МЕЖДУНАРОДНЫЙ ФОРУМ ПО ИНФОРМАЦИИ

2018 **Том 43** **№ 1**

Выходит ежеквартально **Основан в 1975 г.**

СОДЕРЖАНИЕ

Москва **Всероссийский институт научной и технической информации РАН (ВИНИТИ РАН)**

Измерение импакта в оценках исследований: подробное обсуждение методов, эффектов и проблем, касающихся измерений влияния*

Луц БОРНМАН

(Lutz BORNMANN)

Отделение научных и инновационных исследований, Общество Макса Планка, г. Мюнхен, Германия

Импакт в науке – это одна из наиболее важных тем в наукометрии. Последние разработки показывают фундаментальное изменение в измерениях импакта – от воздействия на науку до влияния на общество. Поскольку измерение импакта в настоящее время находится в состоянии далеко идущих изменений, в данной статье описываются недавние разработки и имеющиеся в этой области проблемы. В связи с этим обсуждаются результаты ключевых публикаций (имеющих дело с измерением влияния). Обсуждается, как вообще измеряется импакт в рамках науки и за ее пределами, какие эффекты измерений импакта сказываются на научной системе и какие проблемы связаны с измерением импакта. Проблемы, ассоциирующиеся с измерением импакта, создают основной фокус данной статьи: наука характеризуется неравенством (различием), случайностью, аномалиями, правом совершать ошибки, непредсказуемостью и высоким значением чрезвычайных событий, которые могут исказить измерение влияния. Наукометрия, как производитель подсчетов импакта, и лица, принимающие решения как их потребители, должны помнить об этих проблемах и соответственно рассматривать их в генерации и интерпретации библиометрических результатов.

ВВЕДЕНИЕ

Во всем мире правительства обсуждают вопрос относительно того, как они должны распределять общественные деньги по различным областям (таким, как создание и поддержка инфраструктуры, образование детей и молодежи и защита мира природы в национальном и международном плане) [1]. Распределение денег по различным областям всегда, безоговорочно и явно, создает проблему влияния, которого можно будет достичь путем инвестиции [2]. Возникают вопросы: будет ли инвестиция в защиту мира природы создавать лучшую среду для человечества и увеличивать разнообразие ее видов? Имеет ли инвестиция в образование позитивное влияние на рост национальной экономики, привлекая хорошо образованную молодежь и взрослых на рынок труда? Поскольку наука конкурирует в обществе с другими областями за получение общественных денег, она также сталкивается с вызовом демонстрации своего значения для общества [3]. Основное исследование, в частности, подвергается в этих целях тщательной проверке: ученые могут оценить его значение для общества, а политики могут сделать это с трудом [4,5].

Политики фокусируются на решениях проблем реального мира (таких, как использование воды и энергии или защита окружающей среды) и хотели бы знать степень того, что исследование вносит в проводимую в этих областях работу. Таким образом, они хотят знать не только об общем влиянии исследования на общество, но также и о специфическом влиянии на социально релевантные проблемы [6, 7]. Современное общество подразделяется на различные сегменты (такие, как экономика, наука и правосудие), которые относительно автономны, действуют как отдельные сущности и имеют свои собственные правила [8, 9]. Такие наукометриче-

* Перевод Bornmann L. Measuring impact in research evaluations: A thorough discussion of methods for, effects and problems with impact measurements.— https://link.springer.com/content/pdf/10.1007%2Fs10734-016-9995-x-pdf (Автор расширяет понятие импакт-фактора на широкий круг наукометрических показателей. — *Прим. ред.*)

ские измерения влияния на исследование согласуются с понятием автономной природы различных частей общества. Измерение влияния исследования способом, который касается специфических социально релевантных проблем, является противоположным понятию автономии и может означать концентрацию внимания на предметах, на которых наука, как правило, не фокусируется (так как, например, больше уже нечего исследовать в этой области) [10].

В ряде стран (включая Австралию, Бельгию, Францию, Италию, Новую Зеландию и Великобританию) устанавливаются национальные системы оценки, требующие, чтобы наука отчитывалась за финансирование и показывала, что инвестированные в нее деньги не были потрачены впустую [11, 12]. Эти системы оценивают влияние академических учреждений не только на научный прогресс, но также и на экономику, окружающую среду, оборону и здравоохранение. С 2014 г. Великобритания, например, имеет структуру научного качества (Research Excellence Framework), заменившую схему научной оценки (Research Assessment Exercise – RAE), которая оценивает университеты с помощью рецензирования коллегами, анализа ситуаций и метрик [13, 7]. С одной стороны, национальные системы используются для оценки внутреннего научного исследования в глобальном контексте и того, было ли достигнуто влияние на науку и вне ее. Основной вопрос состоит в том, действует ли исследование лучше или хуже среднего мирового значения. С другой стороны, системы используются для распределения денежных ассигнований научных фондов, которые таким образом распределяются более строго в соответствии с критериями конкуренции, чем в странах, где нет подобной системы.

Поскольку измерение влияния научной продукции находится в состоянии ожидания больших изменений [14], данная статья описывает последние разработки в этой области. С этой целью в статье обсуждается, как обычно влияние измеряется в рамках науки и вне ее (раздел «Измерение влияния»), каковы последствия (эффекты) измерения влияния на научную систему (раздел «Эффекты измерения») и какие проблемы связаны с измерением влияния (раздел «Проблемы измерения влияния»). Проблемы, ассоциируемые с измерением влияния, представляют основной фокус внимания статьи: будет показано, что наука отмечается различием, случайностью, аномалиями, правом делать ошибки, непредсказуемостью и высоким значением чрезвычайных событий, которые ведут к связанным с измерениями влияния проблемам, концентрирующимся на уровнях высокой агрегации и оценивающим продолжительность (непрерывность) в исследовании.

В следующих разделах полный обзор литературы исследований, имеющих дело с влиянием на науку и вне ее, не предлагается, а предлагается обсудить последние разработки в области измерений влияния на основе отобранной ключевой литературы. Она была идентифицирована: 1) в базах данных литературы (например, Web of Science, Thomas Reuters и Scopus, Elsevier) с использованием таких терминов, как «влияние цитирования», «оценка исследования» и «библиометрия», и 2) в соответствующих обзорах литературы [например, 15,16].

ИЗМЕРЕНИЕ ВЛИЯНИЯ

Даже если метрики и анализ ситуаций чаще используются для оценки исследования в различных странах, основой современной оценки исследования являлся и остается процесс рецензирования коллегами, в котором они взаимно оценивают свой научный выход [17]. Рецензирование коллегами является самым старым методом оценки исследования, и его использование тесно связано с развитием современной науки [18, 19]. Только тогда, когда результаты исследования проконтролированы квалифицированными экспертами той же самой области в целях гарантии, что они соответствуют определенным стандартам, исследовательская область может сделать надежные и обоснованные утверждения. Однако метод рецензирования коллегами сталкивается с трудностями своего ограничения, когда надо отрецензировать большое количество (научных) единиц. В национальных системах оценки, как правило, почти каждый университет и исследовательское учреждение неуниверситетского типа в стране подвергаются тщательной проверке. Следовательно, чтобы оценить большее число научных единиц дополнительно к рецензированию используются количественные методы, которые применяют показатели для измерения выхода и/или влияния исследования [15, 16].

Наиболее важным количественным методом в настоящее время является библиометрия, где для измерения влияния используется цитирование публикаций [20]. Например, Австралийский научно-исследовательский совет (Australian Research Council – ARC) в основном применяет анализ цитирования для оценки исследований в естественных науках. И только в области инжиниринга, общественных и гуманитарных науках ARC использует рецензирование коллегами выборочных публикаций или научных выходов, поскольку анализ цитирования не подходит для этих дисциплин [21]. Библиометрия является предпочтительным методом оценки исследования, в первую очередь из-за того, что статьи и книги – это самые важные продукты науки. Кроме этого, подсчеты цитирования обеспечивают информацию относительно того, насколько эти продукты полезны для ученых, работающих в одних и тех же или схожих дисциплинах [22].

Пока еще нет никакого стандартного метода, как библиометрия, который может надежно и обоснованно измерять выгоду исследования для общества (т.е. более широкое влияние исследования) [23, 24]. В наукометрии продолжается поиск «эквивалента цитирования», с помощью которого могут измеряться «такие виды деятельности, как работа в индустрии, вклады в правительственные отчеты или передача исследовательских выходов другим аудиториям, не коллегам исследователя» [7, с. S. 59]. В настоящее время университеты главным образом используют анализ ситуаций, чтобы оценить важность их исследования для общества (например, REF в Великобритании). Для сторонников этого – «анализ ситуаций – единственный жизнеспособный путь оценки влияния; они предлагают потенциал для представления сложной информации и предостерегают от большей концентрации внимания на количественных метриках в отношении влияния анализа ситуаций» [25, с. 49]. Однако использование анализа ситуаций в оценке критикуется, поскольку такой подход может применяться очень избирательно, в случае, если университеты сообщают только те результаты, которые имеют самое лучшее влияние [2]. Кроме того, анализ фактических примеров является дорогостоящим и требует время на подготовку [13, 21].

Проблема нахождения подходящего метода, с помощью которого можно измерять влияние на общество, показывает, что в итоге его более трудно измерять, чем научное влияние [24]. Эта трудность возникает прежде всего из-за того, что существует много различных целе-

вых групп для влияния на общество (тогда как научное влияние – это всегда влияние на науку). Влияние на общество может быть влиянием на разработчиков политики, бизнес или другие (более специфические) части общества [1]. Следовательно, для измерения влияния на общество необходимо в каждом примере принять решение относительно того, каково на самом деле это влияние [2]. В прошлом два метода измерения влияния на общество на основе эквивалента цитирования возникли как особенно полезные: оценка цитирований в патентах (технологическое влияние) [26, 27] и в клинических руководствах (медицинское влияние) [28, 29]. Оба подхода имеют следующие преимущества: 1) влияние на общество может измеряться подобным способом и в научном влиянии (и поэтому доступны логичные методы анализа данных, разработанные за десятилетия наукометрического исследования; 2) тот факт, что они основаны на цитировании, означает, что доступны не являющиеся противодействующими, достаточно объективные и обширные данные; 3) патенты и руководства доступны для оценки в относительно свободной, удобной форме и, по сравнению с другими данными, могут оцениваться с помощью рационального использования усилий.

Наравне с клиническими руководствами и патентами, в последние годы появился другой источник данных, с помощью которого может измеряться влияние исследования в обществе: альтернативные метрики (альтметрия). «Альтметрия … это термин для описания основанных на всемирной сети метрик относительно влияния научного материала, с акцентом на выходы (продукты) социальных медиа как источники данных» [30]. Термин был предложен Примом и др. [31]. В альтметрии также учитываются подсчеты точек зрения, загрузок, кликов, записей, сохранений, твитов, частей, лайков, рекомендаций, тегов, постов, трекбеков, дискуссий, книжных закладок и комментариев. Например, альтернативные метрики представлены для единичных публикаций в такой базе данных, как Scopus (Elsevier), или издателем, таким как PLOS (Public Library of Science) [32, 33]. В общем, альтернативные метрики – это логарифмические данные, измеряющие отдельные упоминания публикаций (такие как загрузки) за определенный период времени [34].

Средство «Altmetric for Institutions» компании Altmetric позволяет учреждениям проследить, проконтролировать и сообщить о широком влиянии исследования (http://www.altmetric.com/institutions.php). Оно подсчитывает, анализирует и (с некоторой обработкой) представляет онлайн упоминания публикаций, выпущенных учреждением на таких платформах как Mendeley, Twitter, CiteLike и Facebook. Однако альтметрия анализирует больше, чем упоминания на платформах социальных медиа. Её анализы также распространяются на документы правительственной политики и другие источники упоминаний научных статей. В частности, документы правительственной политики являются важным источником, так как влияние исследования на разработку политики может, таким образом, стать определяемым количественно [35, 36]. Этот систематический способ измерения влияния на разработку политики представляет потенциальный интерес, в частности, в социальных и гуманитарных науках [37]. В этих дисциплинах почти невозможно применять традиционную библиометрию, и исследователи в области наукометрии заняты поиском альтернативных способов оценки исследования [38].

Согласно Туэйтесу [7], исследователи в области здравоохранения особенно заинтересованы во влиянии, которым их исследование пользуется за рамками науки [39]. Ученые сферы здравоохранения хотели бы оказывать влияние на медицинских практиков с помощью своего исследования. Кроме того, правительство весьма заинтересовано во влиянии исследования на медицину, так как стоимость исследования в этой сфере очень высока. Поскольку влияние, не только в области здравоохранения, часто измеряется по прошествии нескольких лет, то были внесены предложения в исследование влияния на общество, чтобы измерять стремления (учреждений) достичь влияния и соответственно их поощрять. Вместо «мониторинга влияния» было бы лучше осуществлять «мониторинг прогресса в сторону влияния» [2]. Многие голландские организации, вовлеченные в обеспечение качества, сотрудничают в проекте, названном «Оценка исследования в контексте» («Evaluating Research in context – ERiC»), который определил своей целью разработку методов оценки влияния на общество [40]. Одним из значимых результатов этого проекта является то, что продуктивное взаимодействие служит необходимым требованием для исследования – иметь влияние на общество: «должно иметь место некоторое взаимодействие между исследовательской группой и держателями капитала» [40, с. 10]. Такие взаимодействия могут быть в форме личного контакта (например, совместные проекты или сети), публикаций (например, образовательные и оценочные отчеты) [41] и артефактов (например, выставки, программное обеспечение или сетевые сайты). Все эти взаимодействия могут считаться видами деятельности по достижению влияния на общество.

ЭФФЕКТЫ ИЗМЕРЕНИЯ ВЛИЯНИЯ

Поскольку библиометрические показатели получили общее признание в научной политике и достигли релевантного применения в оценке исследования, то сообщалось о стратегиях адаптации ученых, что является результатом использования этих показателей для решения вопросов финансирования [42, 43]. Борнман [44] предложил назвать эти стратегии «мимикрией в науке». Стратегии адаптации могут бросить тень общего сомнения на систему оценки исследования: закон Гудхарта гласит, что «как только измерение становится целью, оно перестает быть хорошим измерением» [45, с. S66]. Существует риск, что ученые больше знают о маркетинге своих научных продуктов, чем о контенте исследования. Ученые применяют стратегии, которые дают им возможность согласиться с библиометрическими оценками и сэкономить финансы для своего собственного исследования. Некоторые из этих стратегий следующие:

(1) Исследователи не суммируют результаты их проекта в одной публикации, а распределяют по многим публикациям, чтобы создать видимость более высокой продуктивности [46, 47]. Такая публикационная стратегия рассматривается как нарезка «салями». Однако наличие нескольких статей на основе одного и того же проекта не всегда является попыткой «обмануть» систему путем создания видимости продуктивности и может иметь независимые, методологические и практические предпосылки, которые совершенно легитимны. (2) Поскольку ученые, как правило, стремятся к оценке мирового уровня, они пытаются опубликовать свои статьи в престижных журналах [6]. Они выбирают не те журналы, которые наиболее подходят для их рукописей, а те, которые пользуются наиболее высоким уважением со стороны их коллег. В дополнение к этим изменениям

относительно публикационного поведения также имеется и (3) давление на ученых с целью «изменить фокус исследования для лучшего выравнивания с основным потоком или с наиболее уважаемыми областями» [45, с. S64], (4) большое нежелание со стороны авторитетных ученых работать с начинающими карьеру исследователями (так как они часто публикуются в менее престижных журналах), и (5) существующая в учреждениях тенденция назначать на должность тех ученых, которые имеют длинный список публикаций, вместо (более молодых) талантливых исследователей, не имеющих такого списка [45].

Установление измерения влияния на общество в качестве составляющей оценки исследования – наряду с измерением влияния на науку – приведет к тому, что ученые начнут направлять свою работу больше в сторону групп людей вне рамок науки. Например, ученые могут представлять отчеты о результатах собственного исследования в специальную группу читателей вне рамок науки (например, политиков). Тогда могут быть использованы альтернативные метрики – как в случае с традиционным цитированием – для измерения влияния отчетов вне науки. Если есть желание внедрить систему измерения выхода продукции (число отчетов) и влияния (число упоминаний отчетов), подобную той, что существует в национальной системе оценки, то необходимо заранее посмотреть, будет ли такой вид измерения иметь подразумеваемые эффекты. (1) Является ли ожидаемая адаптация в поведении ученых подходящей, например, тот факт, что они пишут отчеты для других статусных групп в обществе в дополнение к производству своих научных статей? (2) Вносит ли наука больший вклад в разрешение реальных мировых проблем (таких как использование энергии или защита окружающей среды) с помощью этих новых продуктов и выигрывает ли от этого экономика, окружающая среда, оборона и здравоохранение в стране (см. выше)? Играют ли в результате отчеты выгодную роль для прогресса в сфере технологии, здравоохранения и т.д., чего бы не ожидалось без этих отчетов? Согласно Моргану [2], развитие подходящей методологии оценки, которая принимает в расчет эти и подобные вопросы, является «основной сутью дела» [2, с. 75], которую необходимо рассмотреть.

ПРОБЛЕМЫ ИЗМЕРЕНИЯ ВЛИЯНИЯ

Далее обсуждается несколько проблемных областей измерения импакта. Люди, имеющие дело с влиянием данных, должны знать об этих проблемах и учитывать их при получении и интерпретации результатов.

Оценки требуют времени и денег, независимо от того, как они осуществляются. Научная система с оценкой будет намного успешнее, чем научная система без оценки, для того, чтобы компенсировать нехватку времени и денег. Поскольку измерение влияния на общество появилось как новый (количественный) элемент в оценках в последние годы [48], то возникает вопрос, насколько оно действительно важно, чтобы включить измерение влияния на общество в национальную систему оценки в условиях баланса стоимости и получения пользы от оценок.

Наука является частью общества, отмеченного неравенством, случайностью, аномалиями, правом на ошибку, непредсказуемостью и высоким значением чрезвычайных событий, обсуждаемых далее. Как намерен проиллюстрировать данный раздел, эти характеристики могут привести к проблемам с измерением влияния научной производительности, которое фокусируется на научных единицах на более высоком уровне агрегации и предполагает продолжение в научной деятельности.

Неравенство

Примеров неравенства в науке очень много. Почти в каждом наборе статей (представленных ученым или учреждением) имеется большое количество статей, которые вообще не цитируются или мало цитируются, и только немногие являются высокоцитируемыми статьями [49]. Как показывают результаты в [50], статьи, вносящие вклад в научный прогресс в дисциплине, преимущественно опираются на несколько, ранее опубликованных, важных статей, а не на статьи, сделавшие небольшой вклад. Однако не только на уровне отдельных статей, но также и на уровне индивидуальных ученых, мы видим большое неравенство, влияющее на публикационный выход и цитирование. Авторы работы [51] исследовали 15 153 100 публикующихся ученых (определители отдельных авторов) в БД Scopus. Их результаты показывают, что только 150 608 (< 1%) из них что-то публиковали в каждый год этого 16-летнего периода – непрерывное, продолжающееся присутствие в литературе (Uninterrupted, continuous presence –UCP). Это небольшое ядро ученых с UCP цитируется гораздо больше других, и они (ученые) отвечают за 41,7% всех статей в тот же период и 87,1% всех статей с > 100 цитированиями в тот же самый период».

Право на ошибку и значение чрезвычайных событий

Наука характеризуется не только ассиметричным распространением, но и правом на ошибки. В соответствии с Поппером [52], научные вопросы, гипотезы и проблемы разрешаются с помощью проб и ошибок (а не эмпирическим подтверждением ранее сформулированных гипотез). Когда проблема изучается учеными, то попытки по принятию решения проверяются эмпирически, тогда как слабые альтернативы определяются и отстраняются (удаляются). Поэтому научное исследование всегда склонно ошибаться и ассоциируется с правом использовать метод проб и ошибок, принимать риски и непризнанные (неортодоксальные или интуитивные) пути [53]. Научный прогресс, основанный на методе проб и ошибок, не является кумулятивным; т. е. продолжающееся производство единиц знания, но происходящее в связи с чрезвычайными событиями, которые Кун [54] назвал научными революциями. Важные публикации в дисциплине ведут к полностью новому мышлению, которое отражается изменениями в используемой таксономии [55]. Таксономия до революции фундаментально отличается от таксономии после революции.

Аномалии

Там, где неравенство, ошибка и чрезвычайные события являются важными компонентами в системе, мы можем предположить, что аномалии также играют ключевую роль. Как показывают исследования Борнмана и др. [50] и Ионидиса и др. [51], наука в основном определяется несколькими элементами (такими, как публикации и ученые), а не количественной массой. Даже журналы, которые постоянно анализируются для литературы базы данных Web of Science (Thomson Reuters), отбираются на основе предположения, что научный прогресс может быть представлен небольшим количеством ядерных журналов и, следовательно, нет необходимости для включения огромного множества журналов [56]. Поскольку библиометрические анализы, как правило, предпринимаются на более высоком уровне агрегации (таком, как учреждения или страны), наблюдается воз-

действие аномалий (несколько высокоцитируемых статей) на целое или они рассматриваются пока в качестве проблем. Например, Гёттингенский университет достиг хорошего положения только в предшествующем выпуске Лейденского ранжирования (которое использовало основанный на среднем цитировании показатель для ранжирования учреждений), потому что смог опубликовать одну очень высокоцитируемую публикацию [57]. В отношении анализов на более высоком уровне агрегации, где данные очень отклоняются, несколько аномалий ответственны за результаты, но их огромное влияние на общий результат (такой, как влияние среднего цитирования на учреждение) часто не видно. Учреждение приобретает высокое значение в измерении влияния не из-за большого числа нанимаемых им на работу ученых и их публикаций, а благодаря всего лишь нескольким успешным ученым и их небольшому количеству высокоцитируемых статей.

