластеров на поверхности слоя необходимо задать сеть, состоящую из цепочек кластеров, покрывающих поверхность каждого субслоя. Расстояние между осевыми линиями цепочек (ширина цепочки) и центрами кластеров внутри цепочки при самой плотной упаковке должны удовлетворять условию $R_n = 2\xi Y_n$.

Для соотнесения векторов субслоя с определенным кластером, необходимо найти координаты центров оболочек кластеров в пространстве $\mathbf{R^M}$. Для удобства вычислений перейдем в сферическую систему координат. В *n*-мерном пространстве сферические координаты центров кластеров определяются следующим образом:

$$\begin{cases} x_1 = Y * \cos\theta_1 \\ \vdots \\ x_1 = Y * \cos\theta_1 * \prod_{i-1}^{M-1} \sin\theta_{i-1}, \text{ при } i = 2,3,4,...M \\ \vdots \\ x_M = Y * \sin\theta_1 * \sin\theta_2 * \sin\theta_3 \quad ...\sin\theta_{M-1} \end{cases} \qquad (1)$$

Выберем субслой с номером *n*, и определим $\delta_{1,1}$ – шаг по углу θ_1 (θ_1 – угол между осью L (например, L номер 1) и вектором, определяющим положение центра кластера) для цепочки кластеров с номером 1. Поясним нахождение величины $\delta_{1,1}$. На рисунке показаны ось L=1 и произвольно выбранная ось с номером F, индексы 1,2, 3 и т.д. одновременно обозначают цепочки кластеров и их проекции на плоскость OLF.



Проекции цепочек кластеров на дугу между произвольными осями L и F.

Выберем значение всех углов θ_i равными нулю, и получим следующий набор координат:

$$\begin{cases} x_{1,0} = Y_n \\ \vdots \\ x_{i,0} = 0 \\ \vdots \\ x_{M,0} = 0 \end{cases}$$

Точка с данными координатами (обозначим её A) лежит на оси L=1. Далее рассмотрим ближайшую к этой точке цепочку кластеров, у которой угол между осью L=1 и вектором, задающим центр первой осевой линии равен $\theta_1=\delta_{1,1}$ (см. рисунок), а угловые координаты относительно остальных осей равны 0.

$$\begin{cases} x_{1,1} = Y_n \cos\delta_{1,1} \\ x_{2,1} = Y_n \sin\delta_{1,1} \\ \cdot \\ x_{i,1} = 0 \\ \vdots \\ x_{M,1} = 0 \end{cases} \qquad (2)$$

Далее запишем уравнение для метрики, определяющей расстояние между вектором, задающим осевую линию первой цепочки кластеров, и точкой A, лежащей на оси *L=1*:

$$\left(x_{1,1} - x_{1,0}\right)^2 + \left(x_{2,1} - x_{2,0}\right)^2 + ... +$$
$$+\left(x_{i,1} - x_{i,0}\right)^2 + ... + \left(x_{j,1} - x_{j,0}\right)^2 = \left\{2\xi Y_n\right\}^2$$

$$Y_n^2 \sin^2\delta_{1,1} + \left(Y_n\cos\delta_{1,1} - Y_n\right)^2 = \left\{2\xi Y_n\right\}^2$$

$$2Y_n^2\left(1-\cos\delta_{1,1}\right) = \left\{2\xi Y_n\right\}^2$$

$$\cos\delta_{1,1} = 1 - \frac{\left\{2\xi Y_n\right\}^2}{2Y_n^2} = 1 - 2\xi^2 \qquad (3)$$

Угол θ_1 будет изменяться с дискретным шагом $\delta_{1,1}$ от 0 до $\pi/2$. Общее число цепочек кластеров будет равно $K_M=\pi/2\delta_{1,1}$ с округлением до минимального целого значения.

Определим $\delta_{1,2}$ – шаг по углу $\theta_{2,1}$ ($\theta_{2,1}$ – угол для оси L=2) для цепочки кластеров с номером 1. Выберем значения всех углов θ_i, (кроме $\theta_1=\delta_{1,1}$ и $\theta_{2,1}=\delta_{1,2}$) равными нулю, и, используя описанный выше подход, получим:

$$\cos\delta_{1,2} = 1 - \frac{\left\{2\xi Y_n\right\}^2}{2Y_n^2\sin^2\delta_{1,1}} = 1 - \frac{2\xi^2}{\sin^2\delta_{1,1}} \qquad (4)$$

Угол $\theta_{2,1}$ будет изменяться с дискретным шагом $\delta_{1,2}$ от 0 до $\pi/2$. Общее число кластеров будет равно $K_{M\text{-}1,1}=\pi/2\delta_{1,2}$ с округлением до минимального целого значения.

