Д.А. Девяткин, Ю.Э. Даник, И.А. Тихомиров, А.В. Швец

Метод выделения направлений научных исследований на основе анализа полных текстов публикаций[1]

Аннотация. В статье предложен новый метод выявления направлений научных исследований, а также способ автоматической настройки его параметров. В методе вводится функция оценки близости научных публикаций, которая учитывает тематическую близость текстов, ссылочную связанность и соавторство. При расчете тематической близости текстов также используются тезаурусы и онтологии, что в отличие от других методов позволяет учитывать специфику предметных областей и выявлять многословные словосочетания с синтаксическими и семантическими связями, наиболее точно характеризующие направления научных исследований.

Ключевые слова: наукометрический анализ, поддержка научной деятельности, направления научных исследований, кластеризация коллекций научных публикаций.

Введение

В условиях лавинообразного роста количества публикаций в российских и зарубежных журналах перед исследователями стоит проблема поиска релевантных статьей по отдельным направлениям/темам. Основными инструментами, обеспечивающими механизм оценки и анализа сверхбольших коллекций научных публикаций, являются модули системы Web of Knowledge – ISI Essential Science Indicators (ESI) [1] и InCites [2]. Модули ESI и InCites позволяют получить информацию о некоторых ключевых научных исследованиях в мире, выявить тенденции развития науки. Метод, используемый в ESI, заключается в анализе ссылочной структуры публикаций, выявлении наиболее цитируемых публикаций и построении на их основе фронтов исследований – тематических направлений, соответствующих передовому рубежу современных исследований. Модуль InCites также применяет метод анализа ссылочной информации публикаций и собирает статистические показатели, которые позволяют определять для различных организаций предметные области с наибольшим числом публикаций. Такие методы не подходят для объективной оценки состояния и динамики научных направлений, поскольку позволяют анализировать лишь небольшое число передовых направлений. К тому же методы не подходят для анализа российских исследований, поскольку модули системы Web of Knowledge используют зарубежные базы данных, в которых представлено небольшое число российских научных публикаций (менее 5% от всего объема публикаций, выпускаемых в России). Таким образом, существует объективная потребность в разработке новых инструментов анализа направлений научных исследований.

1. Методы выделения направлений научных исследований

На текущий момент можно выделить несколько групп методов выделения научных направлений: методы анализа цитатных сетей [3, 4], методы полнотекстовой кластеризации документов и гибридные методы.

Рассмотрим некоторые из последних исследований в области разработки методов класте-

[1]Работа выполнена при финансовой поддержке РФФИ (грант №14-29-05008-офи_м).

ризации публикаций и выделения научных направлений.

В [3] определение перспективных направлений исследований проводится путем анализа совместной встречаемости слов, т.е. измерения близости текстов публикаций по использованной в них терминологии. Демонстрируется применение данного метода для выявления закономерностей и тенденций в области информационной безопасности. Источником данных послужила база SCI [1], ключевые слова извлекались из списка терминов, выделенных авторами публикаций. Для того чтобы проследить динамические изменения в области информационной безопасности, авторы публикации предлагают целый набор методов технического картографирования или технологий разметки публикаций. В результате сделан вывод о том, что в области информационной безопасности имеется некоторый устоявшийся набор тем исследований, и в то же время быстро возникают новые направления.

Другой подход – анализ графа цитирования научных публикаций – применялся в [5] для обнаружения новых научно-исследовательских фронтов среди огромного количества научных работ в области регенеративной медицины. Авторы предлагают разделять сеть цитирования на кластеры с использованием метода топологической кластеризации, отслеживать позиции документов в каждом кластере и визуализировать сеть цитирования и список ключевых слов для каждого кластера. Показано, что анализ среднего возраста публикаций и отношений родитель-потомок (образование подкластеров) для каждого из выделенных кластеров может быть полезным при выявлении последних тенденций в исследуемой области науки. Кроме того, наличие новых кластеров знаний можно отслеживать с использованием таких топологических мер, как степень внутрикластерной связности и коэффициент участия. Предполагается, что публикации по одной теме должны иметь высокую степень связности и наоборот. Результаты исследования показывают состоятельность данного метода для обнаружения зарождающихся исследовательских фронтов в регенеративной медицине, что подтверждают и эксперты в предметной области. Также авторы публикации делают предположение о том, что можно отследить появление в будущем значимых документов, с высоким индексом цитирования, при помощи анализа показателя betweenness centrality в графе ссылочности. Этот показатель характеризует влияние того или иного узла на распространение информации внутри графа.