Случайность и непредсказуемость

Кампанарио [58] опубликовал ряд достойных свидетельств существования серендипности в науке. Он изучил 205 комментариев относительно классики цитирования из высокоцитируемых статей в недавней истории науки. «Авторы 17 комментариев относительно классики цитирования (8,3%) упоминают некоторый вид серендипности при осуществлении исследования, о котором сообщалось в высокоцитируемой статье». Термин «серендипность» тесно ассоциируется с Робертом К. Мертоном, опубликовавшим совместно с Элинор Барбер книгу на эту тему [59]. Очень многое, что представляется важным в науке, является результатом счастливого совпадения, возникающего по неизвестным причинам или причинам, которые почти невозможно найти рациональным способом. Открытие рентгеновских лучей и пенициллина – яркие примеры этого [60]. Существование (и важность) случайных элементов в научной работе показывает, что есть очевидные параллели между приобретением знания в науке и эволюционными процессами в природе [61, 62]. В то время как живые организмы адаптируются к генетическим изменениям в природе, научные открытия являются результатом (среди прочего) исследований, которые случайно оказываются подходящими. Исследователи обнаруживают то, чего они даже и не искали.

Это проистекает из случайных элементов в процессе создания знания, так что важный прогресс в науке часто непредсказуем. В науке имеется много примеров, где важность результатов определенного исследования проявляется только спустя десятилетия после их опубликования [63, 64]. Во время публикации лишь несколько ученых (по крайней мере, рецензенты рукописи) ожидали, что она будет иметь какое-либо значение для научного сообщества. Например, «Предел Шокли-Квайссера» [65] описывает ограниченную эффективность солнечных элементов (батареек) на основе процессов поглощения и реэмиссии (повторного излучения) [66]. Принятие рукописи статьи с учетом последующего цитирования было первоначально встречено с колебаниями. Однако эта статья стала одной из высокоцитируемых в своей области. Цитирование статьи, опубликованной Шокли и Квайссером [65], развивалось сравнительно синхронно с быстро растущей областью исследования, касающегося солнечных элементов и фотогальваники.

Однако, даже тогда, когда результаты представлялись непосредственно в дисциплину и только позже оказывались особенно важными, специалисты той же самой области часто не признавали их широкое значение и, следовательно, не цитировали эти статьи. Маркс и Борнман [67, 68] показали в двух библиометрических исследованиях о научных революциях, что ряд публикаций, которые оказались важными для научной революции, цитировались редко. Поэтому важное открытие часто приписывается автору, который связывает различные «нити» знания, необходимого для открытия, в значимую дорогу, а не тем, кто внес вклад или опубликовал решающий эмпирический результат или теорию. Борнман и Маркс [69] предложили назвать этот процесс «Принципом Анны Карениной», где различные и необходимые нити знания, полученные из теоретических и эмпирических процессов, соединены вместе для создания революционного открытия.

Важность неравенства, аномалий, случайности, непредсказуемости, ошибок и революционных событий в процессе научного открытия часто наблюдается в литературе как свидетельство того, что наука не может планироваться, управляться или измеряться и, следовательно, было бы излишним измерять влияние в рамках структуры оценок исследований [70]. Особо выделяются две черты оценок, которые ведут к проблемам с измерениями влияния видов научной деятельности, действующих на фоне неравенства, аномалий, случайности, непредсказуемости, ошибок и революционных событий. Эти черты следующие: (1) единицы, обычно оцениваемые в измерении влияния, и (2) непрерывность, желательная в выполнении исследования.

Оцениваемые единицы

Измеряемыми влиянием часто являются учреждения (такие, как университеты) или страны (такие, как страны БРИКС). Ранжирования университетов, такие как Times Higher Education World University Rankings (http://www.timeshighereducation.co.uk), которые регулярно сообщают о достижениях университетов, служат хорошим примером [71]. Как показывают результаты Борнмана и др. [72], основанные на университетах из Лейденского ранжирования (Leiden Ranking), только 4,3% расхождений во влиянии институционального цитирования может быть отнесено за счет различий между университетами. Остальные, то есть 95,7%, связаны с расхождением внутри университетов (т.е. с факультетами, исследовательскими группами и отдельными учеными). Разнородность выполнения исследований, являющаяся результатом неравенства, аномалий, случайности, непредсказуемости, ошибок и революционных событий, следовательно, так явно выражена в рамках университетов, что кажется спорным измерять влияние на институциональном уровне. Вместо этого более подходящим представляется оценивать выполнение исследования на основе небольших единиц, отвечающих за важные результаты, которые ведут к значительному научному прогрессу.

Возможной причиной популярности анализов влияния на институциональном уровне и на уровне страны несомненно является доступность данных. Гораздо сложнее собирать данные по научно-исследовательским группам и отдельным ученым, чем по учреждениям и странам. Данные по отдельным ученым трудно определить, так как разные люди могут иметь одни и те же фамилии [73]. Авторы статей всегда приводят названия учреждений, в которых они работают, и страны, где располагается учреждение, но они редко дают какую-либо информацию об исследовательской группе, отделении и т.п. Другой причиной популярности анализов

учреждений и стран безусловно является политический фокус внимания на эти единицы [71]. Правительство страны видит себя на уровне конкуренции с другими странами и поэтому призывает к релевантным, на уровне страны, исследованиям [74].

Непрерывность

Вообще говоря, оценка науки – это непрерывный процесс. Институциональные оценки повторяются каждые несколько лет, чтобы выяснить, действительно ли постоянное выделение финансовых ресурсов создаст непрерывный выход (продукции) и влияние. Ожидается, что каждая новая оценка учреждения покажет, что научный выход и влияние предыдущей оценки, по крайней мере, были достигнуты и значительно превышены. Поэтому ученых просят производить постоянный и растущий поток важных научных результатов, которые могут быть опубликованы в высоковлиятельных журналах и со временем достичь высокого влияния. Однако характерные, описанные выше черты науки (неравенства, аномалии, случайность, непредсказуемость, ошибки и революционные события) не дают оснований ожидать линейного отношения между входом и выходом в научной работе. Если предположить, что такое отношение должно быть линейным, то есть опасность, что несоответствие между сильным желанием оценки и реальностью процесса исследования в результате выразится в неправильном научном поведении приспособить реальность к желанию [53]. Чтобы противостоять неудовлетворенности в отношении этого несоответствия, в расчет принимается неправильное поведение в оценке или для оценки.

Общая теория Мертона [75] относительно неправильного поведения в обществе может использоваться для объяснения связи между неудовлетворенностью и отклоняющимся поведением в науке. Мертон [75] делает различие между тремя факторами в социальной структуре, чтобы объяснить отклоняющееся поведение в обществе (такое, как преступность в США): (1) определенные желания и ожидания представляют важные культурные цели в обществе; (2) нормы, устанавливающие правила того, как эти цели должны быть достигнуты, и (3) распределение ресурсов, необходимых для достижения целей. По Мертону [75], отклоняющееся поведение проявляется тогда, когда социальная структура делает невозможным достижение разделяемых обществом культурных целей (в США, например, это личное благосостояние) с помощью социально принятых средств (таких, как честная работа). Отклоняющееся поведение, вероятно, может встречаться у лиц в определенной группе, если они могут достичь целей, продиктованных обществом, только с большим трудом, и если они используют законные средства. Это возникает в ситуации, где определенные признаки успеха слишком переоцениваются группой, но только небольшая часть этих признаков может приобретаться легитимными средствами благодаря способу распределения ресурсов.

На фоне этого общего механизма объяснения отклоняющегося от нормы поведения научное неправильное поведение является результатом ситуации, в которой ученые сталкиваются с целями, выдвинутыми постоянными оценками («выигрыш наверняка»), которые могут быть достигнуты только с большим трудом или вообще не достигнуты в рамках когнитивных и социальных норм в науке («выигрыш с преодолением ограниченных способов активности»). Результаты исследования фальсифицируются или подгоняются, чтобы произвести для оценок новые результаты, которые публикуются в достойных уважения журналах. МакДжилврей [45] описывает попытку выигрыша в национальной системе оценки: «возможно, наиболее вопиющий пример попытки «нападения» на систему имел место в 2006 г. в Новой Зеландии. Один из ведущих университетов переклассифицировал множество штатных сотрудников, особенно тех, кто представлял основанный на производительности научный фонд (Performance-Based Research Fund – PBRF) – сотрудников, имеющих право быть избранными на должность, но производящих мало активных исследований. С помощью переклассификации инертные ученые были отстранены от таких предметов как экономика и биология и переведены в такие области, как философия и исследования в религии, что позволило бы университету улучшить свое положение в первых областях. Увеличение числа новозеландских философов вызвало любопытство обозревателей PBRF, которые в конце концов отменили классификации» [45, с. 66].

Обращение к отклоняющемуся поведению (такому, как фабрикация и фальсифицирование научных результатов и переназначение людей, осуществляющих оценку) можно отнести к следующему: (1) чрезмерный акцент на целях при постоянных оценках, которые устанавливаются с помощью определенных признаков успеха в исследовании (такого, как более высокое влияние цитирования); (2) обычно заниженная важность правил, предназначенных для применения к процессам достижения целей (таким, как достижение высокого влияния цитирования), и (3) ограниченная доступность финансовых и людских ресурсов, с помощью которых возможно достижение этих целей.

ЗАКЛЮЧЕНИЕ

Вне всякого сомнения, без науки не будет никакого прогресса в обществе. Результаты научного исследования проявляются в новой технологии, большем понимании нашей планеты и вселенной и в области лечения населения, что позволяет нам жить дольше и сохранять здоровье [23]. Ученые изучают причины изменения климата и работают над более безопасными и надежными решениями по обеспечению нас энергией. Они значительно улучшили точность прогнозирования погоды и внесли весомый вклад в осуществление контроля над инфекционными заболеваниями. Несмотря на эти достижения, современное общество [76, 77] надеется, что наука и другие сферы общества ответственны за состояние результатов своих продуктов и рассматривают специфические проблемы, как убеждено правительство, с большой самоотверженностью. Поскольку эти специфические требования относительно науки представляют сравнительно новый феномен (возникший в последние десятилетия), то встает вопрос, как они меняют или уже изменили науку. С позиции Лухмана [8,9], мы можем ожидать, что наука примет эти вызовы и с еще большим усилием направит на них свое воздействие; однако это будет возможно только при методах, инструментах и практиках, которыми всегда оперирует наука. Поэтому исследование как деятельность, с точки зрения Лухмана [8,6], не будет фундаментально меняться.

Даже если способ, с помощью которого делаются новые открытия, не меняется, то вероятно, что способ их публикации будет подвергаться изменению. Если влияние исследования не только на науку, но и в более широком плане на многие сферы общества измеряется, то будет полезно сформулировать некоторые виды

ориентации для соответствующей читательской аудитории. Сформулированные определенным образом тексты будут иметь больший шанс создания влияния, чем типичные научные тексты (т. е. статьи, опубликованные в научных журналах). Такая подготовка должна осуществляться другими лицами, как, например, журнальные работники – сотрудники сферы науки. Однако такие сотрудники чувствуют себя комфортно (на своем месте) в собственных подсекциях общества, массмедиа [78] и действуют в соответствии с применяемыми здесь правилами, чтобы достичь широкой аудитории. Если такая подготовка предпринимается самими учеными, то их публикационные привычки должны изменяться [41].

Данная статья изучает измерение импакта и исследует причины того, почему влияние науки измеряется особым методом. Она концентрируется на проблемах измерения влияния и его не специально разработанных последствий. Обе области должны в настоящее время заслуживать особого внимания, так как измерение влияния проводится в более широком масштабе, чем несколько лет тому назад. Требуется всестороннее наукометрическое исследование для точного, эффективного и систематического определения и градации основанных на метриках систем оценки. Однако это исследование касается не только разработки надежных и обоснованных показателей, но также воздействий и проблем, вызванных этими системами. Главный вопрос должен состоять в том, насколько эти системы могут реально отражаться на совершенствовании науки. Например, в этих исследованиях различные страны можно сравнивать по тому, используют ли они национальные системы оценки или нет. Проблема в проведении таких исследований будет состоять в том, чтобы отделить эффект системы оценки на производительность страны от других моментов (например, большее или меньшее получение денег от государства на исследование).

Благодарность. Выражается признательность за финансирование открытого доступа, обеспеченного Обществом Макса Планка.

ЛИТЕРАТУРА

1. *Khazragui H., Hudson J.* Measuring the benefits of university research: Impact and the REF in the UK// Research Evaluation. — 2015. — Vol. 24, No. 1. — P. 51–62. — doi:10.1093/reseval/rvu028.

2. *Morgan B.* Research impact: Income for outcome// Nature. —2014. — Vol. 511, No. 7510. — S72–S75. — doi:10.1038/511S72a.

3. *Cohen G., Schroeder J., Newson R., King L., Rychetnik L., Milat A. J., et al.* Does health intervention research have real world policy and practice impacts: Testing a new impact assessment tool// Health Research Policy and Systems. — 2015. — Vol. 13, No. 12. — doi:10.1186/1478-4505-13-3.

4. *Bornmann L.* Measuring the societal impact of research// EMBO Reports. — 2012. — Vol. 13, No. 8. — P. 673–676.

5. *Bornmann L.* What is societal impact of research and how can it be assessed? A literature survey// Journal of the American Society of Information Science and Technology. — 2013. —Vol. 64, No.2. — P. 217–233.

6. *Finkel A.* Perspective: Powering up citations// Nature. — 2014.—Vol. 511, No. 7510. — S77. — doi:10.1038/511S77a.

7. *Thwaites T.* Research metrics: Calling science to account// Nature. — 2014. — Vol. 511, No. 7510. — S57–S60. — doi:10.1038/511S57a.

8. *Luhmann N.* Theory of society (Vol. 1). — Stanford, CA: Stanford University Press, 2012.

9. *Luhmann N.* Theory of society (Vol. 2). — Stanford, CA: Stanford University Press, 2012.

10. *Douglas H.* Pure science and the problem of progress// Studies in History and Philosophy of Science Part A. — 2014. — Vol. 46. — P. 55–63. — doi:10.1016/j.shpsa.2014.02.001.

11. *Abramo G., D'Angelo C.* Evaluating research: From informed peer review to bibliometrics// Scientometrics. — 2011. —Vol. 87, No. 3. — P. 499–514. — doi:10.1007/s11192-011-0352-7.

12. *Derrick G. E., Pavone V.* Democratising research evaluation: Achieving greater public engagement with bibliometrics-informed peer review// Science and Public Policy. — 2013. — Vol. 40, No. 5. — P. 563–575. — doi:10.1093/scipol/sct007.

13. *King's College London and Digital Science.* The nature, scale and beneficiaries of research impact: An initial analysis of Research Excellence Framework (REF) 2014 impact case studies. —London: King's College London, 2015.

14. *Bornmann L.* Is there currently a scientific revolution in scientometrics?// Journal of the Association for Information Science and Technology. —2014. — Vol. 65, No. 3. — P. 647–648.

15. *de Bellis N.* Bibliometrics and citation analysis: From the science citation index to cybermetrics. — Lanham, MD: Scarecrow Press, 2009.

16. *Hicks D., Melkers J.* Bibliometrics as a tool for research evaluation/ A. N. Link, N. S. Vonortas (Eds.), Handbook on the theory and practice of program evaluation (pp. 323–349). — Northampton, MA: Edward Elgar, 2013.

17. *Bornmann L.* Scientific peer review// Annual Review of Information Science and Technology. — 2011. — Vol. 45. — P. 199–245.

18. *Geisler E.* The metrics of science and technology. — Westport, CT: Quorum Books, 2000.

19. *Virelli L. J.* Scientific peer review and administrative legitimacy// Administrative Law Review. —2009. — Vol. 61, No. 4. — P. 723–780.

20. *Hicks D., Wouters P., Waltman L., de Rijcke S., Rafols I.* Bibliometrics: The Leiden Manifesto for research metrics// Nature. — 2015. — Vol. 520, No. 7548. — P. 429–431.

21. *Sheil M.* Perspective: On the verge of a new ERA// Nature. —2014. — Vol. 511, No. 7510. — S67.— doi:10.1038/511S67a.

22. *Moed H. F.* Citation analysis in research evaluation. — Dordrecht: Springer, 2005.

23. *Campbell P., Grayson M.* Assessing science// Nature. — 2014. — Vol. 511, No. 7510. — S49. — doi:10.1038/511S49a.

24. *National Research Council. Furthering America's Research Enterprise.* —Washington, DC: The National Academies Press, 2014.

25. *Wilsdon J., Allen L., Belfiore E., Campbell P., Curry S., Hill S. et al.* The metric tide: Report of the independent review of the role of metrics in research assessment and man-

agement. — Bristol, UK: Higher Education Funding Council for England (HEFCE), 2015.

26. *Austrian Science Fund.* Rethinking the impact of basic research on society and the economy. — Vienna: Austrian Science Fund, 2007.

27. Kousha K., Thelwall M. Patent citation analysis with Google// Journal of the Association for Information Science and Technology. — doi:10.1002/asi.23608. (в печати)

28. *Lewison G., Sullivan R.* The impact of cancer research: How publications influence UK cancer clinical guidelines// British Journal of Cancer. —2008.—Vol. 98, No. 12. — P. 1944–1950.

29. *Thelwall M., Maflahi N.* Guideline references and academic citations as evidence of the clinical value of health research// Journal of the Association for Information Science and Technology. — 2015. — doi:10.1002/asi.23432.

30. *Shema H., Bar-Ilan J., Thelwall M.* Do blog citations correlate with a higher number of future citations? Research blogs as a potential source for alternative metrics// Journal of the Association for Information Science and Technology. — 2014. — Vol. 65, No. 5. —P. 1018–1027. — doi:10.1002/asi.23037.

31. *Priem J., Taraborelli D., Groth P., Neylon C.* Altmetrics: A manifesto. — 2010. — http://altmetrics.org/manifesto/.

32. *Liu C. L., Xu Y. Q., Wu H., Chen S. S., Guo J. J.* Correlation and interaction visualization of altmetric indicators extracted from scholarly social network activities: Dimensions and structure // Journal of Medical Internet Research. — 2013. — Vol. 15, No. 11. — P. 17. — doi:10.2196/jmir.2707.

33. *Zahedi Z., Costas R., Wouters P.* How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications // Scientometrics. — 2014. — Vol. 101, No. 2. — P. 1491–1513. — doi:10.1007/s11192-014-1264-0.

34. *Haustein S.* Readership metrics/B. Cronin, C. R. Sugimoto (Eds.), Beyond bibliometrics: Harnessing multidimensional indicators of performance (pp. 327–344). — Cambridge, MA: MIT Press, 2014.

35. *Bornmann L., Haunschild R., Marx W.* Policy documents as sources for measuring societal impact: How is climate change research perceived in policy documents? — 2016. — http://arxiv.org/abs/1512.07071.

36. *Liu J.* New source alert: Policy documents. — 2014. — http://www.altmetric.com/blog/new-source-alert-policy-documents/.

37. *Hug S. E., Ochsner M., Daniel H.-D.* Criteria for assessing research quality in the humanities— A Delphi study among scholars of English literature, German literature and art history// Research Evaluation. —2013. — Vol. 22, No. 5. — P. 369–383.

38. Hammarfelt B. Using altmetrics for assessing research impact in the humanities// Scientometrics. — 2014. — doi:10.1007/s11192-014-1261-3.

39. *Thonon F., Boulkedid R., Delory T., Rousseau S., Saghatchian M., van Harten W., et al.* Measuring the outcome of biomedical research: A systematic literature review// PLoS ONE. — 2015. — doi:10.1371/journal.pone.0122239.

40. *ERiC.* Evaluating the societal relevance of academic research: A guide. — Delft: Delft University of Technology, 2010.

41. *Bornmann L., Marx W.* How should the societal impact of research be generated and measured? A proposal for a simple and practicable approach to allow interdisciplinary comparisons// Scientometrics. —2014. — Vol. 98, No.1. — P. 211–219.

42. *Evidence Ltd.* The use of bibliometrics to measure research quality in UK higher education institutions. — London: Universities UK, 2007.

43. Lawrence P. A. The politics of publication. Authors, reviewers and editors must act to protect the quality of research// Nature. — 2003. — Vol. 422, No. 6929. — P. 259–261.

44. *Bornmann L.* Mimicry in science?//Scientometrics. — 2011. — Vol. 86, No. 1. — P. 173–177. — doi:10.1007/s11192-010-0222-8.

45. *McGilvray A.* Research assessment: The limits of excellence// Nature. — 2014. — Vol. 511, No. 7510. — S64–S66. — doi:10.1038/511S64a.

46. *Bornmann L., Daniel H.-D.* Multiple publication on a single research study: Does it pay? The influence of number of research articles on total citation counts in biomedicine// Journal of the American Society for Information Science and Technology. —2007. — Vol. 58, No. 8. — P. 1100–1107.

47. *Mallapaty S.* Q&A Jane Harding: Individual approach// Nature. —2014. — Vol. 511, No. 7510. — S82. — doi:10.1038/511S82a.

48. *Ovseiko P. V., Oancea A., Buchan A. M.* Assessing research impact in academic clinical medicine: A study using Research Excellence Framework pilot impact indicators// BMC Health Services Research. — 2012. — doi:10.1186/1472-6963-12-478.

49. *Seglen P. O.* The skewness of science// Journal of the American Society for Information Science. — 1992. — Vol. 43, No. 9. — P. 628–638.

50. *Bornmann L., de Moya-Anego´n F., Leydesdorff L.* Do scientific advancements lean on the shoulders of giants? A bibliometric investigation of the Ortega hypothesis // PLoS ONE. — 2010. — Vol. 5, No.10, e11344.

51. *Ioannidis J. P. A., Boyack K. W., Klavans R.* Estimates of the continuously publishing core in the scientific workforce// PLoS ONE. — 2014. — Vol. 9, No. 7, e101698. — doi:10.1371/journal.pone.0101698.

52. *Popper K. R.* The logic of scientific discovery (2nd ed.). — New York, NY: Basic Books, 1961.

53. *Bornmann L.* Research misconduct—Definitions, manifestations and extent// Publications. — 2013. — Vol. 1, No. 3. — P. 87–98.