Далее необходимо взять угол $\theta_1=2\delta_{1,1}$, и определить $\delta_{2,2}$ – шаг по углу $\theta_{2,2}$ ($\theta_{2,2}$ – угол для оси 2) для цепочки кластеров с номером 2. Аналогично проделанному ранее получим:

$$\cos\delta_{2,2}=1-\frac{\{2\xi Y_n\}^2}{2Y_n^2\sin^2 2\delta_{1,1}}=1-\frac{2\xi^2}{\sin^2 2\delta_{1,1}} \qquad (5)$$

Угол $\theta_{2,2}$ будет изменяться с дискретным шагом $\delta_{2,2}$ от 0 до $\pi/2$. Общее число кластеров будет равно $K_{M\text{-}1,2}=\pi/2\delta_{2,2}$ с округлением до минимального целого значения.

Аналогичным образом берем угол $\theta_1=3\delta_{1,1}$, и определяем $\delta_{2,3}$ – шаг по углу $\theta_{2,2}$ ($\theta_{2,3}$ – угол для оси L=2) для цепочки кластеров с номером 3.

$$\cos\delta_{2,3}=1-\frac{\{2\xi Y_n\}^2}{2Y_n^2\sin^2 3\delta_{1,1}}=1-\frac{2\xi^2}{\sin^2 3\delta_{1,1}} \qquad (6)$$

Далее получаем, что для оси L=1 шаг по углу составляет $\delta_{1,1}$, а число цепочек $K_1=\pi/2\delta_{1,1}$ (с округлением до наименьшего целого). Для первой цепочки кластеров относительно оси L=2 угловой шаг составит $\delta_{1,2}$, для второй цепочки кластеров – $\delta_{2,2}$, для третьей цепочки кластеров – $\delta_{2,3}$, и т.д. до цепочки K_M.

$$\cos\delta_{1,1}=1-2\xi^2;\ \cos\delta_{1,2}=1-\frac{2\xi^2}{\sin^2\delta_{1,1}};$$

$$\cos\delta_{2,2}=1-\frac{2\xi^2}{\sin^2 2\delta_{1,1}};\ \cos\delta_{3,2}=1-\frac{2\xi^2}{\sin^2 3\delta_{1,1}}$$

$$\cos\delta_{K_1,2}=1-\frac{2\xi^2}{\sin^2 K_M\delta_{1,1}}$$

Далее переходим к оси L=3. Определим шаг по углу θ_3 (θ_3 – угол для оси L=3). Здесь должна быть реализована следующая процедура. Выбираем угол $\delta_{1,1}$, затем берем угол $\delta_{1,2}$ и определяем углы $\delta_{1,3}$, $\delta_{2,3}$, и т.д. как это было проделано ранее для углов $\delta_{1,2}$, $\delta_{2,2}$ и т.д. при условии, что углы $\theta_{3,1}$, $\theta_{3,2}$, $\theta_{3,3}$, и т.д. дискретно изменяются с шагом $\delta_{1,3}$, $\delta_{2,3}$, и т.д. от 0 до $\pi/2$. Далее выбираем угол $\delta_{1,1}$ и угол $\delta_{2,2}$ и проделываем описанную выше процедуру. Затем выбираем угол $\delta_{1,1}$ и угол $\delta_{3,2}$ и проделываем описанную выше процедуру до $\delta_{K_1,2}$, пока не будет достигнут угол $\pi/2$. Затем берем угол $2\delta_{1,1}$ и повторяем все вновь, потом берем $2\delta_{1,1}$ и т.д. до тех пор, пока θ_1 не достигнет $\pi/2$.

$$\cos\delta_{1,3}=1-\frac{\{2\xi Y_n\}^2}{2Y_n^2\sin^2\delta_{1,1}\sin^2\delta_{1,2}}=1-\frac{2\xi^2}{\sin^2\delta_{1,1}\sin^2\delta_{1,2}} \qquad (7)$$

Используя описанный подход, рекурсивно можно найти угловые координаты всех оболочек кластеров для данного субслоя. Очевидно, что число оболочек во всем субслое будет настолько велико, что нахождение координат их центров является технически неразрешимой задачей. Однако отметим, что её решение для кластеризации векторов внутри слоя этого и не требует.

