Выявление направлений исследований и научных коллективов на основе анализа полнотекстовых коллекций научных публикаций

Identifying research areas and corporate author based on the analysis of full-text scientific collections

Д. А. Девяткин, А. В. Швец, И. А. Тихомиров
Институт системного анализа ФИЦ ИУ РАН,
Москва, Россия

Dmitry Devyatkin, Alexander Shvets and Ilya Tikhomirov
Institute for Systems Analysis FRC CSC RAS,
Moscow, Russia

В докладе представлен гибридный подход к автоматическому выделению направлений исследований и научных коллективов, основанный на анализе полных текстов статей. Этот подход был успешно апробирован на больших полнотекстовых коллекциях научных публикаций по нескольким предметным областям. Показаны преимущества представленного подхода перед стандартными методами, основанными на анализе баз цитирования.

The authors proposes to apply a hybrid approach to identifying research area and author communities, based on the analysis of full texts. This approach has been successfully applied to large-scale multi-field science collections in several scientific fields. The advantages of the proposed method over the standard citation-based ones is demonstrated.

Первичный анализ научной области часто подразумевает выявление направлений исследований, а также устойчивых авторских коллективов. Для автоматизированного решения этой задачи обычно используются такие инструменты как Thomson Reuters ESI[1], Thomson Reuters InCites[2], Elsevier SciVal[3] и ScienceResearch[4]. Выделение научных направлений и коллективов в этих системах основано на анализе ссылочной структуры зарубежных цитатных баз. Этот подход имеет следующие недостатки.

1. Отсутствует возможность выявления новых, недавно появившихся научных направлений. Такие направления характеризуются небольшим числом вовлеченных исследователей и низким уровнем цитирования.
2. Фрагментация выявленных направлений. Низкая культура цитирования в России часто приводит к тому, что представители разных научных школ, работающих в одном научном направлении, слабо цитируют друг друга.
3. Неточности при соотнесении имен авторов с реальными исследователями могут приводить к построению некорректных научных коллективов.
4. Формирование искаженного общего представления об анализируемой области. Работы российских ученых по некоторым направлениям исследований слабо представлены в зарубежных цитатных базах: туда попадает менее 10% от всех российских публикаций.

Преодолеть указанные недостатки могло бы использование дополнительной информации, содержащейся в полных текстах публикаций. Поэтому возникла потребность в инструменте для выявления научных направлений и авторских коллективов на основе анализа полнотекстовых коллекций публикаций отечественных ученых.

Все современные подходы к выявлению направлений научных исследований основываются на различных методах кластеризации больших данных, однако отличаются используемой мерой близости публикаций: сходство по наличию совместных цитирований [1, 2], тематическая близость текстов [3, 4], гибридные меры [5, 6]. Наиболее перспективным считается [7] использование

[1] http://thomsonreuters.com/essential-science-indicators
[2] http://thomsonreuters.com/incites
[3] http://www.elsevier.com/online-tools/research-intelligence/products-and-services/scival
[4] http://www.scienceresearch.com

гибридной меры близости публикаций. В предлагаемом новом подходе такая мера вычисляется на основе трех компонентов: близость по тематическому сходству текстов O_t, близость по присутствию совместных цитирований O_l и по наличию общих авторов O_a (см. Рис. 1). Информация, необходимая для вычисления гибридной меры близости публикаций, автоматически извлекается из полных текстов публикаций.