54. *Kuhn T. S.* The structure of scientific revolutions (2nd ed.). — Chicago IL: University of Chicago Press, 1962.

55. *Wray K. B.* Kuhn's evolutionary social epistemology. — Cambridge: Cambridge University Press, 2011.

56. *Garfield E. The* history and meaning of the journal impact factor// Journal of the American Medical Association. — 2006. —Vol. 295, No. 1. — P. 90–93.

57. *Waltman L., Calero-Medina C., Kosten J., Noyons E. C. M., Tijssen R. J. W., van Eck N. J., et al.* The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation // Journal of the American Society for Information Science and Technology. — 2012. — Vol. 63, No. 12. — P. 2419–2432.

58. *Campanario J. M.* Using citation classics to study the incidence of serendipity in scientific discovery// Scientometrics. — 1996. — Vol. 37, No.1. — P. 3–24. — doi:10.1007/bf02093482.

59. *Merton R. K., Barber E. G.* The travels and adventures of serendipity: A study in historical semantics and the sociology of science. — Princeton: Princeton University Press, 2004.

60. *Ban T. A.* The role of serendipity in drug discovery// Dialogues in Clinical Neuroscience. —2006. — Vol. 8, No. 3. — P. 335–344.

61. *Feist G. J.* The psychology of science and the origins of the scientific mind. — New Haven, CT: Yale University Press, 2006.

62. *Gieryn T. F.* Boundaries of science / S. Jasanoff, G. E. Markle, J. C. Petersen, T. Pinch (Eds.), Handbook of science and technology studies (pp. 393–443). — London: Sage, 1995.

63. *Ke Q., Ferrara E., Radicchi F., Flammini A.* Defining and identifying sleeping beauties in science// Proceedings of the National Academy of Sciences. — 2015. — doi:10.1073/pnas.1424329112.

64. *van Raan A. F. J.* Sleeping beauties in science// Scientometrics. — 2004. — Vol. 59, No. 3. — P. 467–472.

65. *Shockley W., Queisser H. J.* Detailed balance limit of efficiency of p–n junction solar cells// Journal of Applied Physics. — 1961. — Vol. 32, No. 3. — P. 510. — doi:10.1063/1.1736034.

66. *Marx W.* The Shockley-Queisser paper—A notable example of a scientific sleeping beauty// Annalen der Physik. —2014. —Vol. 526, No. (5–6). — A41–A45. — doi:10.1002/andp.201400806.

67. *Marx W., Bornmann L.* How accurately does Thomas Kuhn's model of paradigm change describe the transition from a static to a dynamic universe in cosmology? A historical reconstruction and citation analysis// Scientometrics. — 2010. — Vol. 84, No. 2. — P. 441–464.

68. *Marx W., Bornmann L.* The emergence of plate tectonics and the Kuhnian model of paradigm shift: A bibliometric case study based on the Anna Karenina principle// Scientometrics. — 2013. — Vol. 94, No. 2. — P. 595–614. — doi:10.1007/s11192-012-0741-6.

69. *Bornmann L., Marx W.* The Anna Karenina principle: A way of thinking about success in science// Journal of the American Society for Information Science and Technology. — 2012. — Vol. 63, No. 10. — P. 2037–2051. — doi:10.1002/asi.22661.

70. *Schatz G.* The faces of big science// Nature Reviews Molecular Cell Biology. — 2014. — Vol. 15, No.6. — P. 423–426. — doi:10.1038/nrm3807.

71. Hazelkorn E. Rankings and the reshaping of higher education. The battle for world-class excellence. — New York, NY: Palgrave Macmillan, 2011.

72. *Bornmann L., Mutz R., Daniel H.-D.* A multilevel-statistical reformulation of citation-based university rankings: The Leiden Ranking 2011/2012// Journal of the American Society for Information Science and Technology. — 2013. — Vol. 64, No. 8. — P. 1649–1658.

73. *Boyack K. W., Klavans R., Sorensen A. A., Ioannidis J. P. A.* A list of highly influential biomedical researchers, 1996–2011 // European Journal of Clinical Investigation. — 2013. — Vol. 43, No. 12. — P. 1339–1365. — doi:10.1111/eci.12171.

74. *National Science Board.* Science and engineering indicators 2014. — Arlington, VA: National Science Foundation (NSF), 2014.

75. *Merton R. K.* Social structure and anomie// American Sociological Review. — 1938. — Vol. 3, No. 5. — P. 672–682.

76. *Dahler-Larsen P.* The evaluation society. — Stanford: Stanford University Press, 2011.

77. *Power M.* The audit society: Rituals of verification. — Oxford: Oxford University Press, 1999.

78. *Luhmann N.* The Reality of the mass media. — Stanford, CA: Stanford University Press, 2000.

Кластеризация статей на основе семантического сходства*

Шенхуи ВАН
(Shenghui WANG)

Роб КУПМАН
(Rob KOOPMAN)

Исследовательский центр в составе Автоматизированного библиотечного центра с интерактивным доступом, г. Лейден, Нидерланды

Кластеризация документов, как правило, представляет первый шаг тематической идентификации. Поскольку многие методы кластеризации основаны на сходствах между документами, то важно создать представления таких документов с максимальным сохранением их семантики и соответствия эффективному подсчету сходства. Как описывается в Трудах 15-й конференции Международного общества по наукометрии и информетрии [1], метаданные статей в массиве Astro соотносятся с семантической матрицей, использующей векторное пространство для охвата семантики показателей, взятых из этих статей, и впоследствии поддерживающей изучение этих единиц в контексте Малой Ариадны (LittleAriadne). Однако эта семантическая матрица не предполагает подсчет прямых сходств между статьями. В статье подробно описывается формирование семантического представления статьи из единиц связанных с ней. Основываясь на таких семантических представлениях статей, применяются два стандартных метода кластеризации, а именно алгоритм K-Means и алгоритм выявления сообществ, разработанный в Лёвенском университете, определившие два наших решения, далее именуемые OCLC-31 (подразумевая K-Means) и OCLC-Louvain (предполагая Louvain). Также дается подробное описание механизма внедрения и базового сравнения с другими решениями, имеющимися в кластеризации, которым посвящен специальный выпуск журнала Scientometrics, Special Issue of Scientometrics (2017).

ВВЕДЕНИЕ

Темы, подобласти, специальности составляют ядро независимого процесса производства научного знания [2]. Существует множество точек зрения на то, как определять эти единицы когнитивной и социальной организации [3], и идут дискуссии относительно их извлечения автоматизированным и алгоритмизированным способом [4]. На сегодняшний день способом определения *тем* является кластеризация документов. Имеются разные способы установления того, рассматривают ли два документа один и тот же тематический вопрос. Некоторые известные признаки тематического сходства включают ссылки (факт наличия цитируемости одного документа другим) [5], совместные ссылки (факт наличия цитируемости двух документов третьим) [6], библиографическое сочетание (факт объединения двух документов с помощью ссылки в их библиографии) [7] и взаимосвязи часто встречающихся слов (факт объединения двух документов с помощью определенных слов) [8]. Каждый их этих признаков или оттенков может использоваться для построения различных матриц родства или сходства документов, из которых могут быть определены кластеры документов или тем.

В библиометрической литературе преимущества и недостатки различных методов обсуждались неоднократно. Вообще проводят различие между метриками на основе цитирования и на основе текста [9]. Хотя предполагается, что слова менее кодируемы, чем библиографические ссылки, мы разделяем мнение, что слова, особенно в названиях и рефератах, олицетворяют некое знание, постулируемое в статье [10]. Поэтому, в соответствии с программой когнитивной наукометрии [11] и последними библиометрическими исследованиями полных текстов [9], мы утверждаем родство двух документов при условии достаточного разделения ими лексической информации.

В отношении подробно изложенных в статье подходов к кластеризации мы полагаемся на новое семантическое представление статей в целях выявления их сходства. Соответствующий метод и основанный на нем интерфейс интерактивного поиска назван *Ариадной* [1, 12]. Наш под-

* Перевод Wang S., Koopman R. Clustering articles based on semantic similarity. –https://link.springer.com/content/pdf/10.1007%2Fs11192-017-2298-x.pdf

ход очень похож на методы, используемые в информационном поиске, тем, как он действует в пространстве слов. Но в отличие от методов, основанных на пространстве слов из документов, установленном Сэлтоном, нами используется информация ото всех элементов документа (в нашем случае статья) и создается пространство слов для всех таких элементов или показателей. Мотивация к этому основана на предположении, что использование информации из многих различных элементов статьи обусловливает более точное семантическое представление этой статьи. Мы, следовательно, полагаем, что также улучшается базис, по которому определяется сходство/родство статей. Когда используются такие единицы, как авторы, журналы, предметные заголовки (рубрики) или ссылки, одновременно ведется поиск семантического сходства/родства с точки зрения организации производства научного знания, социальной (авторы), коммуникативной (журнал как издательская продукция) или обмена знанием (ссылки).

Вопросы нашего **исследования** следующие: (а) сможем ли мы воссоздать надежное семантическое представление для статей из всех, связанных с ней единиц и (б) определить кластеры статей с применением стандартных методов, основанных на таком семантическом представлении.

В статье сначала описывается метод представления семантики статей на основе включенных в них единиц. Затем перед ознакомлением с особенностями внедрения кратко вводятся два стандартных метода кластеризации – алгоритм K-Means и алгоритм выявления сообществ, разработанный в Лёвенском университете. В конце сравниваются два наших решения с другими способами кластеризации, имеющимися в специальном выпуске журнала Scientometrics, Special Issue of Scientometrics (2017), и далее в статье приводится заключение.

ОТ СЕМАНТИКИ ЕДИНИЦ К СЕМАНТИКЕ СТАТЕЙ

Применительно к нашему подходу адаптируем понятие *статистической семантики* [13,14], основанное на предположении, «слово характеризуется поддерживающим его окружением» [15], или *распределительной гипотезы* из лингвистики [16]: слова, встречающиеся в одинаковых контекстах, склонны иметь одинаковые значения. В *Ариадне* расширяем слова до единиц (таких как авторы, журналы, предметные области, ссылки) таким образом, чтобы каждая единица индексировалась вектором семантического пространства, отражающим ее лексический контекст, т.е. их совместная встречаемость с некоторыми терминами (включая тематические термины, извлеченные из названия и реферата, а также предметные области, устанавливающие пользователя) [17].

Сводная матрица совместной встречаемости показатель-термин может стать чрезмерно большой и пространственной, что сделает дорогостоящим и непрактичным любое крайнее вычисление. Благодаря случайной проекции [18, 19], можно значительно снизить протяженность этого семантического пространства, получив меньшую и с управляемым размером *семантическую матрицу*, при этом максимально сохраняя семантику этих единиц. При помощи всех единиц в виде векторов одного и того же семантического пространства возможно вычислить расстояние или родство любых пар единиц без учета их типа. Такая свобода – уникальная черта *Ариадны*. Она обусловливает контекстуальную точку зрения на объект или запрос как начало исследовательского пути. За более подробным описанием просим обращаться к другим источникам [12,17].

В этой семантической матрице каждая статья поддерживает семантику отдельных единиц. Во время вычислений на большом массиве статистика ведет себя надежно в сходствах единиц. Однако эта семантическая матрица не допускает прямого подсчета сходств статей.

Чтобы кластеризировать статьи и провести сравнение с другими методами, сначала создадим интегрированное представление статьи из связанных с ней единиц. Для этого в каждой статье ищем все связанные с ней единицы в семантической матрице. В итоге получаем набор векторов $V=\{v_1,\ldots,v_n\}$ для каждой статьи, где n – число связанных со статьей единиц, а v_i – вектор единицы e_i. Эти единицы могут быть авторами, предметными областями, журналом, ссылками, тематическими терминами (извлеченными из названия и реферата статьи) и т.д. Каждая статья представлена уникальным набором векторов. Размер массива n может варьироваться, но каждый из векторов внутри массива имеет одинаковую длину, в нашем случае 600 (подробнее см. [17]).

Для каждой статьи строится сейчас новый вектор v', взвешенный центроид ее содержащих векторов

$$v' = \frac{\sum_{i=1}^{n} w_i . v_i}{\sum_{i=1}^{n} w_i}, \qquad (1)$$

где $w_i = \log(N/f_i)^3$, N – общее число статей и f_i – число статей, содержащих единицу e_i. При этом особом взвешивании частотные объекты жестко дисциплинируются, чтобы вносить небольшой вклад в итоговое представление статьи. В результате каждая статья представлена одним вектором из 600 единиц.

Отбор характеристик. Наши результаты, опубликованные в работе Купмана и др. [12], расширяются включением цитирований в качестве дополнительных единиц в семантическую матрицу (подробнее см. [17]). Для рассмотрения роли информации в цитировании в процессе кластеризации будет проведен эксперимент, состоящий из включения или исключения векторов цитирований при вычислении семантических векторов статей (см. (1)). Поэтому для каждой статьи создадим три вектора, один представляет взвешенное среднее арифметическое число всего, за исключением цитирований (т.е. тематических терминов, предметных областей, авторов и журналов, то же, что и Купмана [1]), один – взвешенное среднее арифметическое число только единиц цитирований, и еще один – взвешенное среднее арифметическое число всех типов единиц. В разделе «Проверка консенсуса» будут приведены результаты сравнения.

СТАНДАРТНЫЕ АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ

Следующий шаг, после создания векторов статей, включает определение кластеров статей. Могут применяться разнообразные методы кластеризации. Преимущественно будем экспериментировать на K-Means, поскольку это простой и широкомасштабный метод кластеризации, непосредственно воздействующий на векторные представления статей. Наша цель – проверить, выявляют ли такие семантические представления существенные кластеры.

Методы кластеризации на основе сетей широко применяются наукометрическим сообществом. Поэтому

мы также попытаемся найти решение данной проблемы кластеризации применительно к сети. В ходе дальнейшего процесса обработки таких представлений будем передавать сходства, подсчитанные на основе таких векторных представлений, сходствам сети статей, откуда могут быть выявлены сообщества (кластеры). Наш выбор относительно применения метода выявления сообществ, разработанного в Лёвенском университете [20], объясняется его большой популярностью в использовании наукометрическим сообществом и в то же время адаптацией к моделям данных на основе цитирований. Для нас представляет интерес проверка способности метода Лёвенского университета найти также сообщества, основанные на семантических сходствах статей, а не ссылок между ними.

Теперь кратко опишем эти два стандартных алгоритма и особенности внедрения под наш массив данных.

КЛАСТЕРИЗАЦИЯ С ИСПОЛЬЗОВАНИЕМ K-MEANS

Алгоритм K-Means является самым простым, легким обучающим алгоритмом, способным решать хорошо поставленную проблему кластеризации [21, 22]. Он вполне подходит под выборку большого размера и применим в широком диапазоне прикладных научных областей, в том числе в наукометрии [23].

Учитывая набор точек или наблюдений данных (x_1,x_2, … x_n), где каждая точка данных характеризуется реальным d-пространственным вектором, кластеризация K -means склонна разделять точки данных n на наборы $k(\leq n)$ или кластеры $S= \{S_1, S_2, \ldots S_k\}$, чтобы сумма квадратов внутри кластера становилась минимальной. Другими словами, целью алгоритма K-Means является поиск

$$\arg\min \sum_{i=1}^{k} \sum_{x \in S_i} \left\| x - \mu_i \right\|^2, \qquad (2)$$

где μ_i является центроидом (средним числом) точек в S_i.

Данный алгоритм требует априори определенного числа кластеров. Он начинает работать с первоначального набора k центроидов $m_1^{(1)}, \ldots, m_k^{(1)}$ и измеряет изменения двух этапов [21]:

Этап присвоения: Каждая точка данных присваивается кластеру, среднее число которого выявляет наименьшую сумму квадратов внутри кластера*

$$S_i^{(t)} = \{x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \ \forall j, 1 \leq j \leq k\}, \quad (3)$$

где каждый x_p приписывается к точному одному $S^{(t)}$, даже в случае его приписывания к двум или более.

Этап обновления: Подсчитываются новые средние числа, будущие центроиды точек данных новых кластеров.

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j \ . \qquad (4)$$

Алгоритм завершается при отсутствии новых присвоений, что приводит к образованию (локального) оптимального числа даже без гарантии глобального оптимального числа.

The Mini Batch K-Means [24] представляет вариант алгоритма K-Means, использующий мини-пакеты для сокращения времени вычисления, одновременно стремящийся оптимизировать саму цель функции. Алгоритм берет малые пакеты (выбранные случайно) из набора данных для каждого действия. Затем присваивает кластер каждой точке данных в пакете в зависимости от предыдущих расположений центроидов кластера. Обновляет расположения центроидов кластера на основе новых точек из пакета. Данное обновление является градиентным нисходящим обновлением, которое осуществляется значительно быстрее, чем обычное обновление Batch K-Means.

Использование мини-пакетов значительно сокращает количество вычислений, необходимых для приближения к локальному решению, при этом качество их результатов падает. На практике это различие в качестве может быть очень маленьким [25]. Поэтому принято решение воспользоваться вариантом Mini-Batch K-Means, предоставленным из открытого библиотечного источника по машинному обучению, для построения кластеров статей из массива *Astro,* в котором каждая статья является точкой данных в семантическом пространстве длиной в 600 единиц, как описано в разделе «От семантики единиц к семантике статей».

КЛАСТЕРИЗАЦИЯ С ИСПОЛЬЗОВАНИЕМ МЕТОДА ВЫЯВЛЕНИЯ СООБЩЕСТВ ИЗ ЛЁВЕНСКОГО УНИВЕРСИТЕТА

Рассматриваем каждую статью как узел сети, и здесь есть связь между двумя статьями, когда они очень похожи. Практически, в нашем случае, связываем каждую статью с ее 40 самыми похожими/родственными статьями верхнего ранга на основе сходств по косинусу, вычисленных по их векторному представлению. Это приводит к сети сходств статей, в которой могут быть выявлены кластеры или сообщества. Задача заключается в разделении сети на сообщества тесно связанных узлов, с дисперсными связями или без них между узлами, принадлежащими различным сообществам.

Метод Лёвенского университета [20] – это простой, эффективный и хорошо внедряемый метод по идентификации сообществ в крупных сетях. Он широко используется во многих различных областях, включая наукометрию [26-28]. Воспользуемся им для наглядного представления, насколько он более эффективен в сети сходств, чем в сети на основе цитирований, как сообщила группа из межуниверситетского консорциума по исследованиям и развитию мониторинга ECOOM (Expertisecentrum O&O Monitoring).

Сам по себе метод представляет каскадный алгоритм оптимизации, склонный оптимизировать «модулярность» разделения сети. Модулярность – это масштабная величина от -1 до 1, измеряющая плотность на концах узлов внутри сообществ относительно внешних концов узлов сообществ. Ньюман [29] определяет ее следующим образом:

$$Q = \frac{1}{2|E|} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2|E|} \delta(c_i, c_j) \right], \qquad (5)$$

где $|E|$ – общее число концов в сети, k_i – степень узла i, A_{ij} – элемент матрицы смежности (например, вес на концах узлов i и j), а c_i – сообщество, к которому приписывается узел i, и функция δ равна 1, если c_i=c_j, а иначе 0.

* Поскольку сумма квадратов представляет евклидово расстояние в квадрате, то она является интуитивно «ближайшим» средним числом.

Оптимизация осуществляется в два итерационных этапа. На первом этапе метод ищет «маленькие» сообщества локально оптимизируя модулярность. Каждый узел изначально приписывается к различному сообществу, т.е. сообществ столько же, сколько и узлов. Затем для каждого узла *i* определяется степень модулярности путем подсчета движения i из своего сообщества в сообщество каждого соседа *j* от *i*. После того как это значение подсчитывается для всех связанных с ним сообществ *i*, *i* помещается в сообщество, приводящее к наибольшему росту модулярности. Если невозможно получить положительный результат, тогда *i* остается в своем первоначальном сообществе. Данный процесс многократно и последовательно применяется ко всем узлам до тех пор, пока может происходить улучшение модулярности, после этого первый этап завершается.

На втором этапе метод собирает узлы, принадлежащие к одному и тому же сообществу, и строит новую сеть, узлы которой являются сообществами первого этапа. Затем первый этап может повторно применяться по отношению к этой новой сети. Таким образом, он (метод) итерационно оптимизирует локальные сообщества до достижения максимума глобальной модулярности.

В сравнении с K-Means преимущество использования метода Лёвенского университета заключается в том, что число частей или кластеров определяется самими данными. Как и K-Means, метод Лёвенского университета также является приближенным методом, который по сути не гарантирует глобальный максимум модулярности. Но он широко распространен и часто производит хорошее приближение оптимальных сообществ.

ЭКСПЕРИМЕНТЫ

Упоминавшиеся выше два метода кластеризации применялись к массиву *Astro*, содержащему 111 616 статей по астрономии и астрофизике за 2003-2010 гг. (полное описание массива данных см. в [27]).

Эксперименты с K-Means

Определение K на основе псевдоосновной истины

Оценка результатов кластеризации или выявленных сообществ – сложная проблема. Результаты могут быть предоставлены экспертам, которые решат, является надежным или нет каждый кластер или каждое сообщество. В отличие от этого, основная истина, т.е. кластер обращения или расположение сообщества, может использоваться для измерения того, насколько хорошо решение по кластеризации соответствует основной истине. К сожалению, любой способ слишком трудозатратен, если и возможен в нашем случае.

При применении метода K-Means возникает практическая проблема. Основная истина, или первичное знание о данных, поможет определить один из самых важных параметров K-Means, а именно выбор *k*. Отсутствие основной истины вынуждает нас прагматично определять *k*.