Распределение векторов внутри классов (кластеризация)

Рассмотрим текстовый документ *j*, состоящий из w_i значимых терминов. Соответственно из всего множества R^M, отличными от 0 будут только w_i координат, для которых можно записать систему уравнений (1), в ней останутся только строки, соответствующие отличным от нуля x_i. Чтобы рассматриваемый вектор принадлежал определенному кластеру, необходимо выполнение условия:

$$\left(x_{1,c}-x_{1,v}\right)^2+\left(x_{2,c}-x_{2,v}\right)^2+...+$$
$$+\left(x_{i,c}-x_{i,v}\right)^2+...+\left(x_{M,c}-x_{M,v}\right)^2\le\left\{\xi Y_n\right\}^2,$$

где $x_{i,c}$ – координаты центра кластера, а $x_{i,v}$ – соответствующие координаты вектора.

Используя для угловых координат полученные ранее рекурсивные соотношения и систему уравнений (1), можно восстановить по ним соответствующие координаты $x_{i,c}$ центров кластеров, к которым принадлежат вектора. Критерием точности нахождения координат является соотношение: $\left|x_{i,v}-x_{i,c}\right|\le\Delta$, где Δ – величина ошибки (по сути, каждая из $x_{i,c}$ подгоняется к $x_{i,v}$, таким образом, чтобы рассчитываемая с использованием формул (4) величина отличалась от соответствующей координаты вектора не более, чем на Δ). Величину Δ можно оценить используя уравнение:

$$(\Delta)^2+(\Delta)^2+...(\Delta)^2+...+(\Delta)^2\le\left\{\xi Y_n\right\}^2;$$
$$w_i(\Delta)^2\le\left\{\xi Y_n\right\}^2;\ \Delta\le\frac{\xi Y_n}{\sqrt{w_i}}$$

Далее зная, к какому кластеру принадлежит данный вектор, можно кластеризовать все документы по смысловым группам с заданной ошибкой ξ.

Пусть документ имеет значимый термин, соответствующий в рассматриваемом пространстве координате X_1. Тогда, используя условие:

$$x_{i,v}-\frac{\xi Y_n}{\sqrt{w_i}}\le x_{i,c}\le x_{i,v}+\frac{\xi Y_n}{\sqrt{w_i}}$$

подбираем θ_1, используя для этого его (θ_1) наборы угловых координат δ. Если $X_1=0$, то $\cos\theta_1=0$, а $\sin\theta_1=1$. При $\cos\theta_1=0$ координата кластера по оси 1 выбирается из множества:

$$\theta_1=\delta_{1,1},2\delta_{1,1},3\delta_{1,1},...,n_{\theta_1}\delta_{1,1},$$

где $\delta_{1,1}$ – шаг по углу θ_1 (θ_1 – угол по оси 1). Отметим, что необходимо найти $n_{\max}\delta_{1,1}\approx\pi/2$ или $n_{\max}\approx\pi/2\delta_{1,1}$ с округлением до минимального целого числа. Если $\cos\theta_1\ne0$, то θ_1 сразу выбирается из $\theta_1=\delta_{1,1},2\delta_{1,1},3\delta_{1,1},...,n_{\theta_1}\delta_{1,1}$.

Переходим к определению θ_2. Зная $\cos\theta_1$ (подобранный или найденный из значения координаты X_1), находим величину $\sin\theta_1$, а затем величину $\cos\theta_2 = \frac{x_2}{Y\sin\theta_1}$. Далее, используя условие:

$$x_{i,v} - \frac{\xi Y_n}{\sqrt{w_i}} \le x_{i,c} \le x_{i,v} + \frac{\xi Y_n}{\sqrt{w_i}},$$

подбираем θ_2, используя для этого его наборы угловых координат δ для θ_2, но уже с известным $\theta_1 = n_{\theta_1}\delta_{1,1}$ (с ограничением для θ_1), что существенно уменьшает число вариантов для угловых наборов θ_2. Отметим, что предварительного расчета всех угловых наборов для всех углов θ не требуется, так как все значения можно рассчитывать при отнесении вектора к кластеру. Используем для этого описанную ранее рекурсивную методику и полученные с её помощью математические выражения в общем виде для любых θ и δ, что существенно уменьшает объем вычислений и увеличивает скорость кластеризации.