Из числа публикаций, посвященных использованию гибридных методов кластеризации и выделения научных направлений, следует отметить статью [6]. В этом исследовании описывается процесс кластеризации научных публикаций, а затем выделение новых научных направлений и анализ результатов с использованием библиометрических методов. Применяемые в работе гибридные методы кластеризации основаны на оценке библиографической связности публикаций, на интеллектуальном анализе текстов и на выделении значимых публикаций. При кластеризации используется комплексный показатель близости научных публикаций, представляющий собой взвешенную сумму двух компонент: оценки цитирования и оценки текстов. Варьирование весов этих компонент для тонкой настройки метода применительно к фундаментальным, прикладным и социальным наукам доказало свою состоятельность. Выявление новых научных направлений в исследовании ведется путем анализа перекрестного цитирования. Отслеживание кросс-цитирования между наиболее важными узлами в построенной гибридной сети (на основе цитирования и лексического анализа публикаций) за разные периоды времени позволяет получить информацию о возможном появлении новых направлений. По мнению авторов, новые направления характеризуются публикационной активностью, цитируемостью и международным сотрудничеством. Анализ сетей сотрудничества стран наиболее активных в рассматриваемых дисциплинах в сочетании с оценкой набора стандартных наукометрических показателей, образуют основу для анализа обнаруженных новых научно-исследовательских направлений. Некоторые темы, например, могут носить глобальный характер и требовать интенсивного международного взаимодействия, при этом по некоторым направлениям могут иметь особую важность и региональные аспекты, что в свою очередь приводит к образованию подкластеров и географической поляризации.

Каждый из вышеперечисленных методов имеет свои недостатки и ограничения (например, необходимость доступа к цитатным базам,

предварительное извлечение частей документов), и получаемые результаты далеки от объективных – уровень ошибок классификации достаточно высок. Наиболее эффективными являются гибридные методы [6], которые используют для определения направлений научных исследований анализ текстов и цитирований, а также применяют методы библиографического исследования тематик.

2. Гибридный метод выделения направлений научных исследований

Гибридный метод выделения направлений научных исследований, объединяет в себе принципы кластеризации на основе полнотекстового анализа документов, анализа ссылочной связности и оценки близости научных коллективов сравниваемых статей. Разработанный авторами данной статьи метод предполагает расчет ряда оценок сходства для каждой пары документов из коллекции с последующим вычислением интегрального показателя сходства.

Для расчета интегрального показателя сходства $O(d_1,d_2)$ может использоваться линейная свертка:

$$O(d_1,d_2)=\sum_i z_i O_i(d_1,d_2)).$$

Здесь $O_i(d_1,d_2)$ – оценка сходства, d_1 и d_2 – текстовые документы, z_i – настроечные параметры.

На основе этого показателя принимается решение о распределении документов по научным направлениям. В качестве критерия принятия решения вводится пороговое значение близости двух документов H_a.

Был также получен функционал оценки качества разбиения анализируемой коллекции на научные направления и предложен метод настройки параметров алгоритма путем минимизации полученного функционала качества разбиения.

Рассмотрим разработанный метод более подробно. В качестве входных данных алгоритма выступает коллекция полных текстов научных документов, принадлежащих одной отрасли науки. Обозначим это множество из n анализируемых документов как $D=\{d_s, s=1..n\}$.