Среднее очертание данных [30] является мерой, которая может применяться для определения *k*. Это очертание измеряет, насколько близко точка данных соотносится с другими точками данных внутри своего кластера и насколько свободно она соотносится с точками данных соседнего кластера, т.е. кластера, среднее расстояние которого от точки данных минимально. Оно (очертание) ранжируется от -1, указывая на ошибочное присвоение, до 1, соответствующего присвоения, в то время как оценки около нуля подразумевают перекрывающиеся кластеры. Подсчитано среднее очертание выборки из 20 тыс. точек данных при *k* от 10 до 100. Как видно на рис. 1, несмотря на легкий подъем оценки среднего очертания все еще находятся в пределах нуля. Это значит, что любые количества кластеров из этого массива данных сильно перекрываются и четкая граница между кластерами не возможна. Это может отражаться по сути на взаимосвязанных научных коммуникациях между различными тематиками. Другой возможной причиной служит то, что эти статьи могут концентрироваться на различных темах области астрофизики, но и могут пользоваться перекрывающимся словарем, что затрудняет установление четкого различия на основе лексической информации.

Поскольку уже имеется пара кластерных решений одного и того же массива данных от разных групп исследователей, у нас была возможность построить *псевдоосновную истину* на основе консенсуса между доступными кластерными решениями. Собраны четыре кластерных решения, а именно CWTS-C5, UMSIO, ECOOM-BC13 и STS-RG. Косвенно все эти четыре решения, имеющие 93 986 261 пару статей, включая 96 072 статьи (86% всего массива), всегда находятся в одних и тех же кластерах. Используем эти общие пары как псевдоосновную истину. Она никак не является реальной основной истиной, а служит консенсусом, которым можно воспользоваться, чтобы мотивировать наше *k* сделать наилучшее предположение.

Табл. 1 показывает, что кластеры CWTS-C5 располагают наименьшим числом пар статей при наличии наибольшей доли, разделяемой с тремя другими решениями. В то время как кластеры STS-RG вполне противоположны: создание более 940 K пар статей, из которых только 10% совместно разделяются с другими. Так происходит в основном потому, что его три самых крупных кластера уже содержат 61% от всего массива данных. Они создают большое число внутрикластерных пар статей. Но поскольку эти статьи находятся в одних и тех же кластерах, общее число совместных пар не слишком сокращается за счет включения кластеров STS-RG. Отметим, что кластеры STS-RG рождаются скорее при использовании иного метода, чем тот, которым пользуются другие три [31]. При отсутствии кластеров STS-RG имеются 140 M общих пар и включенных 100 K статей. Однако, если подтверждается, что их включение не оказывает достаточного влияния на выбор *k,* то принимается решение включить кластеры STS-RG в строительство нашей псевдоосновной истины.

Такое простое сравнение, представленное в табл. 1, также предполагает наличие центрального массива статей, кластерные присвоения которых скорее стабильны, независимо от того, какой метод кластеризации применяется. Поэтому предполагается, что массив из этих 93 млн. *общих пар*, включающий 96 K статей, может использоваться для оценки новых решений в кластеризации, таких, как наши собственные результаты Лёвенского метода.

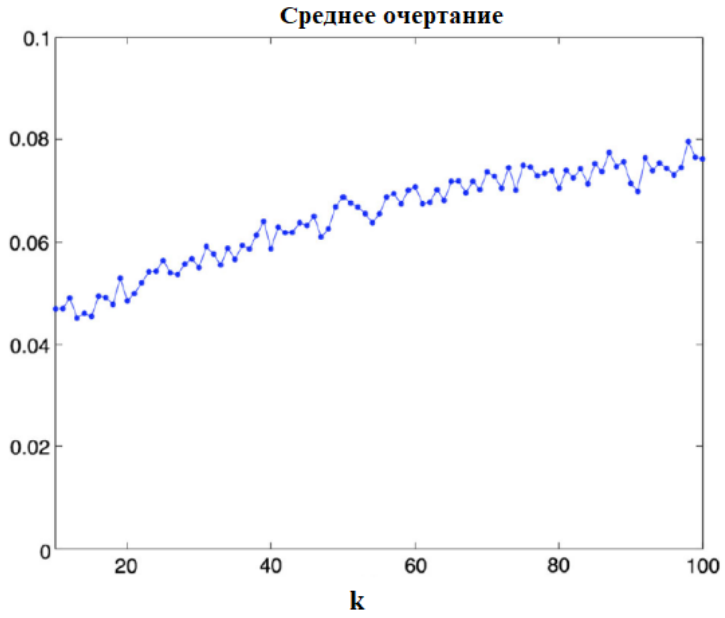

Рис. 1. Среднее очертание более 20 тыс. случайных выборок при k от 10 до 100

Таблица 1

Статистика по четырем решениям кластеризации для псевдоосновной истины

	#Кластер	#Всего пар	Из них общими относительно других являются (%)
CWTS-C5	22	337 151 232	28
UMSIO	22	453 492 311	21
ECOOM-BC 13	13	498 846 580	18
STS-RG	556	940 553 592	10

Исходя из этой псевдоосновной истины, ищется оптимальное k. С одной стороны, эти кластеры k максимально соотносятся с другими четырьмя решениями, т.е. воспроизводством максимальных общих пар. С другой стороны, большие кластеры наказываются, если помещают иррелевантные статьи в одни и те же кластеры. Формально точность (p) и полнота (r) измеряются следующим образом:

$$p = \frac{\#\text{ article pairs in common}}{\#\text{ total produced pairs}},$$

$$r = \frac{\#\text{ article pairs in common}}{\#\text{total shared pairs}}, \quad (6)$$

где #total shared pairs – общее число пар статей в псевдоосновной истине, т.е. 93 млн, # total produced pairs – общее число внутрикластерных пар статей, созданное кластерами k, а # article pairs in common — число пар статей, созданное кластерами k и разделяемое также другими четырьмя решениями. Высокое значение точности (p) означает, что значительная доля созданных пар статей отвечает четырем другим решениям, тогда как высокая полнота (r) указывает на то, что значительная доля общих статей в псевдоосновной истине создается кластерами k. 100% полнота может достигаться помещением всех статей в один кластер, но при этом значительно снизится точность, поскольку подавляющее большинство пар статей внутри кластеров не поддерживается другими четырьмя решениями. Множество небольших кластеров могут улучшать точность, поскольку только они содержат статьи, предположительно находящиеся в одном и том же кластере относительно других четырех решений, тем не менее, многие гипотетически родственные статьи распределяются по различным кластерам, что нарушает полноту.

Чтобы обеспечить баланс между p и r, вычисляется величина $F1$ (https://en.wikipedia.org/wiki/F1_score.), широко используемая сообществом из области информационного поиска:

$$F1 = 2 \times \frac{p \times r}{p + r}. \quad (7)$$

Более того, по отношению к ситуации, подобной с $F1$, склоняемся к гораздо более высокому уровню абстракции, т.е. чем крупнее кластеры, тем это лучше, при условии включения значительного числа иррелевантных статей. Поэтому поощряем бóльший кластер, добавляя параметр среднего размера кластеров в вычисления. Следовательно, наша итоговая оценка массива кластеров рассчитывается следующим образом:

$$adjustedF1 = F \times \log(avgSize). \quad (8)$$

С учетом этого выбираем наилучшее k, которое обеспечивает наивысшую, на основе *adjusted*$F1$, оценку. Как уже упоминалось, позже будем применять эту оценку и к результатам по кластерам из метода Лёвенского университета.

Результаты кластеризации при помощи K-Means

Как упоминалось (см. раздел «От семантики единиц к семантике статей»), строим для каждой статьи три векторных представления: одно усредняет семантические векторы из всех единиц, одно – все единицы, кроме ссылок, одно – только единицы цитирования. Теперь ищем наилучшее *k* для всех трех представлений статей.

Алгоритм K-Means восприимчив относительно вступительного шага, т. е. там, где первоначально располагаются центроиды *k*. Поэтому для *k* от 10 до 60 запускали10 раз алгоритм Mini Batch K-Means, предоставленный библиотекой по языку программирования Python в организации scikit-learn (http://scikit-learn.org/.), и выбирали наилучшее решение с минимальной суммой подсчета квадратов. Затем воспользовались величиной *adjusted*F1, чтобы оценить наши решения, противоположные псевдоосновной истине. Оценки величины изображены относительно *k* на рис. 2.

Если воспользоваться всеми единицами информации, то эта оценка возрастает до тех пор, пока *k* равняется примерно 30, затем снижается, выдавая наивысшую оценку при *k*= 31. Поэтому выбрали наилучшим *k*=31 при условии использования всех единиц для семантического представления статей. Аналогично подтвердилось наилучшее *k*=28 при использовании только цитирований и наилучшее *k*=24, если цитирования не используются. Однако на рис. 2b, при отсутствии использования цитирований, отражается наличие гораздо бóльших флуктуаций, когда наблюдается подобное движение кривой вверх-вниз. Тогда как в случае с использования тольлко цитирований такая кривая едва ли различима.

Табл. 2 подробно показывает качественные оценки этих трех решений кластеризации на основе псевдоосновной истины. Последний столбец таблицы представляет средние опосредованные взаимосвязанные информационные оценки (Adjusted Mutual Information scores – AMI) [32], лежащие между этим и четырьмя другими решениями, а именно CWTS-C5, UMSI0, ECOOM-BC13 и STS-RG. В случае использования только цитирований видно, что итоговые кластеры больше соответствуют другим решениям по кластеризации, чем тем, которые не используют цитирования и чья оценка также является самой низкой. Это не удивительно, поскольку другие решения по кластеризации едва ли полагаются на информацию цитирования. Поэтому, даже при различии способов использования цитирования, информация цитирования все еще вносит сюда достаточный консенсус. Использование всех единиц для повторного представления статей имеет наивысшую оценку *adjusted*F1 и в значительной степени соотносится с другими решениями.

Табл. 3 демонстрирует опосредованные взаимосвязанные информационные оценки (AMI) в отношении этих трех решений и кластеров на основе метода Лёвенского университета. Опять кластеры, основанные только на цитированиях, соотносятся с результатами Лёвенского метода почти в такой же степени, как и кластеры, использующие все единицы. В соответствии с этими измерениями, было принято решение использовать все единицы в качестве окончательной выборки свойств и сохранить эти кластеры (31 кластер) как итоговые результаты метода K-Means, под названием OCLC-31. Масштаб распределения этих кластеров представлен на рис 3a.

Выявление сообществ с использованием метода Лёвенского университета

В отличие от стандартного Лёвенского метода, базирующегося на сети цитирования, воспользуемся Лёвенским методом на основе сети семантических сходств, где каждый узел представляет статью и существует грань между двумя статьями, если они в значительной степени похожи/родственны. Учитывая эксперименты с алгоритмом K-Means, снова используем все единицы для вычисления семантического представления статей. Для каждой статьи подсчитывались 40 самых схожих статей верхнего ранга, чьи значения сходства выше определенного порога (в данном случае 0,6), и рассматривалась связь по принципу родства этой статьи с ее ближайшими 40 объектами верхнего ранга. Раз каждая статья связана со своими одноранговыми объектами, формируется сеть сходств, которая в таком случае больше ориентирована на применение Левёнского метода по выявлению сообществ или кластеров этой сети.

Используем пакет программ networkx Библиотеки языка программирования python (https://networkx.github.io/.) и модуль пакета программ python по определению сообществ, предназначенный для выявления сообществ при использовании Лёвенского метода (http://perso.crans.org/aynaud/communities/.). Это отражается в 32 самых лучших разделах (кластерах), именуемых OCLC-Louvain, самое крупное из которых содержит 9 646 статей, наименьшее – 86, а общее среднее – 3 488 статей (рис. 3b). Контраст его качества по сравнению с псевдоосновной истиной приводится в табл. 2.

Таблица 2

Сравнение качества при отборе различных характеристик

	# кластеры	r	p	*f*1	*Adjusted*F1	AMI
Без цитирования	24	0,53	0,17	0,26	2,16	0,44
Только цитирование	28	0,58	0,18	0,28	2,29	0,47
Все единицы	31	0,56	0,23	0,33	2,69	0,47
OCLC-Louvain	32	0,61	0,21	0,31	2,57	0,49

Таблица 3

Опосредованная взаимосвязанная информация относительно решений

	Без цитирования	Только цитирования	Все единицы	Louvain
Без цитирования	1,00	0,59	0,63	0,56
Только цитирование		1,00	0,69	0,65
Все единицы			1,00	0,67
OCLC-Louvain				1,00

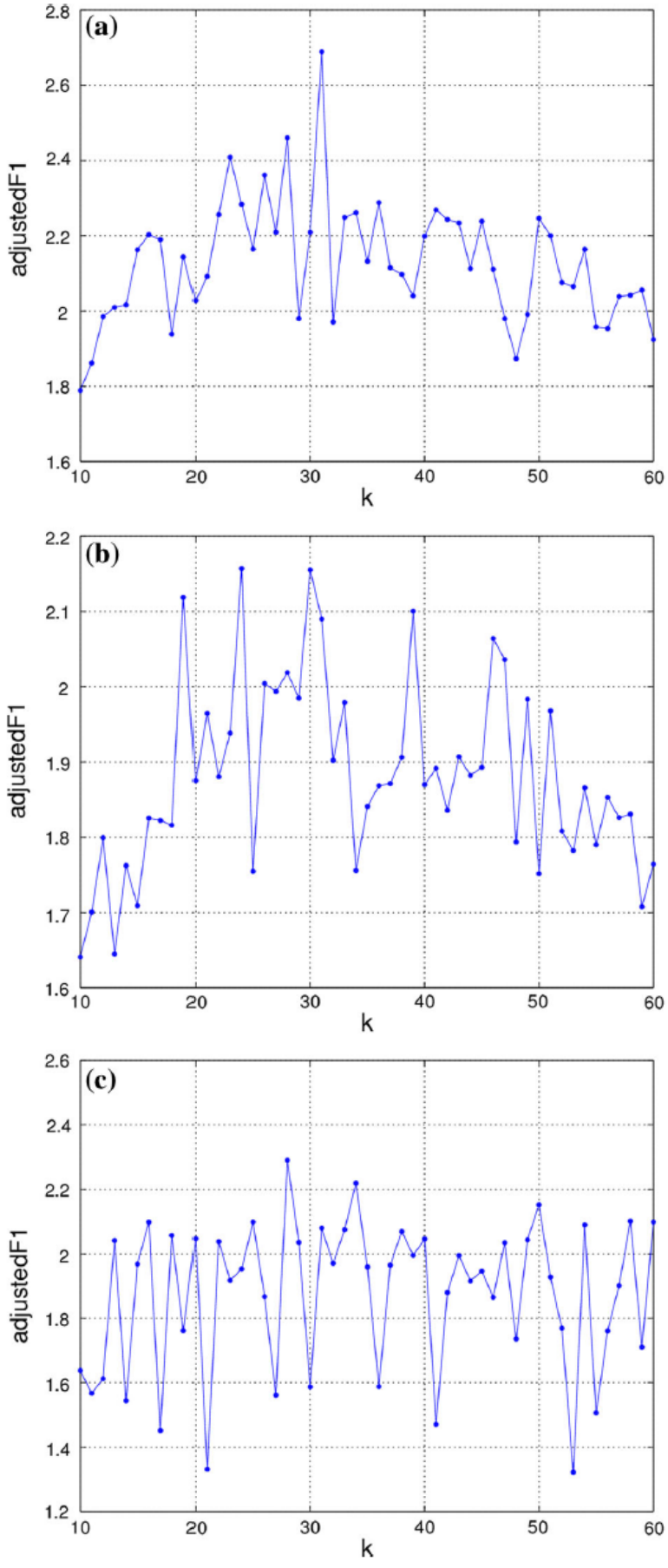

Рис. 2. Поиск наилучшего k на основе *adjustedF*1 с использованием различных наборов объектов (a - все объекты, b - без цитирования, c - только цитирования)

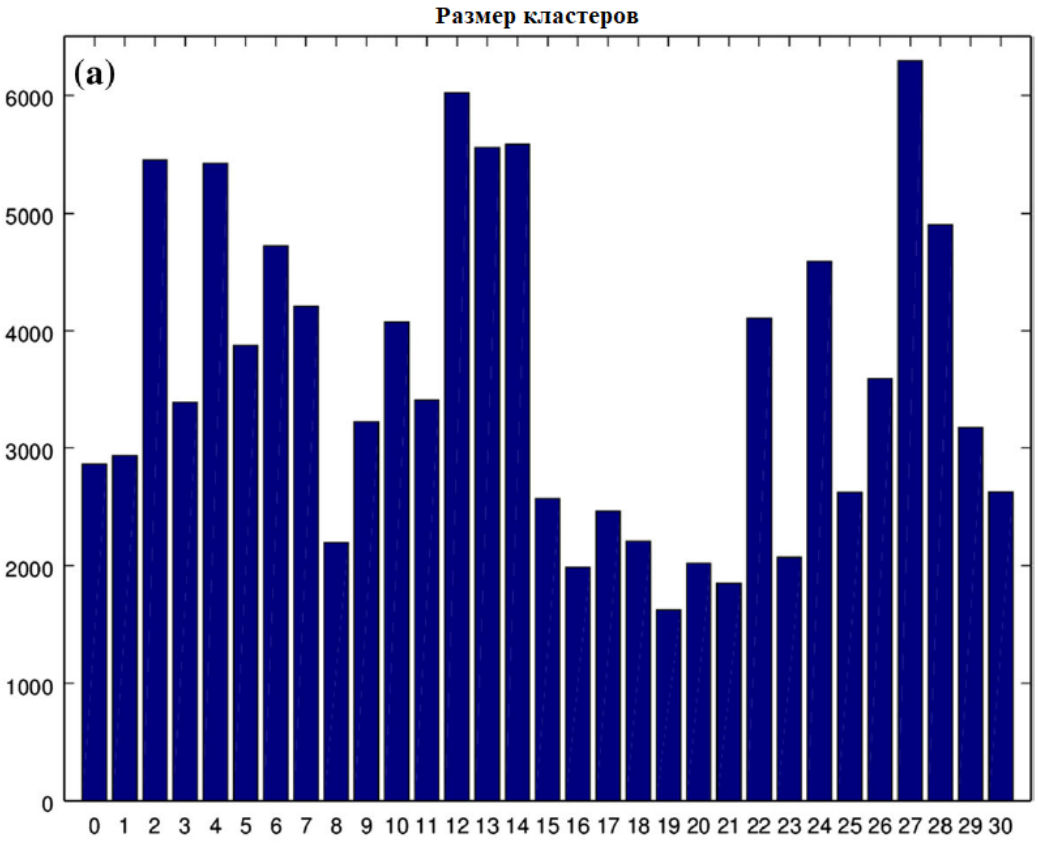

Рис. 3a. Масштаб распределения двух наших решений по кластеризации (метод OCLC-31)

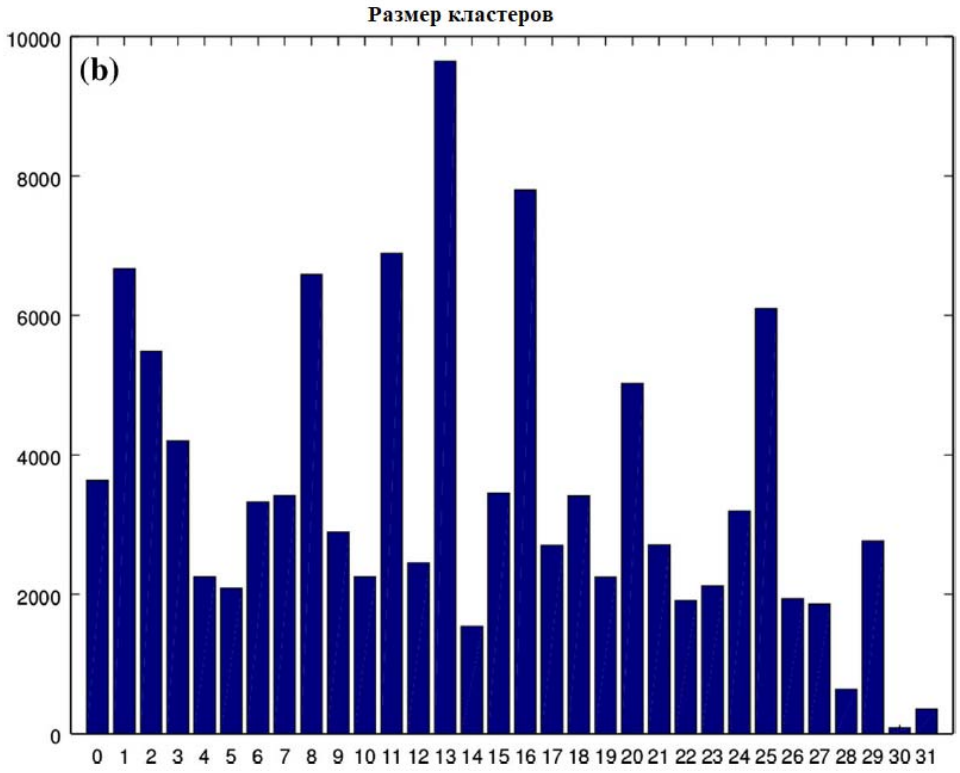

Рис. 3b. Масштаб распределения двух наших решений по кластеризации (метод OCLC-Louvain)

Таблица 4

Проверка консенсуса с использованием опосредованной взаимосвязанной информации (AMI)

	sr	c	u	eb	en	ok	ol
STS-RG (sr)	1,00	0,44	0,46	0,43	0,34	0,41	0,42
CWTS-C5 (c)		1,00	0,77	0,47	0,39	0,56	0,61
UMSIO (u)			1,00	0,47	0,38	0,51	0,55
ECOOM-BC 13 (eb)				1,00	0,46	0,46	0,46
ECOOM-NLP 11 (en)					1,00	0,41	0,39
OCLC-31 (ok)						1,00	0,67
OCLC-Louvain (ol)							1,00
Average AMI	0,42	0,54	0,52	0,46	0,40	0,50	0,52

Лёвенские кластеры действуют подобно кластерам K-Means и действительно больше соотносятся с другими решениями по кластеризации. Однако недостатком в использовании Лёвенского метода является то, что он не распространяется на более крупный массив данных, поскольку сеть сходств представляется дорогостоящей для разработки, использующей метрику расстояния, даже если сам по себе Лёвенский алгоритм относительно масштабируем.