Далее аналогичным образом определяем все остальные координаты.

$$\theta_1^{(vector)} = \arccos\left|\frac{x_1^{(vector)} \pm \frac{\xi * Y}{\sqrt{w_i}}}{Y}\right|,$$

при j=1, $x_1^{(vector)}$ – координата вектора по оси 1;

$\delta_1 = \arccos\left(1 - 2*\xi^2\right)$, при j=1;

$n_{\theta_1} = \min\left\{\frac{\theta_j^{(vector)}}{\delta_1}\right\}_{integer}$, (выбирается целое n_{θ_1}, для которого наименьшая погрешность при округлении);

$\theta_1^{(cluster)} = n_{\theta_1} * \delta_1 \le \frac{\pi}{2}$, где $\theta_1^{(cluster)}$ – угловая координата по оси 1 центра кластера, которому принадлежит данный вектор;

$$\theta_1^{(vector)} = \arccos\left|\frac{x_j^{(vector)} \pm \frac{\xi * Y}{\sqrt{w_i}}}{Y * \prod_{j-1}^{M-1}\sin\left\{n_{\theta_{j-1}} * \delta_{j-1}\right\}}\right|,$$

при $j = 2,3,4,...M$, $x_j^{(vector)}$ – координата данного вектора по оси j;

$$\delta_j = \arccos\left|1 - \frac{2*\xi^2}{\prod_{j-1}^{M-1}\sin^2\left\{n_{\theta_{j-1}} * \delta_{j-1}\right\}}\right|,$$

при $j = 2,3,4,...M$;

$n_{\theta_1} = \min\left\{\frac{\theta_j^{(vector)}}{\delta_j}\right\}_{integer}$, (выбирается целое n_{θ_j}, для которого наименьшая погрешность при округлении);

$\theta_1^{(cluster)} = n_{\theta_j} * \delta_j \le \frac{\pi}{2}$, где $\theta_1^{(cluster)}$ – угловые координаты по осям j=1,2,3, М-центра кластера, которому принадлежит рассматриваемый вектор.

Общий алгоритм кластеризации

1. Обрабатываем всю коллекцию текстовых документов, и создаем матрицу документ – термин.
2. Определяем длины всех N – векторов коллекции документов. Выбираем вектор с максимальной длиной $l_k^{(\max)}$, и вектор с минимальной длиной $l_p^{(\min)}$. До момента выбора минимального и максимального векторов эта операция может осуществляться параллельно.
3. Задаем величину критерия ξ смыслового совпадения документов. И, решая уравнение

$$Y_n^2\left(1-\xi^2\right) - \left(1-\xi^2\right)Y_{n-1}^2 - 4\xi^2 Y_{n-1}Y_n - $$
$$-4\xi Y_{n-1}\sqrt{Y_{n-1}Y_n\left(1-\xi^2\right)} = 0,$$

(где n=2,3,4,…..k), определяем радиусы (Y_n) слоев оболочек кластеров в информационном пространстве (учитывая, что $Y_1 \le l_p^{(\min)} / (1-\xi)$). Для любого Y_n должно выполняться условие $Y_n \le l_k^{(\max)} / (1+\xi)$, n – номер соответствующего слоя. Если $Y_{n+1} > l_k^{\max}/(1+\xi)$, а $Y_n \le l_k^{\max}/(1+\xi)$, то ширина последнего слоя (с номером n+1) определяется условием: $\{l_k^{\max} - Y_n(1+\xi)\}$.

4. Сортируем все вектора по слоям (классам), используя следующее условие:

$$Y_n(1-\xi) \le l_j \le Y_n(1+\xi).$$

5. Используя описанную ранее рекурсивную методику определения координат оболочек кластеров при их покрытии субслоя и распределения векторов внутри классов, определяем угловые координаты кластера (наборы δ для углов θ), к которому относится данный вектор. Такая операция может осуществляться параллельно не только по обработки каждого субслоя, но и по обработке каждого из векторов, отнесенного к тому или иному субслою.
6. Зная координаты центров кластеров, к которым относятся вектора, проводим кластеризацию документов по смысловым группам.

СПИСОК ЛИТЕРАТУРЫ

1. Часовских А. Обзор алгоритмов кластеризации данных // Сайт Хабрахабр. – URL: http://habrahabr.ru/post/101338/ (дата обращения: 12.06.2017).
2. Скуратов А.К., Кошкин Д.Е. Сравнение 12 алгоритмов кластеризации данных применительно к задаче кластеризации тов А.К., Кошки

текстов // Информационные технологии. – 2014. – № 7(215). – С. 6-22.
3. Ершов К.С., Романова Т.Н. Анализ и классификация алгоритмов кластеризации // Новые информационные технологии в автоматизированных системах. – 2016. – № 19. – С. 274-279.
4. Kumar A., Kumar D., Jarial S.K. A novel hybrid K-means and artificial bee colony algorithm approach for data clustering // Decision Science Letters. – 2018. – Vol. 7, Issue 1. – P. 65-76.
5. Abualigah L.M., Khader A.T., Al-Betar M.A., Alomari O.A. Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering // Expert Systems with Applications. – 2017. – № 84. – P. 24-36.
6. Kanimozhi K.V., Venkatesan M. A novel map-reduce based augmented clustering algorithm for big text datasets // Advances in Intelligent Systems and Computing. – 2018. – Vol. 542. – P. 427-436.
7. Ailem M., Role F., Nadif M. Sparse Poisson Latent Block Model for Document Clustering // IEEE Transactions on Knowledge and Data Engineering. – 2017. – №29 (7). – P. 1563-1576.