Каждый научный документ из анализируемой коллекции представлен в виде кортежа $d=<T(d,D),L(d),A(d)>$. Раскроем подробно элементы этого кортежа:

$T(d,D)$ – набор кортежей $t(w,d,D)$ вида $<Fplc(w_i,d),v(w_i,d,D),w_i,\ i = 1..|W(d)|>$, характеризующих лексико-синтаксические дескрипторы w_i документа d (т.е. ключевые слова и словосочетания документа d). Эти дескрипторы выделяются из документа при помощи лингвистических методов обработки текста, которые позволяют выделять семантические и синтаксические связи между словами и формировать словосочетания. Для синтаксического и семантического анализа используется реляционно-ситуационная модель текста, специализированные семантические словари, тезаурусы и анализаторы [7, 8]. Здесь $W(d)$ – множество лексико-синтаксических дескрипторов документа d, $v(w_i,d,D)$ – вес лексико-синтаксического дескриптора w_i в документе d вычисленный по формуле $ltf\text{-}idf(w_i,d,D)$. Также веса дескрипторов могут корректироваться: если слово или словосочетание из лексико-синтаксического дескриптора входит в тезаурус, или является концептом онтологии анализируемой предметной области, то вес его дескриптора повышается. Таким образом, достигается повышение значимости слов и словосочетаний, характерных для исследуемой предметной области. $Fplc(w_i,d)$ – функция, возвращающая номер первого вхождения дескриптора w_i в документ d. Этот показатель используется для расчёта приоритета слова в зависимости от порядка его появления в тексте: предполагается, что чем раньше слово появляется в статье, тем больший вес оно имеет. Согласно правилам и логике оформления научных публикаций, именно в начале статьи приводится обоснование актуальности, новизны исследования, описывается суть работы и характер полученных результатов. После этого уже следуют описание использованных методов и обзор предыдущих исследований.

Помимо показателя $T(d,D)$ научная публикация d описывается множеством из m публикаций из коллекции D, на которые она ссылается, т.е. показателем $L(d)\subset D,|\ L(d)|=m$ (множество ссылок на другие документы).

Документ d также характеризуется множеством из k авторов, работавших над выпуском публикации: $A(d)=\{<a_y,h_f,r_{a_y h_f}>, y=1...k\}$, где a_y – автор, $\{h_f, f=1..q\}$– множество рабочих коллективов, r_{ayhf} – роль автора в коллективе h_f, $f = 1..q$, $r\in\{ld,c,lf,g\}$. Предполагается, что

в коллективе выделяются следующие категории: «лидеры», «ядро», «гости» и «листья».

«Ядро» включает в себя часто публикующихся авторов коллектива, «лидеры» – авторов «ядра», на которых больше всего ссылаются в работах этого коллектива, «листья» – авторов, имеющих небольшое количество публикаций с коллективом, но не имеющих публикаций с другими коллективами, «гости» – авторов, входящих в ядро другого коллектива, но имеющих публикации с данным коллективом.

Множество коллективов, работавших над статьей *d,* обозначается

$$H(d) = \{h_f \mid \exists < a_y, h_f, r_{a_y h_f} > \in A(d)\}.$$

В качестве функции оценки близости научных публикаций предлагается взвешенная сумма оценок по каждому из трех вышеперечисленных критериев.

$$O(d_1,d_2) = z_1 * O_t(d_1,d_2) + z_2 * O_l(d_1,d_2) + z_3 * O_c(d_1,d_2)$$

где $O_t(d_1,d_2)$ – оценка тематической близости текстов публикаций, $O_l(d_1,d_2)$ – оценка ссылочной связности публикаций, $O_c(d_1,d_2)$ – оценка близости публикаций по соавторству. Веса оценок обозначены соответственно z_t, z_l, z_c. Веса будут подбираться автоматически с использованием методов оптимизации, так чтобы обеспечить достижение минимума функционала качества, который будет введен ниже. В качестве критерия отнесения двух документов к одному кластеру выступает решающее правило $O(d_1,d_2) > H_a$, где H_a – некое пороговое значение, которое также будет подбираться в ходе обучения алгоритма.