ПРОВЕРКА КОНСЕНСУСА

Теперь можно использовать стандартные измерения консенсуса, такие как опосредованная взаимосвязанная информация (AMI) [33], чтобы проверить, насколько эти два решения кластеризации соотносятся друг с другом. Табл 4. представляет оценку консенсуса в этих двух решениях, другие пять решений описываются в специальном выпуске журнала Scientometrics, Special Issue of Scientometrics (2017) [*]. Последняя строка таблицы приводит среднюю AMI относительно одного решения кластеризации и всех других решений.

Такие цифры предполагают, что данная модель данных оказывает больше влияния на это решение, чем выбранный алгоритм, поскольку OCLC-31 и OCLC-Louvain имеют второе самое высокое значение в рамках соответствия друг другу (наивысшая форма консенсуса определена между CWTS-C5 и UMSIO, которые также используют ту же самую модель данных). При сравнении STS-RG и ECOOM решений даже с иной моделью данных, наши решения демонстрируют, что они остаются высокосопоставимыми. Более подробное описание сравнения см. в [31].

[*] CWTS-C5 и UMSIO являются кластерными решениями, созданными двумя различными методами, Infomap и the Smart Local Moving Algorithm (SLMA) соответственно, которые непосредственно предназначены для сети цитирования статей. Два решения по кластеризации консорциума ECOOM порождаются применением Лёвенского метода относительно поиска сообществ среди статей с библиографическим сочетанием, в которых ECOOM-NLP11 также содержит информацию по ключевым словам. Кластеры STS-RG генерируются первым проектированием небольшой части массива Astro для полной БД Scopus и сбором их присвоений к кластерам после того, как статьи Scopus целиком кластеризуются с применением SLMA непосредственно в сети цитирования. Более точный подсчет можно посмотреть в [31].

ЗАКЛЮЧЕНИЕ

В статье рассматриваются два метода кластеризации для определения кластеров в массиве данных Astro. В отличие от других методов, представленных в специальном выпуске журнала Scientometrics, Special Issue of Scientometrics (2017), здесь построено семантическое представление статей и предпринята попытка выявить кластеры на основе их семантического сходства. Приведены технические подробности и путь выработки решения по отношению к двум решениям кластеризации, одно основано на алгоритме K-Means, а другое – на методе выявления сообществ, разработанном в Лёвенском университете.

Семантическое представление статей базируется на семантической матрице, в создание которой вносят вклад эти статьи. Каждая единица (тематический термин, предметная область, автор, журнал, ссылка) представлена своим лексическим окружением, извлеченным и значительно сокращенным по сравнению с массивом. Были интегрированы семантические векторы всех единиц, включенных в одну статью в качестве её представления. Эксперименты показали, что такая интеграция семантики отдельных единиц отражает семантику статей, а результаты кластеризации конкурентны относительно других решений кластеризации, основанных главным образом на информации по цитированию.

Хотелось бы подчеркнуть, что два метода кластеризации, рассмотренные здесь, представляют только две возможности, проверенные на таком семантическом представлении. K-Means широко применим и выдает результаты, в большой степени согласующиеся с другими решениями. Преимущество заключается в том, что он применим при отсутствии информации о цитировании. Он может стать первым шагом в кластеризации по разделению статей на основе их лексической информации до того, как произведено глубокое погружение в релевантные подмножества с помощью более гибких и сложных методов кластеризации.

Благодарность. Работа частично профинансирована Европейской кооперацией в области научных и технологических исследований – COST Action TD1210 Knowescape. Хотелось бы поблагодарить Йохана Глезера и Андреу Шарнхорст за подробные замечания относительно ранних редакций данного текста, а также внутреннего рецензента Майкла Хинца и анонимных внешних рецензентов за их важные замечания и предложения.

ЛИТЕРАТУРА

1. *Koopman R., Wang S., Scharnhorst A.* Contextualization of topics—browsing through terms, authors, journals and cluster allocations/ A. A. Salah, Y. Tonta Y., A. A. A. Salah, C. R. Sugimoto, U. Al, (Eds.), Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey, 29 June to 3 July, 2015. — Bogaziçi University Printhouse, 2015. — http://www.issi2015.org/files/downloads/all-papers/1042.pdf.

2. *Bruckner E., Ebeling W., Scharnhorst A.* The application of evolution models in scientometrics// Scientometrics. — 1990. — Vol. 18, No. (1–2). — P. 21–41. — doi:10.1007/BF02019160.

3. *Sugimoto C. R., Weingart S.* The kaleidoscope of disciplinarity//Journal of Documentation. — 2015. — Vol. 71, No. 4. —P. 775–794. — doi:10.1108/JD-06-2014-0082. — http://www.scopus.com/inward/record.url?eid=2-s2.0-84933503812&partnerID=tZOtx3y1.

4. *Gläser J., Glänzel W., Scharnhorst A.* Same data: Different results? Towards a comparative approach to the identification of thematic structures in science // Scientometrics. — 2017. — doi:10.1007/s11192-017-2296-z.

5. *Garfield E.* Citation indexing—Its theory and application in science, technology and humanities. — Philadelphia: ISI Press, 1983.

6. *Small H.* Co-citation in the scientific literature: A new measure of the relationship between two documents// Journal of the American Society for Information Science.— 1973. —Vol. 24.— P. 265–269.

7. *Glänzel W., Czerwon H. J.* A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level // Scientometrics. — 1996. —Vol. 37. —P. 195–221.

8. *Leydesdorff L.* Words and co-words as indicators of intellectual organization//Research Policy. — 1989. — Vol. 18, No. 4. — P. 209–223. — doi:10.1016/0048-7333(89)90016-4.

9. *Boyack K. W., Small H., Klavans R.* Improving the accuracy of co-citation clustering using full text// Journal of the American Society for Information Science and Technology. — 2013. — Vol. 64, No. 9. — P. 1759–1767. — doi:10.1002/asi.22896.

10. *Leydesdorff L., Hellsten I.* Measuring the meaning of words in contexts: An automated analysis of controversies about 'monarch butterflies', 'frankenfoods', and 'stem cells' // Scientometrics. — 2003. — Vol. 67, No. 2. — P. 231–258.

11. *Rip A., Courtial J. P.* Co-word maps of biotechnology: An example of cognitive scientometrics// Scientometrics. — 1984. — Vol. 6, No. 6. — P. 381–400.

12. *Koopman R., Wang S., Scharnhorst A., & Englebienne G.* Ariadne's thread: Interactive navigation in a world of networked information/ B. Begole, J. Kim, K. Inkpen, W. Woo (Eds.), Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, Seoul, CHI 2015 Extended Abstracts, Republic of Korea, April 18–23, 2015, pp. 1833–1838. — ACM, 2015. — doi:10.1145/2702613.2732781.

13. *Furnas G. W., Landauer T. K., Gomez L. M., Dumais S. T.* Statistical semantics: Analysis of the potential performance of keyword information systems// Bell System Technical Journal. —1983. —Vol. 62, No.6. — P. 1753-1806. — doi:10.1002/j.1538-7305.1983.tb03513.x.

14. *Weaver W.* Translation/ W. Locke, D. Booth (Eds.), Machine translation of languages (pp. 15–23). — Cambridge, Massachusetts: MIT Press, 1955.

15. *Firth J.R.* A synopsis of linguistic theory 1930–1955 // Studies in Linguistic Analysis. —1957.—P. 1–32.

16. *Harris Z.* Distributional structure// Word. —1954. — Vol. 10, No. 23. — P. 146-162.

17. *Sahlgren M.* The distributional hypothesis// Rivista di Linguistica. —2008. — Vol. 20, No.1. — P. 33-53.

18. *Koopman R., Wang S., Scharnhorst A.* Contextualization of topics—browsing through the universe of bibliographic information/ J. Gläser, A. Scharnhorst, W. Glänzel (Eds.), Same data—different results? Towards a comparative approach to the identification of thematic structures in science// Special Issue of Scientometrics, 2017.

19. *Achlioptas D.* Database-friendly random projections: Johnson–Lindenstrauss with binary coins//Journal of Computer and System Sciences. —2003. —Vol. 66, No. 4. — P. 671-687.— doi: 10.1016/S0022-0000(03)00025-4.

20. *Johnson W., Lindenstrauss J.* Extensions of Lipschitz mappings into a Hilbert space// Contemporary Mathematics. — 1984. — Vol. 26. — P. 189–206.

21. *Blondel V. D., Guillaume J. L., Lambiotte R., Lefebvre E.* Fast unfolding of communities in large networks// Journal of Statistical Mechanics: Theory and Experiment. — 2008. — Vol. 10. — P10008. (12pp).

22. *MacKay D.* Information theory, inference and learning algorithms, chap. Chapter 20. An Example inference task: Clustering, p. 284-292. — Cambridge University Press, 2003.

23. *Witten I. H., Frank E., Hall M. A.* Data mining: Practical machine learning tools and techniques, third edition edn. The Morgan Kaufmann series in data management systems. — Burlington: Morgan Kaufmann, 2011.

24. *Boyack K. W., Klavans R., Börner K.* Mapping the backbone of science// Scientometrics. —2005. —Vol. 64, No. 3. — P. 351–374.

25. *Sculley D.* Web scale k-means clustering// Proceedings of the 19th International Conference on World Wide Web, p. 1177-1178. — Raleigh, NC, USA, 2016.

26. *Béjar J.* K-means vs mini batch k-means: A comparison. — Tech. rep., Universitat Politècnica de Catalunya, 2013. http://upcommons.upc.edu/bitstream/handle/ 2117/23414/R13-8.pdf.

27. *Zhang L., Liu X., Janssens F., Liang L., Glänzel W.* Subject clustering analysis based on ISI category classification// Journal of Informetrics. — 2010. — Vol. 4, No. 2. — P. 185–193. — doi:10.1016/j.joi.2009.11.005. —http://www.sciencedirect.com/science/article/pii/S1751157709000832.

28. *Glänzel W., Thijs B.* Using hybrid methods and 'core documents' for the representation of clusters and topics. The astronomy dataset// Scientometrics. — 2017. — doi:10.1007/s11192-017-2301-6.

29. *Zhang L., Liu X., Janssens F., Liang L., Glänzel W.* Subject clustering analysis based on ISI category classification// Journal of Informetrics. —2010, —Vol. 4, No. 2. —P. 185–193. — doi:10.1016/j.joi.2009.11.005. — http://www.sciencedirect.com/science/article/pii/S1751157709000832/The ASIS&ISSI "metrics" pre-conference seminar and the Global Alliance

30. *Newman M. E.* Modularity and community structure in networks// Proc Natl Acad Sci USA. —2006. — Vol. 103, No. 23. — P. 8577–8582. — doi:10.1073/pnas.0601602103. — http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=retrieve&db=pubmed&list_uids=16723398&dopt=AbstractPlus.

31. *Rousseeuw P.* Silhouettes: A graphical aid to the interpretation and validation of cluster analysis// Journal of Computational and Applied Mathematics. —1987. — Vol. 20, No. 1. — P. 53–65. — doi:10.1016/0377-0427(87)90125-7.

32. *Velden T., Boyack K., van Eck N., Glänzel W., Gläser J., Havemann F., Heinz M., Koopman R., Scharnhorst A., Thijs B., Wang S.* Comparison of topic extraction approaches and their results// J. Gläser, A. Scharnhorst, W. Glänzel (Eds.), Same data—different results? Towards a comparative approach to the identification of thematic structures in science// Special Issue of Scientometrics, 2017.

33. *Vinh N. X., Epps J., Bailey J.* Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance// Journal of Machine Learning Research. — 2010.— Vol. 11.— P. 2837-2854.

Вычисление семантического сходства научных статей с использованием тематического события и онтологии*

Лиу МИН
(Liu MIING)

Лан БО
(Lang BO)

Жу ЦЗЕПЕН
(Gu ZEPENG)

Государственная профильная лаборатория по разработке сред программирования, Пекинский университет, г. Пекин, Китай

Установление семантического сходства академических документов представляется важным для многих задач, таких как выявление плагиата, автоматизированное технологическое исследование и семантический поиск. Сегодняшние исследования большей частью посвящены семантическому сходству понятий, предложений и коротких фрагментов текста. Тем не менее, семантическое соответствие на уровне документов все еще находится на поверхностном уровне понимания и основывается на статистической информации, отклоняя структуру статей и глобальные семантические значения, что может вызвать отклонение в понимании документа. В статье с позиции нового метода рассматривается проблема семантического сходства документального уровня для академической литературы. Академические статьи представляются как тематические события, использующие многочисленные информационные профили, такие как цели исследования, методологии и предметные области, с целью полного описания научной работы и вычисления близости тематических событий, основанных на онтологии предметной области, чтобы получить семантическое сходство статей. Эксперименты показывают, что наш подход значительно эффективнее по сравнению с известными сегодня методами.

ВВЕДЕНИЕ

Соответствие семантики текста находит широкое применение во многих сферах, таких как машинный перевод, автоматизированное получение ответа на вопрос и интеллектуальный поиск. Оно также имеет важное значение для выявления плагиата, автоматизированного технологического исследования и рекомендации относительно цитирования и анализа направлений исследований в научной предметной сфере. Проблема семантики текста, такая как семантика слов и предложений, получает в последнее время широкое внимание. Тем не менее, мало исследований занимается семантическим соответствием на уровне документов из-за его сложности. Объемные документы обычно имеют усложненную структуру и несут большой массив информации, что затрудняет измерение их семантического сходства и, насколько нам известно, сегодня даже не существует публично доступного массива данных.

Крупные единицы текста состоят из единиц меньшего размера. Семантика большого документа может быть получена путем сочетания семантики текстовых единиц меньшего размера. В последнее время многие исследования придерживаются этой идеи для получения семантического сходства текстовых единиц большего размера. Например, семантическое сходство предложений можно получить из интеграции семантического родства между парами слов из двух предложений [1,2]. Кроме лексической семантики также принимаются во внимание особенности на уровне предложения в глобальном масштабе для получения семантического сходства предложений [1-8]. Однако эти исследования фокусируются только на коротких текстах из-за лексической семантики и особенностей на уровне предложений, которые еще далеки от возможности семантического сходства на уровне документов.

Исследование, концентрирующееся на семантическом сходстве документов, – относительно редкое явле-

* Перевод Ming L., Bo L., Zepeng G. Calculating semantic similarity between academic articles using topic event and ontology. — https://arxiv.org/ftp/arxiv/papers/1711/1711.11508.pdf

ние. Действующие методы сходства на уровне документов в основном скорее опираются на информационный поиск на поверхностном уровне, чем на уровне понимания семантики. Обычные метрики сходства [9-11], например, модель векторного пространства (VSM – vector space model) [12], определяют сходство документов через статистику или морфологию слов, не учитывая структуру документов и значения слов, содержащихся в документах. Модель векторного пространства воспринимает каждый документ как набор слов и измеряет сходства документов, преимущественно основываясь на присутствии слов. Например, имеется два фрагмента текста: «Джек одолжил книгу у учителя» и «Учитель одолжил книгу у Джека». Модель векторного пространства воспринимает оба текста как равные, а по существу они имеют противоположные значения. Латентное распределение Дирихле (Latent Dirichlet Allocation, LDA) [13] может быть адаптировано под детальный анализ документов, основанный на разнице в распределениях тем по документам, который пригодится для измерения семантики на уровне документов. Также имеются исследования [14-19], которые ищут возможность дополнить представление документов внешним знанием, что обогащает контент, добавляя релевантные термины из источников знания. Но подобные методы все еще не преодолели такие проблемы, как сложность вычислений и непрозрачность представления.

Большие документы часто меняют тему в изложении вопроса и акценты внимания, это затрудняет определение семантики их ядра. Тем не менее, есть мнение, что эти темы в любом документе согласованы, и такие корреляции получаются путем всестороннего анализа по разным факторам документа. Поэтому представляем проблему семантики ядра в каждом документе как событие, называемое тематическим событием (Topic Event – TE). TE выглядит структурированным рефератом, извлеченным из каждого документа, содержащим существенные ключевые элементы данного документа.

Ядром семантики научной статьи является авторская исследовательская работа. Формируем TE, основанное на структуре статьи и использовании многих областей в информации, таких как цели исследования (research targets), методологии (methodologies), ключевые слова (keywords) и предметные области (domains), которые могут полно описывать различные фасеты исследовательской работы. Поэтому, семантическое сходство научных статей можно измерять подобием TE (рис. 1). В целях получения высокой точности подсчетов подобия TE разработаем и будем применять также онтологию стиля стиля и онтологию предметной области научной дисциплины. Для придания большей практической ценности нашему подходу предлагаем метод автоматизированного формирования TE. В целях проверки эффективности создаем систему оценок с помощью аннотации вручную, использующую систему антологической сети Ассоциации по вычислительной лингвистике (Association of Computational Linguistics – ACL, США) [20]. Эксперименты демонстрируют, что наши методы максимально эффективны, а результаты более привлекательны для восприятия людьми. Делаем вывод – основные преимущества нашей работы следующие:

– предлагается идея формирования тематических событий в виде семантических представлений объемных документов, а также излагается общий метод вычисления подобия тематических событий;

– развиваются и строятся онтологии стиля исследования и предметной области для научных статей, эти два источника знания значительно облегчают процедуры извлечения семантики и измерения подобия тематического события;

– предоставляется способ построения автоматизированного тематического события на основе онтологии без указания данных определенной области, он применяется в построении тематических событий в документах по вычислительной лингвистике;

– впервые вводится система семантического соответствия документов с подробными аннотациями, она выступает в роли основополагающей истины в оценке соответствия семантики исследований на уровне документов.

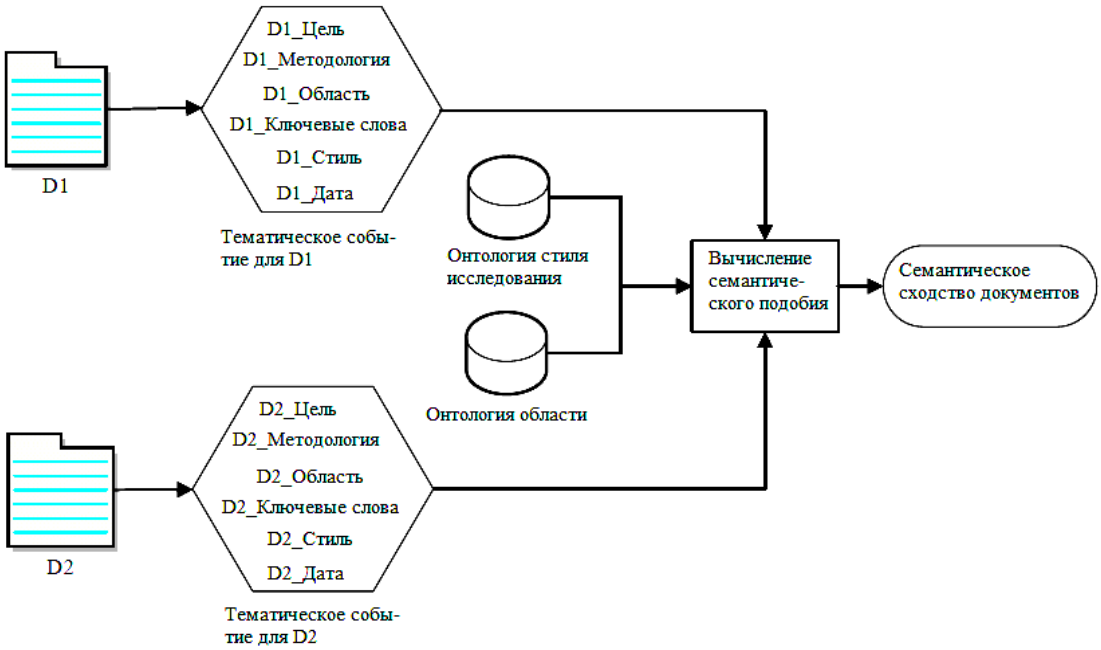

Рис. 1. Соответствие семантики документа на основе TE

Далее статья организована следующим образом. В разделе «Связанная работа» описываются аналогичные работы. В разделе «Тематические события» кратко излагается тематическое событие в научной области. В разделе «Автоматизированное построение тематического события» описывается сам процесс. В разделе «Вычисление подобия тематического события» приводится способ вычисления подобия тематического события. В разделе «Оценка эксперимента» сначала строятся онтология предметной области и система оценок в области вычислительной лингвистики, а затем даются оценки эксперимента. Завершается статья разделом «Заключение».

СВЯЗАННАЯ РАБОТА

Исследование семантического сходства на уровне документа представляет собой новую зарождающуюся область, идея которой может проясниться из самого понятия и методов семантического сходства небольших текстов. Семантика на уровне понятия служит основой понимания семантики документа, а сходство на уровне небольших текстов является самой актуальной областью, проливающей свет на семантическое сходство на уровне документов.

Вообще, семантическое сходство на уровне документа может измеряться методами на основе знания или на основе массива. Методы на основе знания в основном используют расстояние между понятиями в источниках знания, чтобы продемонстрировать их семантическое сходство [21, 22]. Лин [23] и Ресник [24] измеряли такое семантическое сходство соотношением информационного содержания, по крайней мере, двух понятий из общей подсуммы этих понятий. Вместо непосредственного изучения графического расстояния в источниках знания ряд исследований [25, 26] воспроизводил векторы понятий в соответствии с набором свойств онтологии для понятия семантического сходства. Методы на основе массива предполагают, что слова с одинаковым значением часто встречаются в схожих контекстах, латентный семантический анализ (Latent Semantic Analysis, LSA) [27] представляет слова в виде компактных векторов через сингулярное разрушение в поле матрицы, а глобальная векторная модель для представления слов GloVe [28] снизила стоимость вычисления за счет прямого направления отличных от нуля элементов в поле матрицы. Тэрни [29] измерял семантическое сходство с помощью точечной взаимоисключающей информации. Web можно также расценивать как огромный массив, а метод автоматизированного извлечения слов Google Distance [30] использует ряд слов, совместно встречающихся на сетевых страницах, для семантического сходства понятия. Миколов [31] предложил подход вовлечения слов с помощью нейронной сети для охвата семантики слов, встречающихся в окне с фиксированным размером.