Материал поступил в редакцию 19.09.17.

Сведения об авторах

ЖУКОВ Дмитрий Олегович – доктор технических наук, профессор, заместитель по научной работе директора Института комплексной безопасности и специального приборостроения, профессор кафедры КБ-8 «Информационное противоборство» Института комплексной безопасности и специального приборостроения федерального государственного бюджетного образовательного учреждения высшего образования «Московский технологический университет»
e-mail: zhukovdm@yandex.ru

ГОЛОВИН Сергей Анатольевич – доктор технических наук, профессор, заведующий кафедрой математического обеспечения и стандартизации информационных технологий Института информационных технологий федерального государственного бюджетного образовательного учреждения высшего образования «Московский технологический университет»
e-mail: sgolovin@itstandard.ru

АНДРИАНОВА Елена Гельевна – кандидат технических наук, доцент, доцент кафедры корпоративных информационных систем Института информационных технологий федерального государственного бюджетного образовательного учреждения высшего образования «Московский технологический университет»
e-mail: andrianova@mirea.ru

РАЕВ Вячеслав Константинович – доктор технических наук, профессор, профессор кафедры инструментального и прикладного программного обеспечения Института информационных технологий федерального государственного бюджетного образовательного учреждения высшего образования «Московский технологический университет»,
e-mail: vkr3708@gmail.com

ПОЗДНЕЕВ Борис Михайлович – доктор технических наук, профессор, директор института информационных систем и технологий, заведующий кафедрой информационных систем федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный технологический университет «Станкин»
e-mail:bmp@stankin.ru

База данных (БД) ВИНИТИ РАН

Федеральная база отечественных и зарубежных публикаций по естественным, точным и техническим наукам, генерируется с 1981 г., обновляется ежемесячно, пополнение составляет около 1 млн документов в год. Тематическое наполнение соответствует реферативному журналу ВИНИТИ. Для поиска одновременно по всем или нескольким тематическим фрагментам генерируется единая Политематическая БД.

БД ВИНИТИ РАН в сети INTERNET

Сервер ВИНИТИ - http://www.viniti.ru – обеспечивает on-line доступ к Базе данных ВИНИТИ РАН круглосуточно без выходных.

На основе БД ВИНИТИ РАН предоставляются следующие услуги:

- Диалоговый поиск научно-технической информации **в режиме on-line**;
- **Демо-версия**, позволяющая ознакомиться с основными функциями поисковой системы, составом данных, формами представления документов и получить навыки работы с системой;
- **Поисковые эксперты** ВИНИТИ выполнят тематический поиск по разовым или постоянным запросам, а также окажут **консультационные услуги**.

БД ВИНИТИ РАН на CD-ROM

Любые наборы тематических фрагментов БД ВИНИТИ или их разделов за любой период с 1981 г., а также **проблемно-ориентированные выборки** из БД ВИНИТИ по актуальным направлениям научных исследований могут быть предоставлены на договорной основе **в поисковой системе (ИПС) "Сокол"**, работающей под управлением Microsoft Windows и обеспечивающей следующие возможности:

- **Чтение** документов в режиме последовательного просмотра или выборочно по оглавлению за весь период заказанной ретроспективы
- **Поиск** документов по автору, заглавию, источнику, ключевым словам или словосочетаниям, реферату, рубрикам, году издания, стране, языку и т.д. (всего более 20 признаков)
- **Словарь** системы поможет правильно подобрать термины для поиска и выбрать глубину их усечения.
- **Для уточнения поиска** можно дополнительно использовать год издания документа, язык текста документа, рубрики, шифры тематических разделов БД.
- Выполненные **запросы можно сохранять** для их последующего использования и/или редактирования.

125190, г. Москва, ул. Усиевича, 20, БД ВИНИТИ РАН.
Отдел взаимодействия с потребителями – (499) 155-45-25, (499) 152-58-81
E-mail: csbd@viniti.ru, sales@viniti.ru

WWW: http://www.viniti.ru