Оценка тематической близости текстов публикаций представлена в виде

$$O_t(d,d') = \frac{\sum_{w \in W(d) \cap W(d')} |ltf(w,d)\,p(t(w,d,D),d) - ltf(w,d')\,p(t(w,d',D),d')|\,idf(w,D)}{\sum_{w \in W(d) \cap W(d')} |ltf(w,d)\,p(t(w,d,D),d) + ltf(w,d')\,p(t(w,d',D),d')|\,idf(w,D)}$$

Здесь вычитаемое – модифицированное манхэттенское расстояние между множествами лексико-синтаксических дескрипторов сравниваемых документов [9]. Суть модификации состоит в умножении веса каждого лексического дескриптора на функцию приоритета этого дескриптора *p(t,d)*. В рамках дальнейшей работы предполагается экспериментально проверить несколько вариантов реализации функции *p(t,d)*:

1. $p(t(w,d,D),d) = \frac{1}{Ae^{B*Fplc(w,d)}}$ – приоритет *wi, i=1..|W(d)|*, в документе *d* в зависимости от позиции в тексте, где *A* и *B* – настроечные константы.

2. $p(t(w,d,D),d) = 1 - \frac{(Fplc(w,d)-1)}{|\Omega(d)|-1}$, где $\Omega(d)$ - общая совокупность числа вхождений терминов в документе *d*. Иными словами, документ – это вектор дескрипторов длины $\Omega(d)$; *Fplc(w,d)* – порядковый номер первого появления слова/словосочетания в документе, изменяется от 1 до $|\Omega(d)|$.

Оценка ссылочной связности публикаций $O_l(d_1,d_2)$ рассчитывается как относительное число совместных ссылок документов d_1 и d_2.

$$O_l(d_1,d_2) = \frac{|L(d_1) \cap L(d_2)|}{|L(d_1) \cup L(d_2)|}$$

Для определения близости документов по соавторству предлагается следующая несимметричная оценка, в которой учитывается, сколько авторов документа *d* состоит в одних и те же коллективах, что и авторы документа *d'*.

Вводится множество

$$G(d,d') = \left\{ a = <a_y, h_f, r_{a_y h_f}> \in A(d) \mid h_f \in H(d') \right\}$$

$$O_c(d,d') = \frac{\sum_{a \in G(d,d')} u_{r_{a_y h_f}}}{\sum_{a \in A(d)} u_{r_{ld}}}$$

где $u_r \in U=\{u_{ld},u_c,u_{lf},u_g\}$ – в зависимости от роли автора в коллективе.

Используя метод аналитической иерархии (МАИ) [10] определяются значения весовых коэффициентов $U=\{u_{ld},u_c,u_{lf},u_g\}$, $\sum_r u_r = 1$, где $r \in \{ld, c, lf, g\}$ – роли авторов в коллективе (лидер, ядро, лист, гость). Для определения весовых коэффициентов по МАИ требуется при-

влечь экспертов, которые проведут попарные сравнения важности категорий по 9 балльной шкале. На основе этих оценок будет построена обратно-симметричная матрица сравнений и рассчитан вектор приоритетов категорий (с проверкой согласованности оценок экспертов). В дальнейшем эти веса будут учитываться при расчете числа близких по коллективам авторов для двух документов. Веса могут варьироваться от коллекции к коллекции.

Используем полученную меру сходства научных документов для кластеризации набора *D* разработанным методом кластеризации больших коллекций текстовых документов. На основе интегральной оценки близости и выбранного порога близости H_a, производится разбиение документов на кластеры.

Далее опишем метод автоматического подбора параметров для гибридного метода выделения направлений научных исследований. Критерием качества кластеризации выступает значение функционала качества разбиения [11]:

$$Q(W,B,D)=\frac{2\sum_{g,v}^{n}|b_{gv}-w_{gv}|}{n(n-1)}\to\min,$$

где *W* – матрица разбиения коллекции документов *D* на кластеры, *B* – матрица эталонного разбиения коллекции *D, n* – размер коллекции документов D. Элемент матрицы разбиения a_{ij} = 1, если документы i и j находятся в одном кластере.

Определение параметров работы метода, выполняется при помощи минимизации этого функционала с ограничениями $0<z_i<1$, $0<H_a<1$, $0<H_b<1$, $i = 1..3$, для этого используется безградиентный метод оптимизации BOBYQA [12].