Проблема определения семантического сходства небольших текстов заключается в разработке перехода от семантики на уровне слов к семантике уровня небольшого текста. Общий подход представляет использование взвешенной суммы семантического двусловного сходства (слово в слово), а авторы работ [1,2] пользовались методом избыточного присвоения для формирования семантического сходства предложения. Бэйни и др. [3] измеряли семантическое сходство сниппетов текста, передающих представление о знании. Д. Реймидж с соавторами [6] построил граф понятия, используя слова из каждого сниппета, затем измерил это сходство двух графов понятий. Недавно конференция SemEval рассмотрела задачу семантического сходства текста в условиях проявления особого интереса к семантике небольшого текста, модели регрессии для прогнозирования оценки сходства были приняты большинством участников [5,8], также изучались лексические и синтаксические особенности. Модель охвата тем документа – Paragraph vector, аналогичная вовлечению слов, также предлагается с нейронными сетями, чтобы измерять семантическое сходство небольших текстов.

Исследование, непосредственно занимающееся семантическим сходством документов, встречается нечасто и недостаточно хорошо разработано. Традиционные исследования относительно сходства, такие как TF-IDF [12], преобразуют документы в векторы через подсчет слов и измеряют сходство документов через сходство векторов. Научные статьи можно рассматривать как полуструктурированные тексты, содержащие, помимо основного текста, многоструктурные аннотации. Сходство научных статей может измеряться при помощи аннотированной информации. Мартин и др. [33] объединяли структурную информацию, такую как авторы и ключевые слова, с широко применяемыми измерениями сходства научных статей на основе текста. Авторы работ [9-11] рассматривают статьи и ссылки внутри их в качестве информационной сети. Сходство статей становится сходством двух единиц в информационной сети. К сожалению, упоминавшиеся ранее обычные методы сходства направлены скорее на индексирование документов на поверхностном уровне, чем на понимание семантики на уровне документа.

Некоторые исследования склонны добавлять внешнее знание в получение семантического представления документов. Работы [14,15,17] обогатили контент включением релевантных терминов из источников знания, что ориентировано на улучшение качества результатов кластеризации документов. Источники, такие как [18,19], извлекали отношения «троек» из источника документов и добавляли отношения элементов из вспомогательного (дополнительного) знания для формирования тройного графа в целях улучшения качества документа. Шумахер и Понцетто [16] предложили семантическую модель на основе графа для представления контента документа, которая добавила знание к представлению документа за счет привязывания единиц документа к БЗ (базе знаний) DBpedia. Подобные методы требуют точных отношений между объектами и создают обогащенную знанием модель документа, а семантическое сходство подсчитывается при помощи коэффициента протяженности графа. Однако упомянутым методам не хватает интерпретации. Более того, есть редкие объекты, такие как люди, организации и наименования мест внутри контента научных статей, которые затрудняют использование подобных методов.

Большие документы, такие как научные статьи, как правило, имеют несколько центров концентрации внимания и огромное множество слов. LDA [13] получает семантическое сходство различных понятий через темы и рассматривает каждый документ как распределение по отношению к набору тем. Таким образом, LDA может применяться в семантическом анализе объемных документов. Мухаммад Рафи [17] определял коэффициент сходства на основе тематических представлений в вопросах кластеризации документов. Документы преобразуются в тематические представления на основе кодированного знания, а сходство в паре документов выглядит как корреляция между общими моделями представлений.

М. Чжан с соавторами [34] обогатил документ скрытыми темами из внешнего массива и измерял сходство документа в вопросах классификации текста с помощью сходства в распределениях тем. Кроме рассмотрения тематической модели исследователи в работах [35,36] измеряли семантическое сходство документов на основе дивергенции распределения тем, которую можно рассчитать через расстояние Кульбака-Лейблера, а метод на основе LDA лучше подходит для решения вопросов относительно семантического сходства научных статей.

ТЕМАТИЧЕСКОЕ СОБЫТИЕ

Что можно использовать для передачи основной семантики объемного документа? Эта задача сложная и не будет выгодна с точки зрения простой аккумуляции развернутого понятия семантики. Чтобы получить глобальное понимание документа, необходимо извлечь ключевую информацию из огромного количества слов и сформировать ядро семантики документа.

Структура тематического события

В академической среде статьи используются для передачи прогресса в исследованиях. Большая часть научных статей имеет нормативные форматы и регулярные структуры, а исследовательская работа включает одни и те же профили, которые можно изобразить в виде универсальной структуры тематического события.

Наиболее существенными разделами научной статьи являются ее цели, методы и результаты, передающие основную информацию исследования. Ключевые слова могут легко представлять ядро семантики и знакомить читателей с общим пониманием исследования. Область научных проблем указывает направление научно-исследовательской деятельности, а тип научно-исследовательской работы отражает стиль научного исследования и его особенности. Даты публикации свидетельствуют о различных этапах научно-исследовательского процесса. Считаем вышеуказанные факторы первоначальными элементами тематического события, структура тематического события изображена на рис. 2.

Элементы, отмеченные *, важны и существенны, замены для них не существует, тогда как оставшиеся элементы оптимальны. *Eid* и *Did* являются основными идентификациями тематического события и соответствующей статьи. *Style* (стиль) отражает тип исследования, а такие элементы как *Domain* (область), *Target* (цель), *Methodology* (методология), *keywords* (ключевые слова) являются терминологиями, извлеченными из статей, тогда как элементы *Conclusion* (заключение), *Background* (предпосылки), *Performance* (эффективность) и *Forecast* (прогноз) служат основными субъектами в статьях.

Тематическое событие={

* ***Eid***: *Непосредственный определитель документа для тематического события;*

* ***Did***: *Непосредственный определитель документа для соответствующей статьи;*

* ***Target***: *Сама цель исследования;*

* ***Methodology***: *Средства, используемые в исследовании;*

* ***Domain***: *Непосредственная категория научной сферы;*

* ***Style***: *Определенный тип исследовательского метода;*

****Keywords***: *Ключевые слова, характеризующие статью;*

****Date:*** *Фиксированная дата публикации;*

Metadata:

{автор, организация, цитирование, публикация, издание}

Name: *Оптимальное название разрабатываемой системы;*

Object: *Полезные объекты для исследования;*

Tools: *Оптимальные для исследования инструменты;*

Feature: *Отличительные особенности исследовательского метода;*

Conclusion: *Важные заключительные предложения;*

Background: *Важные поддерживающие проекты;*

Forecast: *Важные предложения для дальнейшего исследования;*

Performance: *Эффективность исследования;*

Dataset: *Оптимальные массивы данных, используемые в исследовании;*

}

Рис. 2. Структура тематического события

Онтология стиля исследования

Стиль исследовательской работы подразумевает важную семантику. Он может отражать разнообразие исследований с точки зрения трудности выполнения, способов и типов. Например, E1 и E2 – это две исследовательские работы, определенные следующим образом:

E1: Авторы изучают методы, касающиеся определенной проблемы, и сжато излагают их в научной статье.

E2: Авторы фокусируются на определенной проблеме и предлагают решение, процесс и результат решения излагаются в научной статье.

Есть существенные различия между E1 и E2. E1 – это статья обзорного типа, тогда как E2 – тип статьи, касающийся проблемы решения. В целом у E2 больше мотивации и трудностей, чем у E1, и у них разные значения. E1 подходит начинающим ученым в целях получения базового знания, а E2 приносит больше пользы для воодушевления опытных людей. Поэтому тип научной статьи является важным фактором представления ее семантики. Для выражения знания, содержащегося в типах научного стиля, сначала создаем категории стиля тематического события, сформированного на рис. 3, с использованием проекта protégé (свободный, открытый редактор онтологии, а также система построения баз знаний) [37]. Каждый отдельный стиль каждой научно-исследовательской работы отражается и объясняется в табл.1. Онтология стиля исследования предполагает взаимосвязи между разными стилями исследовательской работы, которые можно использовать для измерения семантики между различными научными исследованиями.

Таблица 1

Подробности онтологии стиля исследования

Тип	Пояснение	Пример
Теоретическое происхождение (этимология)	Предложение оригинальных подходов	Латентное распределение Дирихле (Latent Dirichlet Allocation – LDA)
Улучшение методологии	Улучшение некоторых методологий или теорий	Улучшение тематических моделей LDA для микроблогов через массив Твиттера и автоматическое наименование
Сфера применения	Применение некоторых систем или методов	TEXTRUNNER: Открытое извлечение информации по сети
Проблема решения	Решение некоторых проблем существующими методами	Извлечение биологического события при помощи соответствия подграфа
Обзор	Обзор некоторых исследовательских вопросов	Обзор извлечения события из текста
Анализ	Анализ некоторых вопросов	Сравнение подходов относительно широкомасштабного анализа данных
Открытие феномена	Представление некоторых выводов	Роль ученых – лидеров в эволюции научных сообществ

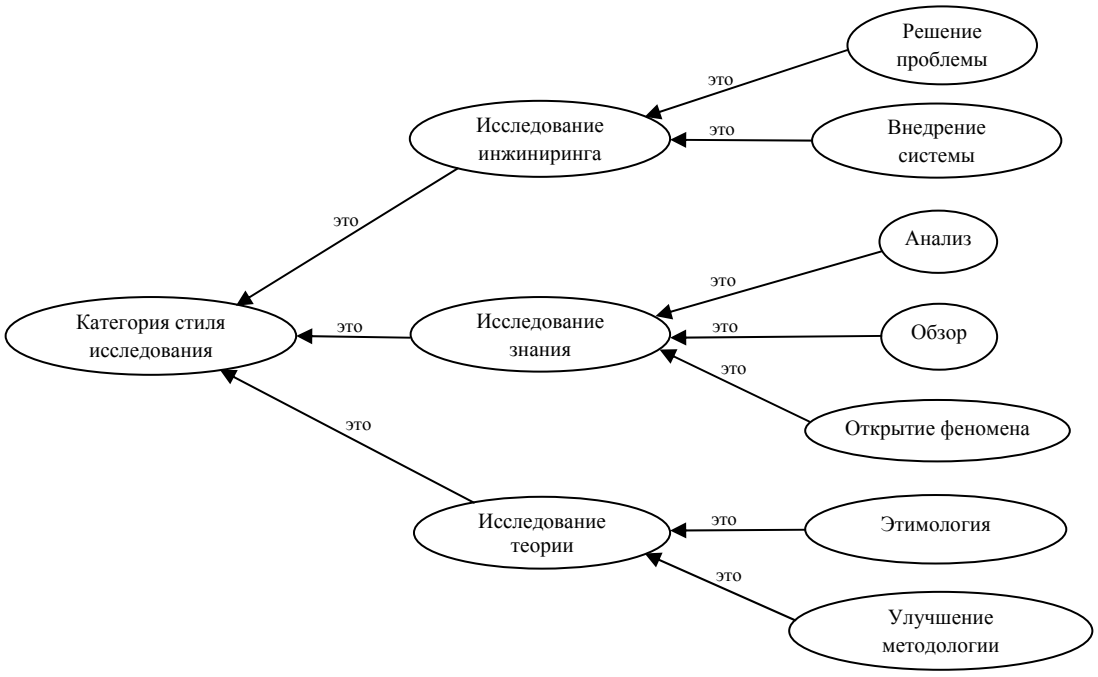

Рис. 3. Иерархия онтологии стиля исследования

АВТОМАТИЗИРОВАННОЕ ПОСТРОЕНИЕ ТЕМАТИЧЕСКОГО СОБЫТИЯ

Обзор

В соответствии с рис. 2 из документа нам необходимо извлечь *Target, Methodology, Domain, Style, keywords* и *Date* для формирования тематического события. Настоящая работа по извлечению структурированных представлений из событий в основном делает акцент на текстах новостных сообщений, использующих объекты наименования, временные выражения и значения, встречающиеся в предложениях цели, как на кандидатах для объектов события. Распространенная проблема извлечения события считается проблемой классификации относительно данных наименования. Однако данные наименования редко встречаются в целях модели извлечения, и почти никакой объект наименования не может служить в качестве кандидата для терминологии отдельной области.

Научная литература имеет ряд особенностей, которые представляют уникальные вызовы и уникальные возможности для признания элемента события. В научных статьях содержится множество структурированных аннотаций, таких как цитирования, авторы, дата публикации, ключевые слова и журналы, которые явно могут быть извлечены и использованы для обогащения тематического события [38]. Стоит отметить, что многие, вовлеченные в тематическое событие, единицы скрыты в неструктурированном контенте научных статей. Основная работа извлечения заключается в определении терминологий, таких как *target, methodology, domain* и *style* в содержании статьи. Для решения проблемы отсутствия данных относительно наименования события предлагается онтология и метод извлечения на основе модели.

Научные статьи, как правило, имеют понятные темы и цели, внутри имеется много повторяющихся синтаксических структур, содержащих ключи для извлечения события. Процесс формирования тематического события отображен на рис. 4. Сначала научные статьи делятся на разделы, и затем отбираются разделы наиболее важные для извлечения тематического события. Далее проводится основная обработка текста на естественном языке, такая как расчленение предложения и тегирование частей речи каждого включенного предложения в отобранных разделах. Потом отбираются все именные конструкции каждого предложения в качестве вероятных единиц тематического события, и составляется список терминологий по онтологии области. После этого из нескольких элементов, соответствующих модели, выбираются самые лучшие темы события. В конце извлеченные элементы события вносятся в онтологию области для расширения события родственными семантическими элементами. Обычно онтология области устанавливает семантическое сходство терминологий области, что помогает поддерживать терминологии, а также их взаимосвязи в определенной области. В данной статье онтология области может обеспечивать внешнее знание для семантического понимания документов и поддерживать процедуру по созданию тематического события.

Признание элементов события

Обратимся к подробностям – каждая научная статья делится на несколько частей в соответствии с ее планом, а разделы *Название, Реферат, Введение* и *Заключение*, предположительно, содержат глобальное описание исследовательской работы целиком без большого числа необязательных пояснений. Затем определяются важные предложения этих разделов путем триггеринга слов. Важные предложения могут содержать элементы события, всего определено 95 таких слов для извлечения элементов.

Для охвата претендентов на роль элементов события в важных предложениях усиливаем онтологию области и некоторые процессы обработки текста на естественном языке. Терминологический список, полученный из области онтологии, используется при поиске вероятных элементов события. Однако другая главная проблема заключается в том, как искать многие незнакомые фразы в сравнительно новых научных статьях. Для решения этой задачи проводим тегирование частей речи каждого предложения, а затем используем все именные конструкции в качестве претендентов на роль элементов события, чтобы охватить новые неизвестные фразы.

Признание цели и методологии. После получения претендента терминологии следующим шагом является подтверждение того, какой претендент считается самым лучшим элементом события в каждом предложении. Создаем модели извлечения *Target* и *Methodology*. Они состоят из набросков моделей (премоделей) и постмоделей, являющихся моделями, часто встречающимися до и после элементов события. Некоторые используемые модели извлечения *Target и Methodology* представлены в табл. 2. Всего насчитывается свыше 550 моделей извлечения *Target и Methodology*.

Например, вовлеченное предложение *«В статье мы предлагаем расширенный подход машинного обучения по извлечению связей»* определяется первоначальными словами «предлагаем» из раздела Введение. Оно соответствует предыдущей модели target, т.е. «подход к». Таким образом, терминология *«извлечение связей»* выбирается в качестве цели этой статьи.

Признание стиля исследования. Из табл. 1 видно, что статьи различных научных стилей обладают отличающимися наименованиями характеристик. Многие виды научных статей в области вычислительной лингвистики могут быть распознаны по названию статьи. Например, каждое название типа Проблема решения является строгим, оно начинается с аббревиатуры наименования программного обеспечения и связано с последующим названием через пунктуацию «:» или «-», для наглядности – «TEXTRUNNER: Открытое извлечение информации по сети», «USER: Сжатая сетевая система извлечения связей». Большая часть названий имеет характерные слова для разграничения своих научных стилей, создаются модели определения научного стиля тем в табл. 3.

Расширение семантики на основе онтологии

Многие элементы тематического события, такие как цель исследования и область, адаптированная методология и способы, исследовательский объект и массив данных, тесно соотносятся. Вообще, исследовательские цели составляют центральные вопросы научных статей и соответствующие тематические события, а область, к которой принадлежит научная статья, определяется ее исследовательской целью. Когда устанавливается цель научной статьи, применяется онтология области с целью выявить область, к которой принадлежит научная статья.

После того, как установлена цель научной статьи, семантические сходства между целью и каждым понятием конкретной области подсчитываются на основе онтологии области (рис. 5 отражает область вычислительной лингвистики как пример онтологии области). Понятие области, имеющее максимальное семантическое сходство с понятием цели, выбирается в качестве области соответствующей научной статьи.

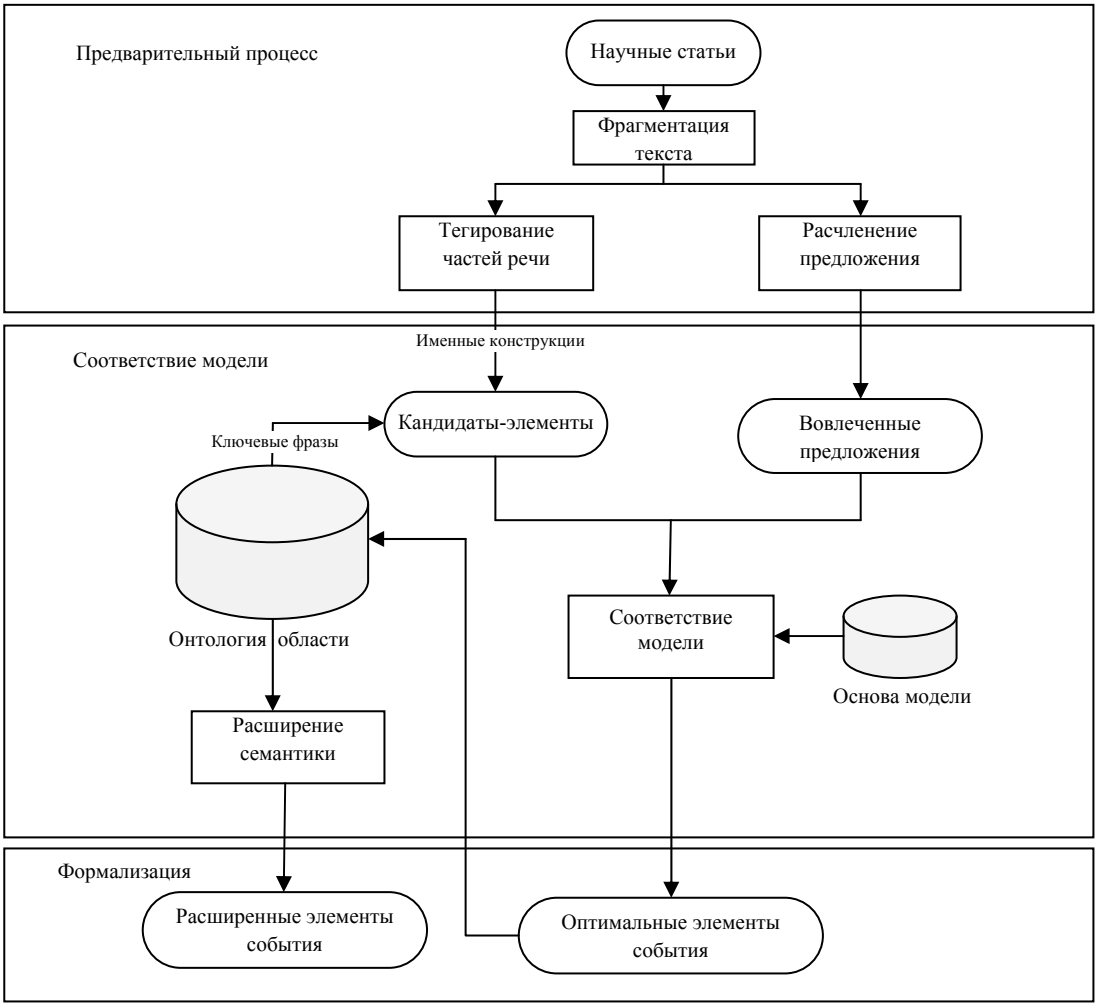

Рис. 4. Процесс извлечения тематического события

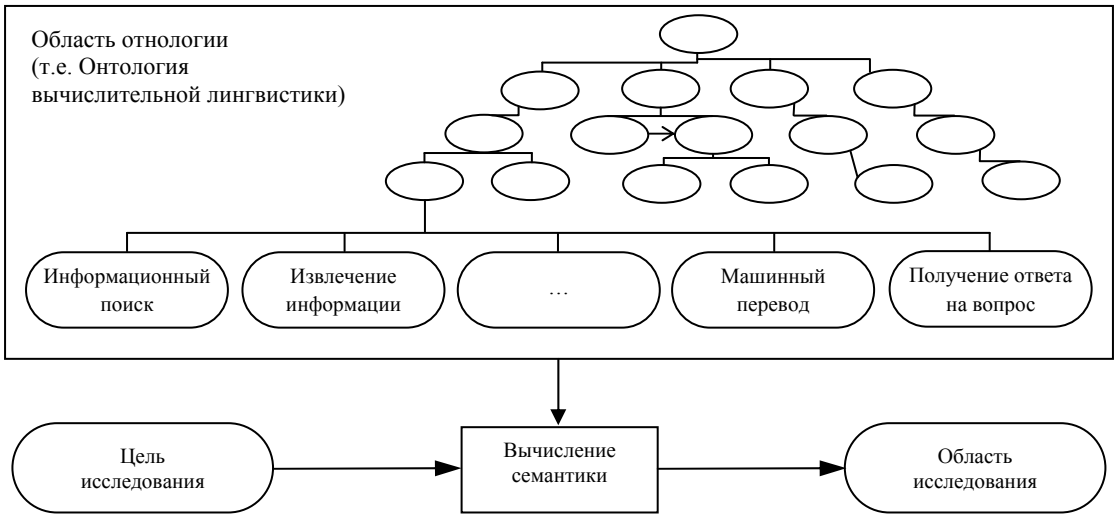

Рис. 5. Архитектура расширения семантики на основе онтологии

Таблица 2

Типичные модели извлечения для элементов тематического события (ТЕ)

Элемент ТЕ	Модели, предшествующие элементу ТЕ	Модели, следующие за элементом ТЕ
Target	Проблема	Обзор
	Задача	Оценка
	Система для	Отслеживание
	Обзор по	Система
	Способ для	Именной
	Подход к	Процесс
Methodology	Путем использования	Основной подход
	Тот, кто использует	Метод для
	Тот, кто нанимает Имеет преимущество	Алгоритм по Способы для
	Методы для	Применяется к
	Посредством	Выполняет гораздо лучше

Таблица 3

Типичные модели стилей исследования

Категория исследования	Модель
Этимология	Модель
	Модель для
Улучшение методологии	Улучшающая
	Улучшенная
	Улучшение
Сфера применения	-,:,корпус
Проблема решения	Использующая, основанная, сфокусированная
	Подход, распространяющая, гибридная
	Реализуемая путем, изучающая, внедряемая
	Использующая, методы, извлекающая
	С помощью, использование, через, способ, измерения
Обзор	Обзор
	Обзор
	Обзор
	Введение
Анализ	Сравнение
	Оценивающая
	Оценка
	Оценивающая
	Анализ
	Проблемы
	Будущее для
Открытие феномена	Изучающая
	Изучение для

ВЫЧИСЛЕНИЕ ПОДОБИЯ ТЕМАТИЧЕСКОГО СОБЫТИЯ

Общий подход

Поскольку тематическое событие может представлять семантику документа, то семантическое сходство документов может достигаться через сходство тематических событий. В данной статье скрытая внутри каждого документа информация извлекается из тематического события и устанавливается внутренняя релевантность понятий из онтологии области.