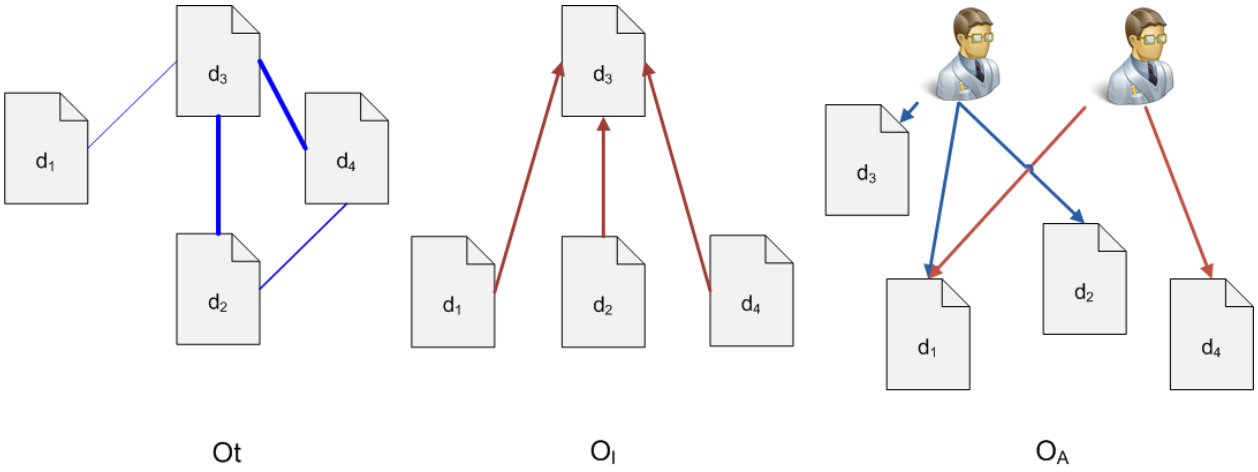

Рис. 1– Меры близости научных публикаций

Представленный подход был применен для выявления научных направлений в [4] на коллекции зарубежных публикаций из предметной области «Регенеративная медицина» и в [8] на работах российских исследователей по теме «Искусственный интеллект и принятие решений». Научные направления, полученные в результате этих экспериментов, представлены в таблицах 1 и 2.

Таблица 1– Научные направления,
построенные по статьям предметной области «Регенеративная медицина»

Название направления	Ключевые слова и словосочетания
Brainstroke	stroke, brdu test, behavioral, endothelial cell, cell brdu, psa-ncam, cortical cell, reactive, regeneration neuronal, endogenous, neurogenesis, endostatin, cortical, expansion nonhematopoietic, ischemicneuron
Chimeric antigen receptor cell therapy	pbls, receptor chimeric, cd19-speci, immunity, antitumor, adoptive ifn, malignancy b-cell, aapcs, protein fusion, infusion cell
Inducedpluripotentstemcells	gene signature, photoreceptor, epithelium retinal, retinal, suppressor, tumor, cell ips, technology ips, retinal cell, ips-derived, ipscs pigmented
Woundhealingburns	wound chronic, wound heal, keratinocyte, epidermal cell, fusenig keratin, epidermis region, bulge cotsarelis, stratus, skin mouse, migration keratinocyte, sheath, root size, wound

Таблица 2 – Научные направления, построенные по статьям из предметной области «Искусственный интеллект и принятие решений»

Название направления	Ключевые слова и словосочетания
Многокритериальная оптимизация	выбираемый вектор, граница Парето, задача выбора, иррефлексивный, многокритериальная оптимизация, многокритериальная среда, многокритериальный выбор, множество Парето, недоминировать, неизвестное множество, отношение предпочтения, Парето-оптимальный, предпочтение лпр, принцип Эджворта-Парето, проблемы сужения, сужение множества.
Семантический поиск	именная группа, интеллектуальный поиск, информационный поиск, коллекция документов, компьютерная лингвистика, семантический поиск, синтаксический анализ, слова запроса.
ДСМ-метод порождения гипотез	каузальная полнота, методы миля, операция сходства, дсм-рассуждение, фактическое противоречие, синтез процедур, дсм-метод, база фактов, абдуктивный, аксиома полноты, автоматическое порождение, правдоподобный вывод.
Многокритериальная порядковая классификация	вербальный анализ, метод классификации, номинальная классификация, многокритериальный, порядковая классификация, непротиворечивый, задача классификации, шкала критериев.
Интегрированные экспертные системы	задачно-ориентированный, ат-технология, задачно-ориентированная методология, прикладные иэс, инструментальный комплекс, поддержка построения, инструментальная база, многоагентный, модель представления, технология построения.
Кластерный анализ и распознавание образов	выбор метрики, Евклид-Махаланобис, кластеризация, машинная графика, размещение кластеров, сеть кохонена, искусственный интеллект, число кластеров, задача кластеризации, задача распознавания, матрица ковариации.
Автоматизация полета вертолета	автоматизация, вертолет, летательный аппарат, полет.

Корректность полученных направлений исследований была подтверждена экспертами в соответствующих научных областях.

Основным подходом к выделению научных коллективов является использование методов кластеризации на графах соавторства [9]. Однако неправильное соотнесение имен авторов с реальными исследователями может приводить к искажению графа соавторства и, как следствие, к некорректным результатам работы. Как и для задачи выявления научных направлений, повысить качество выделения коллективов можно при помощи гибридной меры близости авторов. Экспериментальная проверка такого гибридного подхода производилась на полнотекстовой коллекции научных публикаций по физико-математическим наукам. Фрагмент одного из выявленных научных коллективов представлен на рис. 2.