Разработанный гибридный метод выделения направлений научных исследований использует не только информацию о близости текстов научных публикаций и общих ссылках, но и информацию о соавторстве, что улучшает качество выделения научных направлений. Предложенный способ автоматической настройки параметров разработанного метода позволяет избежать эмпирического подбора.

3. Анализ динамики направлений научных исследований

Анализ динамики развития научных направлений позволяет выявить зарождающиеся научные направления, определить тренды развития отрасли науки. В ходе анализа период, которому соответствует анализируемая выборка научных документов, разбивается на n пересекающихся окон. При помощи гибридного метода производится выделение научных направлений внутри каждого из окон.

Пересечение окон требуется в случае резкого «сдвига» тематики научного направления. Величина окна и пересечения окон подбирается эмпирически для каждой отрасли науки.

Между направлениями, входящими в соседние окна, формируются ссылочные связи, а также связи на основе соавторства и близости текстов. С использованием этих связей для каждого направления из окна *i* выполняется поиск предка в окне *i-1*. Строится динамика развития направлений, выделяются направления, ответвляющиеся от родительского, объединяющиеся научные направления, а также зарождающиеся научные направления, для которых не нашлось предка в окне *i-1*.

После проведения процедуры разбиения документов на кластеры можно определить значимость документа в конкретном научном направлении. Для этого используются следующие показатели [4, 13].

1. Связность узла внутри научного направления

$$z=\frac{K_i-\overline{K_{s_i}}}{\sigma_{K_{s_i}}},$$

где K_i –число ссылок из документа i на документы в направлении s_i, $\overline{K_{s_i}}$- среднее значение K по всем узлам s_i, $\sigma_{K_{s_i}}$- стандартное отклонение K в s_i. Если $z \geq 2.5$, то документ можно отнести к важным документам кластера.

2. Коэффициент участия в других направлениях научных исследований

$$P=1-\sum_{s=1}^{N_M}\left(\frac{K_{is}}{k_i}\right)^2,$$

где K_{is}- число ссылок в документе i на документы в кластере s, k_i – число ссылок в документе i. Направления с большим коэффициентом взаимного участия могут быть объединены.

3. Коэффициент участия в данном направлении научных исследований

$$P_s=\sum_{s=1}^{N_M}(\frac{K_{is}}{k_i})^2$$

показывает, какую долю среди всех ссылок публикации составляют ссылки на публикации научного направления.

После вычисления указанных оценок становится возможным отсортировать документы внутри научного направления, оценить степень их участия в направлении, выбрать наиболее значимые для этого направления публикации.

Заключение

Разработан новый гибридный метод выделения направлений научных исследований. В рамках метода используются одновременно данные статистического анализа и лингвистическая информация, тезаурусы и онтологии. Это позволяет учитывать специфику предметных областей и выделять узкие направления научных исследований, что выгодно отличает метод от аналогов, предоставляющих лишь широкие области научных исследований.

Разработанный гибридный метод выделения направлений научных исследований позволяет проводить кластеризацию без использования цитатных баз и предварительного отбора статей по ключевым словам. Также преимуществом этого метода является возможность группировать содержательно близкие тексты, не обязательно цитирующие друг друга. При расчете близости текстов публикаций используется реляционно-ситуационная модель текста [7] и модифицированное манхэттенское расстояние между множествами лексико-синтаксических дескрипторов сравниваемых документов. Благодаря этому в кластеры не попадают лишние статьи, не относящиеся к направлению, как это происходит при анализе цитатных сетей. Выделение направлений путем кластеризации статей позволяет проводить дальнейший анализ динамики направлений, в отличие от методов, предполагающих построение направлений в отрыве от публикаций, лишь на основе статистических показателей совместной встречаемости слов.

Дальнейшим направлением исследований является экспериментальная проверка метода на больших массивах научных публикаций и широком спектре предметных областей с использованием системы Exactus Expert [14, 15].