Наш метод для краткого вычисления ядра семантики статей использует шесть первых элементов в тематическом событии в соответствии с их характеристиками, таких как *Target, Domain, Style, Methodology, Keywords* и *Date*. Используется онтология тематического события для измерения сходства внутреннего события в различных типах, а онтология области – для семантического сходства внутри терминологий. Подобие тематического события подсчитывается взвешенной суммой сходств элементов в структуре соответствующего события, а также может дополняться метаданными и другими элементами для получения более детального сходства тематического события. Сходство тематических событий E1 и E2 определяется следующим равенством (1):

$$SimTEs(E_1, E_2) = \sum_{i=1}^{6} W_i \times S_i(L_1, L_2) \ , \qquad (1)$$

где w_i – вес элемента i в тематическом событии, S_i – функция сходства элементов i для L_1 и L_2. L_1 и L_2 представляют два тематических события, элементы которых определяются как L = {Target, Domain, Style, Methodology, Keywords, Date}.

Тематические события извлекают ядро семантики каждой научной статьи. Однако значение элементов события не может измеряться их буквальным появлением. Фундаментальное знание, такое как лексическое значение, обязательно для понимания семантики статей. Для получения внутренней семантической связи в терминологиях терминологии в тематических событиях следует связать с узлами понятий в базе знаний, чтобы установить их семантическое сходство. В следующем подразделе вводится метод связывания понятий и мера семантического сходства элементов.

Связывание понятий в онтологии области

Для вычисления сходства внутри терминологий важным вопросом считается связь извлеченных терминологий с их собственными позициями в базе знаний. В научных статьях имеется множество синонимов, многие понятия могут описываться различными терминологиями в разных статьях, такими как *«перекрестный лингвистический поиск»* и *«многоязыковой информационный поиск»*, *«понимание текста»* и *«понимание сообщения»*, *«признание именного объекта»* и *«тегирование именного объекта»*. При формировании онтологии области присваивалось наименование всем известным синонимам каждого понятия в узле, чтобы облегчить связывание объектов.

Многие автоматически извлеченные терминологии несут тривиальные суффиксы и префиксы, что может затруднять связывание объектов. Для решения этой проблемы используется редакционное расстояние, чтобы измерять сходство строк и признавать вариации одного и того же понятия. Поскольку наша онтология формируется для области вычислительной лингвистики, и все узлы понятий извлекаются из массива области, то большинство терминологий будет находить позиции извлеченного понятия в онтологии области. Сначала составляем список терминологий, чтобы формировать онтологию области. Когда появляется терминология, извлекаемая из научных статей, подсчитывается ее редакционное расстояние с каждой терминологией. Узел понятия с минимальным редакционным расстоянием будет считаться узлом терминологии.

Сходство элементов тематического события

Элементы, такие как *Target, Methodology* и *Domain,* являются терминологиями, их семантические сходства можно измерить онтологией области. Семантическое сходство Style может измеряться через онтологию стиля исследования. Сходство Date – их интервалом. Короче говоря, сходство этих элементов в тематических событиях может измеряться следующими методами.

Сходство стиля исследования. Для установления различия в отличающихся типах научной работы, может использоваться онтология стиля исследования, представленная в табл. 1. Сходство *Style* различных типов тематических событий измеряется методом, подобным таковому By и Палмера [22], на основе онтологии стиля исследования, отраженной на рис. 3, и формулой равенства (2):

$$Sim_{Etype} = \frac{2 \times depth(LCS)}{depth(Style_1) + depth(Style_2)}, \quad (2)$$

где $Style_1$ и $Style_2$ обозначают типы двух тематических событий. LCS является наименьшей общей подсуммой узлов двух стилей.

Сходство терминологии. Понятие семантического сходства можно измерять базой знания. Оценивалось несколько методов на основе знаний и был сделан вывод, что метод By и Палмера больше подходит для сходства понятий в этой области. Понятия *Target, Domain, Methodology* и *Keywords* представляют наборы терминологий, вычисляемых методом By и Палмера на основе онтологии области или методов на основе вовлечения слов. Семантическое сходство понятия на основе онтологии подсчитывается равенством (3):

$$Sim_{ec} = \frac{2 \times depth(LCS)}{depth(ec_1) + depth(ec_2)}, \quad (3)$$

где ec_1 и ec_2 представляют терминологии тематического события.

При использовании метода сходства понятия на основе массива для семантического сходства может использоваться сходство векторов терминологии по косинусу. Данное сходство по косинусу определяется равенством (4):

$$Sim_{ec} = \frac{TermVec_1 \bullet TermVec_2}{|TermVec_1||TermVec_2|}. \quad (4)$$

Сходство даты. Вопросы исследования продолжают появляться по мере развития науки, а ученые фокусируются на отдельных научных проблемах на каждом этапе. Предполагается, что научные статьи, имеющие близкие даты публикации, будут становиться схожими, тогда как статьи, опубликованные по датам удаленно друг от друга, будут иметь все меньше общих черт. Таким образом, сходство даты может устанавливаться через временной интервал. Года (*years*) и месяцы (*months*) используются в вычислениях сходства между двумя датами. Сходства даты определяется формулой равенства (5):

$$Sim_{Date} = \frac{1}{1 + \left| (year_1 + \frac{month_1}{12}) - (year_2 + \frac{month_2}{12}) \right|}. \quad (5)$$

ОЦЕНКА ЭКСПЕРИМЕНТА

Формирование массива

Имеется несколько общедоступных массивов данных, применяемых к оценке семантического сходства коротких текстов и предложений, таких как MSPR [3], Michael D.LEE 50 corpus [7] и SEMILAR corpus [5]. Ни один текст из этих массивов не содержит больше 200 слов, что не обеспечивает надежность семантического сходства на уровне документов. Поэтому сформируем массив данных для семантического сходства между документами, используя научные статьи в области вычислительной лингвистики. Набор пар статей создается из системы AAN [20]. Пары статей аннотируются с помощью двухуровневого и пятиуровневого аннотирования в качестве основополагающей истины. Каждая пара статей обозначается 1, если она подобна с точки зрения семантики, или ей присваивается 0, если она семантически разнородна на втором уровне аннотирования. На пятом уровне аннотирования пара статей маркируется интервалами от 1 до 5 в соответствии со степенью ее семантического сходства. Если они (статьи) равноценны с точки зрения семантики, сходство статей будет аннотировано 5, если не имеют никакого семантического сходства друг с другом, то в таком случае сходство анно-

тируется 1. Двенадцать экспертов из нашей лаборатории составляли аннотации и одновременно подтверждали связь 1021 пары документов. Каждая пара статей повторно аннотируется другим лицом после первого аннотирования. Если повторное аннотирование соответствует первому, данное аннотирование признается основополагающей истиной, иначе – третье лицо аннотирует данную пару статей для получения основополагающей истины. В итоге получается аннотированный массив из 1021 пары статей. Сегодня данный массив общедоступен, а его унифицированный указатель ресурса следующий — https://github.com/buaaliuming/DSAP-document-semantics-foracademic-papers/tree/buaaliuming-annotation.

Онтология вычислительной лингвистики

Ресурсы базовых знаний, такие как WordNet, не могут охватить терминологии области. Для вычисления семантических сходств терминологий вручную формируется онтология области, чтобы передать семантику в различных терминологиях.

Понятия, извлеченные из системы AAN [20], используются для формирования вручную онтологии вычислительной лингвистики. На данный момент наша онтология содержит 1195 узлов понятий с иерархией из 9 уровней, которая может постоянно расширяться в будущем. Архитектура онтологии вычислительной лингвистики представлена на рис. 6. Основные взаимосвязи между понятиями в онтологии гипонимичны (противопоставлены). Синонимы рассматриваются и аннотируются в онтологические узлы понятий по ходу формирования онтологии.

Онтология вычислительной лингвистики применяется для измерения сходства между понятиями области вычислительной лингвистики. Учитывая характер области вычислительной лингвистики планируем создать онтологию области в трех частях, а именно: *Тема исследования, Инфраструктура* и *Общие подходы*, каждая из частей подкрепляется нисходящими узлами дополнительной информации. Узел *Общий подход* содержит методологии общего характера, действующие в области вычислительной лингвистики, такие как *машинное обучение, соответствие модели* и *проектирование знания* и т.д. Узел *Тема исследования* включает *фундаментальные языковые процессы*, *исследовательские проблемы* и *объекты исследований. Фундаментальные языковые процессы* предполагают обработку текста на естественном языке, такую как *сегментация слов*, *синтаксический анализ*, *тегирование по частям речи, лемматизация* и т.д. *Исследовательские проблемы* затрагивают насущные исследовательские вопросы, такие как *машинный перевод*, *категория текста*, *извлечение информации*, *информационный поиск* и *распознавание речи* и т.д. Узел *Инфраструктуры* содержит общие инструменты, базы знаний, собрания и организации в области вычислительной лингвистики.

Экспериментальная установка

Эксперименты проводятся на машине DELL OptiPlex390, имеющей размер памяти 8G и центральный процессор I5-2400. Помимо автоматического формирования тематических событий, вручную аннотируются и соответствующие тематические события из научных статей для выявления противоречия эксперимента. Метод на основе LDA выбран в качестве приоритетного. При вычислении семантического сходства терминологий в тематических событиях применяется метод LSA наравне с методом на основе онтологии. Короче говоря, осуществляются следующие методы.

LDA_2013. Метод на основе LDA (2013 г.) [36] является самым тесно связанным исследованием и выбран в качестве приоритетного по сравнению с другими. При выполнении метода на основе LDA некоторые модели LDA с заданными разными параметрами основаны на системе AAN [20]. В представленных ниже результатах по контрасту выбирается модель LDA с 200 темами, что приводит к наилучшему выполнению работы в разных моделях LDA при одних и тех же условиях.

TE_Onto. Метод семантического сходства тематического события работает на точных аннотациях тематического события, а семантические сходства понятия вычисляются по нашей онтологии вычислительной лингвистики.

Auto TE_Onto. Метод семантического сходства тематического события работает на автоматически извлеченных тематических событиях, а семантические сходства понятия также вычисляются по нашей онтологии вычислительной лингвистики. Метод **Auto TE_Onto** используется в сравнении с методом **TE_Onto** в целях измерения влияния автоматического извлечения тематического события.

TE_LSA. Метод семантического сходства тематического события работает на точных аннотациях тематического события, а семантические сходства понятия вычисляются по векторам слов, созданным LSA. Метод LSA извлекает представление каждого слова с помощью операции SVD (singular-value decomposition – сингулярное разложение), а семантическое сходство между терминологиями охватывается сходством общих тем. Когда используется метод LSA при подсчете единицы тематического события, формируется матрица термин-подокументу по всей совокупности аннотаций. Метод **TE_LSA** используется в сравнении с методом TE_Onto в целях измерения работы онтологии вычислительной лингвистики.

Auto TE_LSA. Метод семантического сходства тематического события работает на автоматически извлеченных тематических событиях, а семантические сходства понятия вычисляются по векторам слов, созданным LSA.

TE item weights. Первоначальными элементами тематического события являются *Target*, *Domain*, *Methodology*, *Style, Keywords* и *Date* – важные элементы рис. 2. Исходя из нашего опыта, цель исследования представляет собой существенный вопрос каждой научной статьи. Области исследований, типы исследовательских работ и адаптированные методы научных статей являются важными аспектами, которые придают различный характер каждой исследовательской работе. Определяем вес упомянутых элементов в соответствии с их важностью и, руководствуясь нашими экспериментами, устанавливаем вес для этих элементов, т. е. *Target*, *Domain*, *Methodology*, *Style, Keywords* и *Date*, в размере 0,3, 0,25,0.25, 0,1, 0,05 и 0,05 соответственно.

Evaluation metrics. Выбираем коэффициент корреляции Пирсона для измерения качества оценок семантического сходства. Чем больше коэффициент корреляции Пирсона, тем более коррелируемыми являются прогнозируемые оценки и основополагающая истина. Коэффициент корреляции Пирсона отражен в равенстве (6).

$$Px,y = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left((X-\mu_X)(Y-\mu_Y)\right)}{\sigma_X \sigma_Y} \quad . \qquad (6)$$

Здесь X - прогнозируемая оценка семантического сходства, а Y указывает на значение аннотированного семантического сходства. *cov(X,Y)* представляет ковариацию X и Y. μ_X и μ_Y обозначают средние значения переменных X и Y; σ_X и σ_Y – стандартные отклонения X и Y соответственно.

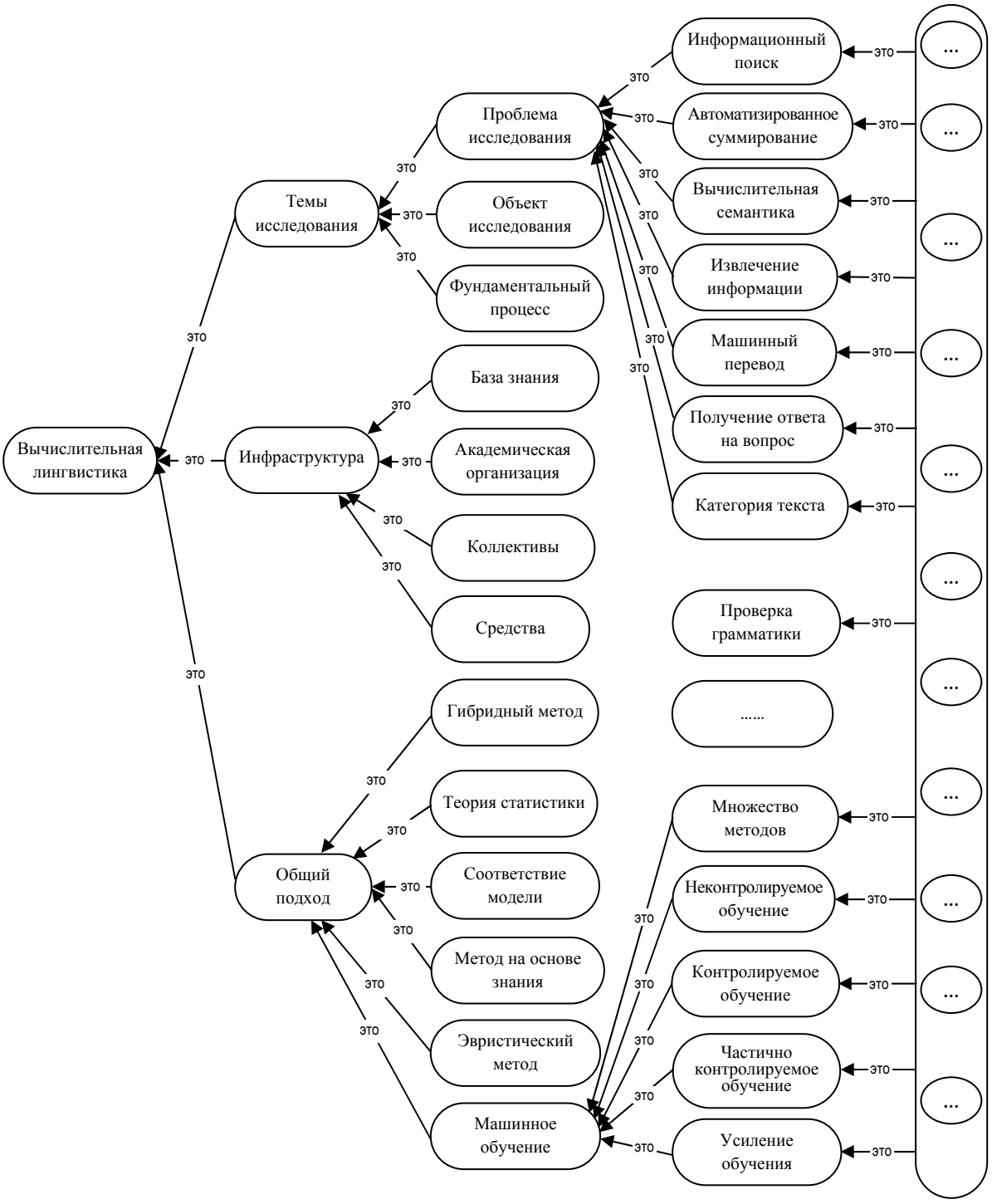

Рис. 6. Базовая структура онтологии вычислительной лингвистики

Поскольку наш массив имеет бинарные аннотации и аннотации пятого уровня, далее устанавливаем разные пороговые величины для прогнозирования того, окажутся ли семантически сходными два документа. Обе оценки Accuracy и F1 могут представлять метрику всей оценки помимо корреляции. Accuracy отражена в равенстве (7).

$$Accuracy = \frac{(TP - TN)}{(TP + TN + FP + FN)} \qquad (7)$$

В равенстве (7) *TP* обозначает число пар документов, предположительно схожих, которые в действительности представляют пары подобных документов. *TN* – число пар документов, различных по предположению, которые в действительности представляют пары различных документов. *FP* обозначает число пар документов, как предполагается, схожих, которые в действительности представляют пары различных документов. *FN* – число пар документов, предполагаемо различных, которые в действительности представляют пары подобных документов.

Accuracy – означает общую способность предсказывать метод, а оценка F1 подразумевает полную эффективность точности и полноты. Оценка F1 представлена в равенстве (8).

$$F\ score = \frac{(1+\beta)P}{(\beta P + R)} \ . \qquad (8)$$

P означает точность, а R - полноту. Вообще, значением β является 1, а оценка $F\,score$ аннотируется как оценка F1 в следующем разделе.

Результаты и обсуждения

Коэффициент корреляции Пирсона

Для проверки качества различных методов подсчитаем коэффициент корреляции Пирсона для 1021 оценки сходства при помощи сопоставления с аннотированной людьми основополагающей истиной.

Сравнение с приоритетным методом. Результаты, приведенные в табл. 4, показывают, что наш метод тематического события на основе онтологии имеет явное преимущество перед приоритетным методом, т.е. методом LDA_2013. Аннотирование на пяти уровнях обладает более детализированными уровнями сходства; а оценки в корреляциях с ним более убедительны, чем оценки в корреляциях при аннотировании на двух уровнях. Наш метод TE_Onto достигает 4,1% (относительного) улучшения относительно приоритетного метода; когда тематические события извлекаются автоматически, наш метод Auto TE_Onto может добиться 5,8% (относительного) улучшения по сравнению с приоритетным методом.

Влияние онтологии. Наш метод TE_Onto гораздо эффективнее, чем метод тематического события на основе LSA, а Auto TE_Onto демонстрирует даже 22,7% преимущества относительно метода Auto TE_LSA, при котором тематические события извлекаются автоматически.

Метод на основе LSA измеряет семантическое сходство понятий с помощью общих слов, характерных для тем документа, или совместной встречаемости слов, тогда как методы на основе онтологии измеряют семантическое сходство понятий через точное знание в онтологии. Представленные выше результаты подтверждают, что ресурсы знаний, такие как онтология области, крайне важны для измерения семантики документа.

Влияние извлечения тематического события. Оценки в корреляциях наших методов по автоматическому формированию тематических событий и аннотированным вручную тематическим событиям являются близкими. Эффективность наших методов автоматического извлечения тематических событий сопоставима с точно аннотированными экспертами тематическими событиями. Метод TE_LSA показывает очень маленькое преимущество перед методом Auto TE_LSA, а метод Auto TE_Onto даже немного лучше, чем метод TE_Onto в корреляции аннотации на пяти уровнях. Это подтверждает, что метод извлечения на основе модели может извлекать необходимую информацию с должной точностью в отдельной области, а процесс автоматического извлечения сопоставим с ручным аннотированием.