Литература

1. ISI Essential Science Indicators, http://thomsonreuters.com/essential-science-indicators
2. InCites, http://thomsonreuters.com/incites
3. Lee W. H. How to identify emerging research fields using scientometrics: An example in the field of Information Security //Scientometrics. – 2008. – Т. 76. – №. 3. – С. 503-525.
4. Shibata N. et al. Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications //Technological Forecasting and Social Change. – 2011. – Т. 78. – №. 2. – С. 274-282.
5. Chen, C., Hu, Z., Liu, S., and Tseng, H.: Emerging Trends in Regenerative Medicine: a Scientometric Analysis in CiteSpace. Expert Opin. Biol. Ther., 12(5), 593–608, (2012)
6. Chen C. et al. Emerging trends in regenerative medicine: a scientometric analysis in CiteSpace //Expert opinion on biological therapy. – 2012. – Т. 12. – №. 5. – С. 593-608.
7. Osipov G. et al. Relational-situational method for intelligent search and analysis of scientific publications //Proceedings of the Integrating IR Technologies for Professional Search Workshop. – 2013. – С. 57-64.
8. Freeling – an open source suite of language analyzers, http://nlp.lsi.upc.edu/freeling
9. Девяткин Д.А., Суворов Р.Е., Соченков И.В. Распределенная тематическая кластеризация текстовых документов в системе Exactus Expert // Труды пятой международной конференции «Системный анализ и информационные технологии» (САИТ-2013). Красноярск, 2013. Т. 1, С 200-207.
10. Saaty, T.L. The Analytic Hierarchy Process/T.L. Saaty.– New York: McGraw-Hill International, 1980.
11. Миркин Б. Г., Черный Л. Об измерении близости между различными разбиениями конечного множества объектов //Foresight. – 2013.
12. Powell M. J.D. The BOBYQA algorithm for bound constrained optimization without derivatives //Cambridge NA Report NA2009/06, University of Cambridge, Cambridge. – 2009.
13. Guimera R., Amaral L.A. N. Functional cartography of complex metabolic networks //Nature. – 2005. – Т. 433. – №. 7028. – С. 895-900.
14. Тихомиров И.А., Смирнов И.В., Соченков И.В., Девяткин Д.А., Шелманов А.О., Зубарев Д.В., Швец А.В., Лешкин А.В., Суворов Р.Е. Exactus Expert: Поисково-аналитическая система поддержки научно-технической деятельности // Труды тринадцатой национальной конференции по искусственному интеллекту с международным участием КИИ-2012. Белгород: Изд-во БГТУ, 2012. т. 4. - С. 100-108.
15. Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov, Ilya Sochenkov, Artem Shelmanov, Alexander Shvets Information Retrieval for R&D Support // Professional Search in the Modern World. Lecture Notes in Computer Science Volume 8830, 2014, pp 45-69

Девяткин Дмитрий Алексеевич. Младший научный сотрудник Института системного анализа Российской академии наук. Окончил Рыбинскую государственную авиационную технологическую академию в 2011 году. Автор 7 научных работ. Область научных интересов: адаптивные методы дистанционного обучения, классификация и кластеризация текстов, интеллектуальный веб-краулинг, искусственный интеллект.
E-mail: devyatkin@isa.ru.

Даник Юлия Эдуардовна. Аспирант Института системного анализа Российской академии наук. Окончила Национальный исследовательский университет Высшая школа экономики в 2014 году. Автор 1 научной работы. Область научных интересов: сингулярные возмущения, модели экономической динамики, оптимальное управление, базы и хранилища данных, методы оценки инвестиций в информационные системы.
E-mail: juliet.d.e.777@mail.ru

Тихомиров Илья Александрович. Ведущий научный сотрудник Института системного анализа Российской академии наук. Окончил Рыбинскую государственную авиационную технологическую академию в 2002 году. Кандидат технических наук. Автор 55 научных работ. Область научных интересов: искусственный интеллект, компьютерная лингвистика, поисковые системы, информационная безопасность, интернет-системы. E-mail: tih@isa.ru.

Швец Александр Валерьевич. Младший научный сотрудник Института системного анализа Российской академии наук. Окончил Сибирский федеральный университет в 2011 году. Автор 15 научных работ. Область научных интересов: компьютерная лингвистика, математическое моделирование, методы оптимизации, искусственный интеллект. E-mail: shvets@isa.ru