Accuracy и F1-score

Accuracy *и F1-score* могут быть метриками общей оценки помимо Корреляции. Accuracy отражает общую способность метода прогнозировать правильный результат; *F1-score* является балансом между точностью и полнотой, что демонстрирует полную эффективность данного метода. В целях применения на практике могут быть установлены различные пороговые величины, чтобы предсказывать, обладают ли два документа семантическим соответствием. Вообще, наилучшая эффективность среди различных порогов считается важным фактором в оценке. В следующих экспериментах заявлены различные пороговые величины. Пары статей предполагаются семантически схожими, если их оценка сходства гораздо выше заявленных порогов.

Сравнение с приоритетным методом. Как показывают рис. 7 и 8, оценка Accuracy в наших методах TE всегда имеет значительное преимущество над методом на основе LDA при разных пороговых величинах.

Наша лучшая оценка *F1-score* – 0,639, тогда как лучшая оценка *F1-score* приоритетного метода – 0,536. Оценка *F1-score* в нашем методе Auto TE_Onto превосходит LDA_2013 в большинстве порогов, а оценки *F1-score* в наших методах TE_Onto с точными тематическими событиями демонстрирует преимущество над приоритетным методом, когда порог составляет менее 0,75.

Влияние онтологии. Сравнивается эффективность метода тематического события на основе LSA с методом тематического события на основе онтологии. Как показывают рис. 9 и10, лучшая оценка Accuracy в методе тематического события на основе LSA равняется 0,712, тогда как лучшая оценка Accuracy в методе тематического события на основе онтологии – 0,768. Методы тематического события на основе онтологии гораздо эффективнее, чем методы тематического события на основе LSA, когда пороговые величины составляют менее 0,250. Подводя итоги, методы тематического события на основе онтологии более выгодны, чем методы тематического события на основе LSA, что соотносится с результатом корреляции Пирсона из предыдущего подраздела. Эти результаты показывают, что онтология играет важную роль в распознавании семантического сходства документов.

Таблица 4

Сравнение корреляций Пирсона с основополагающей истиной

Корреляции	Корреляция при аннотировании на пяти уровнях	Корреляция при аннотировании на двух уровнях
TE_Onto	0,559	0,461
Auto TE_Onto	0,568	0,456
TE_LSA	0,480	0,346
Auto TE_LSA	0,463	0,327
LDA_2013	0,537	0,250

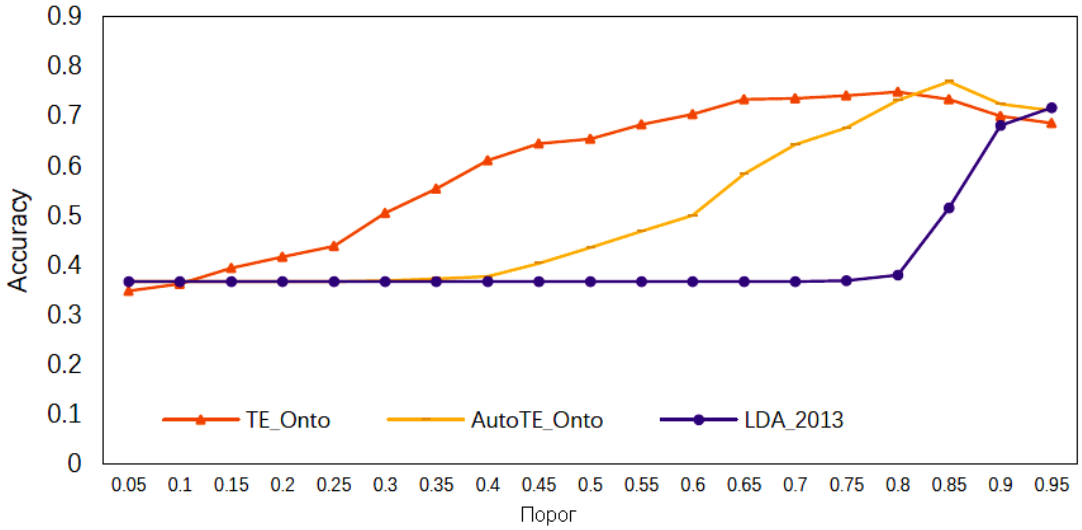

Рис. 7. Сравнение оценки Accuracy в наших методах и приоритетном методе

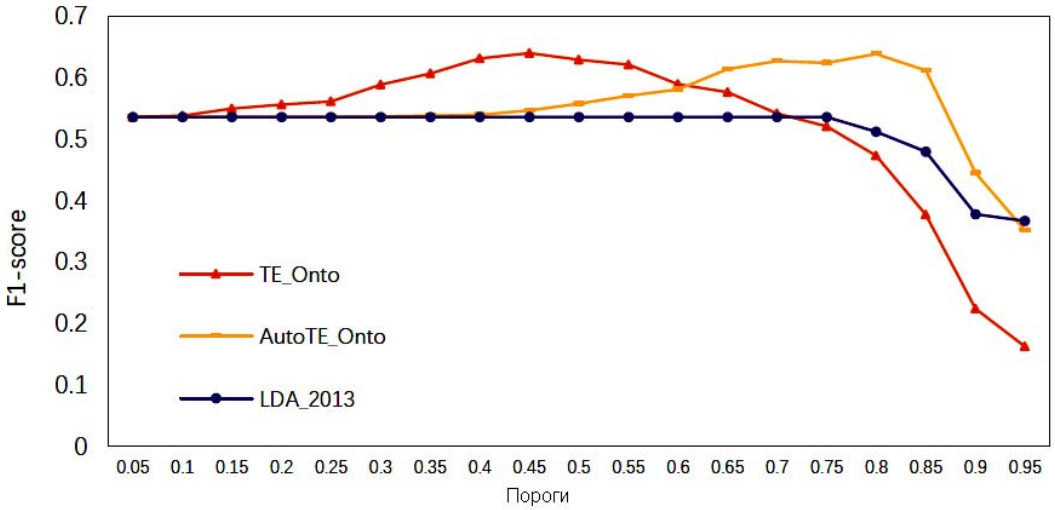

Рис. 8. Сравнение оценки F1 в наших методах и приоритетном методе

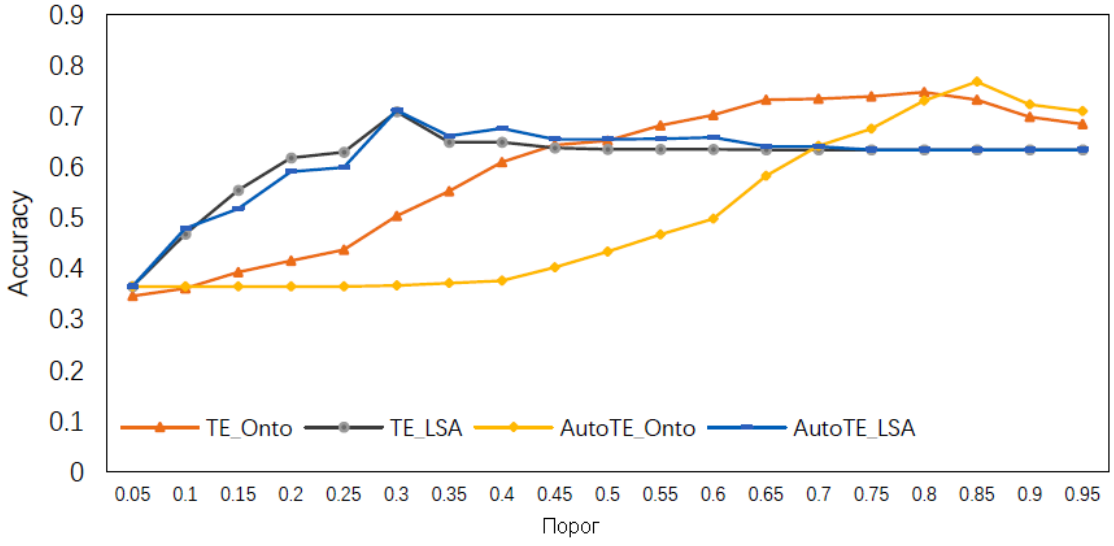

Рис. 9. Сравнение оценки Accuracy в методах тематического события на основе онтологии с методами тематического события на основе LSA

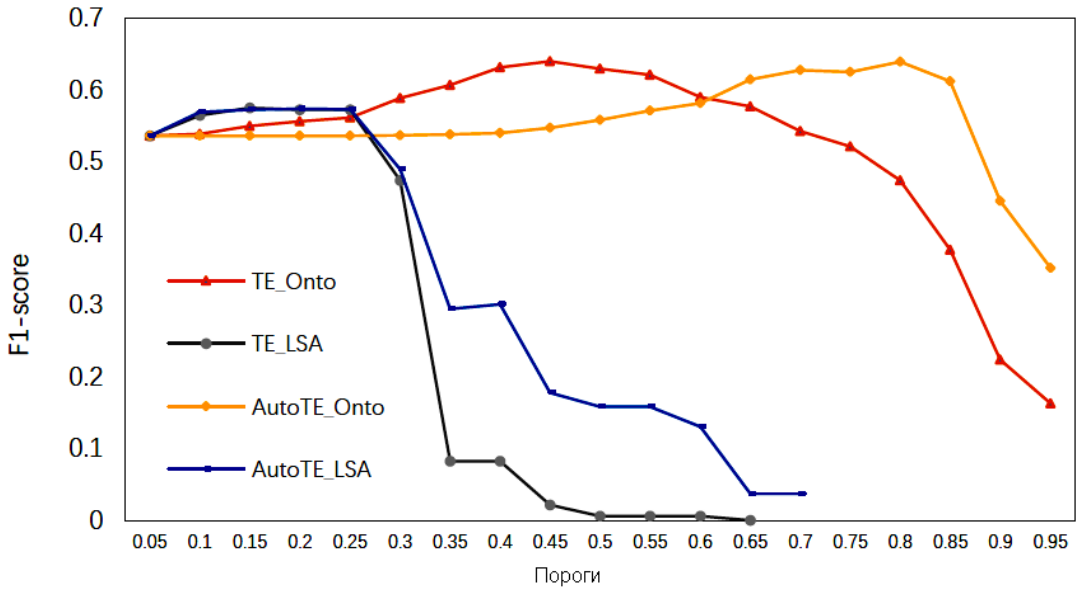

Рис. 10. Сравнение оценки F1 в методах тематического события на основе онтологии с методами тематического события на основе LSA

Обсуждение. Описанные эксперименты демонстрируют, что в целом методы тематического события могут отражать гораздо большую эффективность. Оценки сходства в методах TE_LSA и Auto TE_LSA варьируются от 0,0 до 0,70, и здесь нет никакого реального положительного результата, когда пороги превышают 0,70. Эти низкоуровневые оценки сходства сводят на нет подсчет оценок F1-scores в методе тематического события на основе LSA. Научные статьи, как правило, имеют богатый контекст, содержащий множество дублирующих терминологий, а часто используемые терминологии в узко предметной области имеют тенденцию появляться в нескольких статьях данной области. Более того, научные статьи склонны рассматривать связанные с исследуемой проблемой работы. Метод на основе LDA определяет тематическое сходство с помощью пересечения слов. Следовательно, оценки сходства в LDA_2013 ранжируются относительно высоко – от 0,7 до 1,0, что свидетельствует о его хорошей работе с точки зрения полноты, но слабой работе с точки зрения точности, а также преуменьшает его общую производительность. Наши методы обладают основной семантикой документа через структурированные тематические события, оценки сходства в методах тематического события колеблются от 0,0 до 1,0 относительно сходства пар документов, они больше различаются и в то же время имеют более высокую общую эффективность.

Затраты времени и памяти

Каждый метод нашего эксперимента нуждается в процессе, не требующем работы в интерактивном компьютерном режиме. Трудно измерить и сопоставить затраты этого не интерактивного процесса в различных условиях и процедурах. В этом разделе измеряется рабочее время, затраченное на вычисления семантического сходства на уровне документов в одних и тех же условиях. Среднее время, потраченное на методы тематического события, составляет 0,02 сек, а требуемый объем памяти – около 100 М, тогда как LDA_2013 тратит 4,83 сек и занимает более 8Г объема памяти. Очевидно, что наши методы тематического события более точны, чем традиционное измерение на основе LDA с точки зрения затрат по времени и памяти. Причиной этого служит то, что наши методы тематического события применяют онтологию области для подсчета семантического сходства в отличие от развернутой модели LDA. В целом, наши методы семантического сходства, получающие семантические сходства на основе извлечения и модели, достигают желаемой общей эффективности.

ЗАКЛЮЧЕНИЕ

В статье впервые предлагается модель формирования тематического события для представления семантики документа и измеряется семантическое сходство научных статей при помощи вычисления сходства их соответствующих тематических событий. Очерчивается общая архитектура тематического события и описывается модель ее формирования, а также способы вычисления сходства тематических событий. Формируются онтология стиля исследования и система оценок. Чтобы измерить семантическое сходство понятий, разрабатывается онтология для вычислительной лингвистики. Оценки эксперимента показывают, что наш метод тематического события получает значительное улучшение относительно действующих измерений семантического сходства, а методы тематического события на основе онтологии демонстрируют все преимущества в Корреляции, Accuracy и *F1-score*. Ресурсы знания, такие как онтология области, играют важную роль в семантическом сходстве документов.

Более того, наш метод можно использовать для моделирования семантики в различных стилях научной работы, а автоматическое формирование на основе онтологии и подсчет сходства тематического события согласуются с различными областями, что подразумевает, что наш метод может легко применяться для разных областей науки.

Благодарность. Данное исследование выполнено при поддержке фонда Государственной профильной лаборатории по разработке сред программирования (грант No. SKLSDE-2015ZX-04). Авторы выражают признательность рецензентам за их ценное сотрудничество, которое послужило существенному улучшению авторского текста статьи.

ЛИТЕРАТУРА

1. *Corley C., Rada M.* Measuring the semantic similarity of texts//Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment. — 2005. —P. 13-18.

2. *Rus V., Lintean M., Graesser A. C., McNamara D.S.* Assessing student paraphrases using lexical semantics and word weighting//Proceedings of the 14th International Conference on Artificial Intelligence in Education, Brighton, UK. — 2009.

3. *Banea C, Choi Y, Deng L, et al.* CPN-CORE: A text semantic similarity system infused with opinion knowledge//Proceeding of Second Joint Conference on Lexical and Computational Semantics. Atlanta, Georgia, USA. — 2013. — P. 221.

4. *Dolan B., Quirk C., Brockett C.* Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources// Proceedings of ACL. — 2004. — P. 350.

5. *Agirre E, Banea C.* SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability//[C] SemEval 2015, June.

6. *Ramage D., Rafferty A. N., Manning C. D.* Random walks for text semantic similarity// Proceedings of the 2009 workshop on graph-based methods for natural language processing. Association for Computational Linguistics. — 2009. — P. 23-31.

7. *Rus V., Lintean M., Moldovan C., Baggett W., Niraula N., Morgan B.* The similar corpus: A resource to foster the qualitative understanding of semantic similarity of text// Proceedings of LREC. — 2012. — P. 23-25.

8. *Šarić F., Glavaš G., Karan M., Šnajder J., Bašić B. D.* Takelab: Systems for measuring semantic text similarity // Proceedings of the Sixth International Workshop on Semantic Evaluation. —Association for Computational Linguistics. — 2012. — P. 441-448.

9. *Amsler R.* Application of citation-based automatic classification. Technical report. — The University of Texas at Austin Linguistics Research Center, 1972.

10. *Kessler M.* Bibliographic coupling between scientific papers// Journal of the American Documentation. — 1963. — Vol. 14, No. 1. —P.10-25.

11. *Small H.* Co-citation in the scientific literature: A new measure of the relationship between two documents// Journal of the American Society for Information Science. — 1973. — Vol. 24, No.4. — P. 265-269.

12. *Salton G., Wong A., Yang C. S.* A vector space model for automatic indexing[J]// Communications of the ACM. — 1975.— Vol. 11. — P. 613-620.

13. *Blei D. M., Ng A. Y, Jordan M. I.* Latent Dirichlet allocation// J. of Mach Learn. Res. — 2013. — Vol. 3. — P. 993-1022.

14. *Madylova A., Öğüdücü Ş G.* A taxonomy based semantic similarity of documents using the cosine measure[C] // Computer and Information Sciences. 24th International Symposium on. IEEE. — 2009. — P. 129-134.

15. *Nagwani N. K., Verma S.* A frequent term and semantic similarity based single document text summarization algorithm [J]// International Journal of Computer Applications. — 2011. — Vol. 0975–8887. — P. 36-40.

16. *Schuhmacher M, Ponzetto S P.* Knowledge-based graph document modeling[C]//Proceedings of WSDM. — 2014. — P. 543-552.

17. *Rafi M, Shaikh M. S.* An improved semantic similarity measure for document clustering based on topic maps[J]. — arXiv preprint arXiv:1303.4087, (2013).

18. *Wang Y., Zhang R.-b., Lai J.-b.* Measuring concept similarity between fuzzy ontologies// Fuzzy Information and Engineering. — 2009. — Vol. 2. — P.163-171.

19. *Zhang M., Qin B., Zheng M., et al.* Encoding Distributional Semantics into Triple-Based Background Knowledge Ranking for Document Enrichment// Proceedings of ACL.

20. *Radev D. R., Muthukrishnan P., Qazvinian V.* The ACL anthology network corpus// Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries. Association for Computational Linguistics. — 2009. — P. 54-61.

21. *Leacock C., Chodorow M.* Combining local context and WordNet sense similarity for word sense identification//WordNet, An Electronic Lexical Database. — The MIT Press, 1998.

22. *Wu Z., Palme M.* Verb semantics and lexical selection//Proceedings of ACL — 1994. — P. 133-138.

23. *Lin D.* An information-theoretic definition of similarity// Proceedings of the International Conf. on Machine Learning. — 1998.

24. *Resnik P.* Using information content to evaluate semantic similarity// Proceedings of the 14th International Joint Conference on Artificial Intelligence. — 1995.

25. *Goikoetxea J., Soroa A., Agirre E., et al.* Random walks and neural network language models on knowledge bases[C]//Proceedings of NAACL-HLT. — 2015. — P. 1434-1439.

26. *Faruqui M., Dyer C.* Non-distributional word vector representations [J]. — arXiv preprint arXiv:1506.05230, (2015).

27. *Landauer T. K., Foltz P. W., Laham D.* An introduction to latent semantic analysis// Discourse Processes. — 1998. — Vol. 25, No. 2-3. — P. 259-284.

28. *Pennington J., Socher R., Manning C. D.* Glove: Global vectors for word representation// Proceedings of EMNLP. — 2014.

29. *Turney P.* Mining the web for synonyms: PMI-IR versus LSA on TOEFL// Proceedings of the Twelfth European Conference on Machine Learning. — 2001.

30. *Cilibrasi R.L., Vitanyi P. M. B.* The Google Similarity Distance// IEEE Trans. Knowledge and Data Engineering. — 2007. — Vol. 19, No. 3. — P. 370-383.

31. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient estimation of word representations in vector space. — arXiv preprint arXiv:1301.3781, (2013).

32. *Q. V. Le, Mikolov T.* Distributed representations of sentences and documents. — arXiv preprint arXiv:1405.4053, (2014).

33. *Martín G. H., Schockaert S., Cornelis C., et al.* Using semi-structured data for assessing research paper similarity [J] // Information Sciences. — 2013. — Vol. 221. — P. 245-261.

34. *Zhang M, Qin B., Liu T., et al.* Triple based background knowledge ranking for document enrichment// Proceedings of COLING. — 2014.

35. *Kim J. H., Kim D., Kim S., Oh A.* Modeling topic hierarchies with the recursive chinese restaurant process// Proceedings of the 21st CIKM. — 2012. — P. 783–792.

36. *Rus V., Lintean M., Banjade R., Niraula N., Stefanescu D.* SEMILAR: The Semantic Similarity Toolkit//Proceedings of ACL — 2013. — P. 163-168.

37. *Musen M.A.* The Protégé project: A look back and a look forward. AI Matters// Association of Computing Machinery Specific Interest Group in Artificial Intelligence. — 2015. —Vol. 1, No. 4, June.

38. *Tkaczyk D., Szostek P., Fedoryszak M., et al.* CERMINE: Automatic extraction of structured metadata from scientific literature//International Journal on Document Analysis and Recognition (IJDAR). — 2015. — P. 317–335.

Приглашаем российских и зарубежных авторов к сотрудничеству
в журнале «Международный форум по информации».
Оригинальные статьи и другие материалы (рецензии, письма)
можно присылать на русском или английском языке
по почтовому адресу, указанному в «Памятке для авторов»
или по электронной почте: mfi@viniti.ru.

Ответственный за выпуск *Л. В. Кобзева*

Компьютерная верстка *М. А. Филимонова*

ИД № 04689 от 28.04.2001 г.
Подписано в печать 06.03.2018 г.
Бумага офсетная. Формат 60x841/8. Гарн. литер. Печать цифровая
Усл. печ. л 5,00 Уч.-изд. л. 5,42 Тираж 33 экз.

Адрес редакции: 125190, Россия, г. Москва, ул. Усиевича, д. 20
Тел. (499) 155-44-95

ВНИМАНИЮ ПОДПИСЧИКОВ!

С 2018 года возобновляется издание информационного бюллетеня «Иностранная печать об экономическом, научно-техническом и военном потенциале государств-участников СНГ и технических средствах его выявления» серии «Экономический и научно-технический потенциал» (56741) взамен информационного бюллетеня «Экономика и управление»

Периодичность выхода – 12 номеров в год. Объем 48 уч.-изд. л. в год.

В бюллетене освещаются материалы иностранной печати по широкому спектру вопросов, касающихся сфер экономического и научно-технического развития России и стран СНГ: общие вопросы, финансы, промышленность, рынки, сельское хозяйство, космос, транспорт и связь, природные ресурсы, трудовые ресурсы, внешние торгово-экономические и научные связи

Оформить подписку на информационный бюллетень, начиная с любого номера, можно в ВИНИТИ РАН по адресу: 125190, Россия, Москва, ул. Усиевича, 20,

Телефоны: (499) 151-78-61; (499) 155-42-85

Факс: (499) 943-00-60;

E-mail: contact@viniti.ru; sales@viniti.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