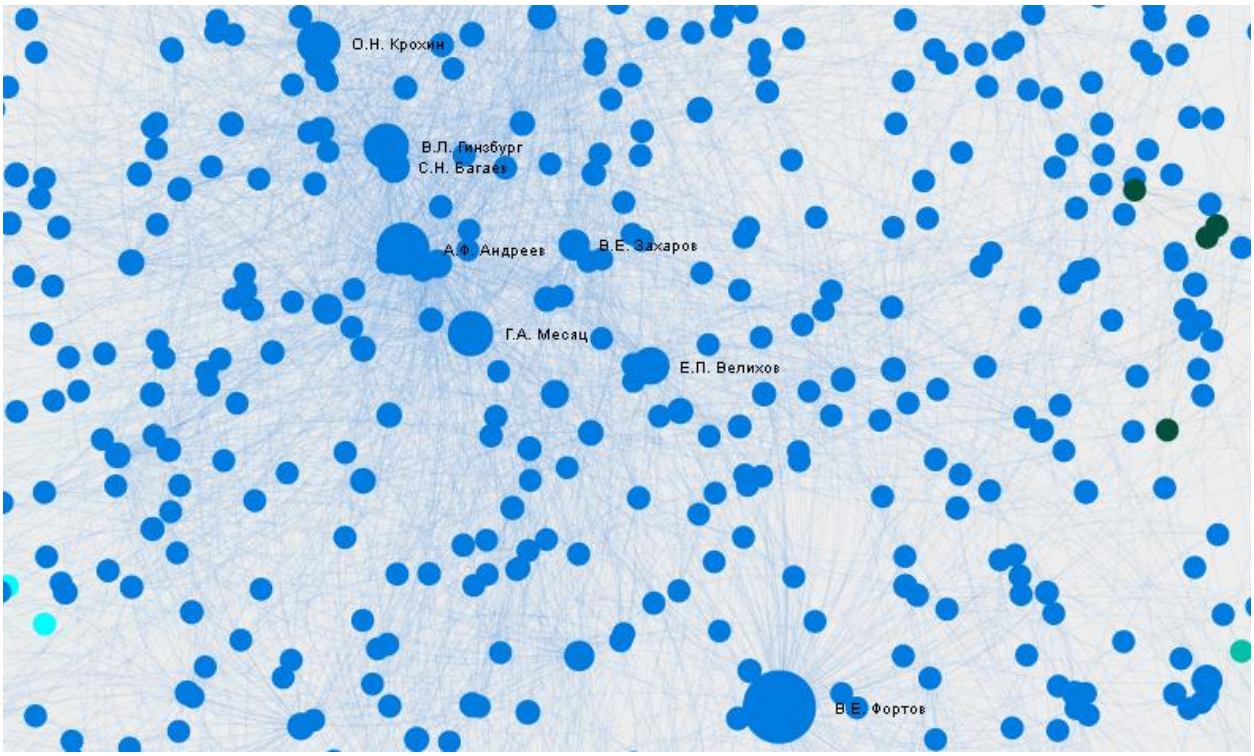

Рис. 2 – Фрагмент авторского коллектива по физико-математическим наукам

Таким образом, работоспособность предложенного подхода была подтверждена экспериментально. К преимуществам этого подхода можно отнести:

1. возможность выделения новых, недавно появившихся научных направлений;
2. высокое качество выделения направлений и коллективов по сравнению с аналогами, что достигается за счет сохранения синтаксических и лексических связей между словами текстов и применения гибридной метрики близости публикаций;
3. не требуется наличие баз цитирования и соавторства: необходимая информация автоматически извлекается из полных текстов документов.

Новый подход позволяет обойти ограничения, с которыми сталкиваются методы анализа баз цитирования. Он был реализован как один из инструментов поисково-аналитической системы Exactus Expert [10].

Литература

1. Lee W. H. How to identify emerging research fields using scientometrics: An example in the field of Information Security //Scientometrics. – 2008. – Vol. 76(3). – p. 503-525.
2. 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. – Vol. 78(2). – p. 274-282.
3. Min Song, Su Yeon Kim. Detecting the knowledge structure of bioinformatics by mining full-text collections. Scientometrics, 96(1), 183-201, (2013)
4. Shvets A. V. et al. The study of systems and methods for scientometric analysis of scientific publications //Scientific and Technical Information Processing. – 2015. – Vol. 42(5). – p. 359-366.
5. Glänzel, Wolfgang. Bibliometric methods for detecting and analyzing emerging research topics. El profesional de la información, 2012, vol. 21, n. 1, pp. 194-201.
6. Thijs B., Glänzel W., Meyer M. Using noun phrases extraction for the improvement of hybrid clustering with text-and citation-based components. The example of “Information System Research //Proc. of the Workshop Mining Scientific Papers: Computational Linguistics and Bibliometrics, 15th International Society of Scientometrics and Informetrics Conference (ISSI), Istanbul, Turkey: http://ceur-ws. Org. – 2015.
7. 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(9). – p. 1759-1767.

8. А.В. Швец, Д.А. Девяткин, Д.В. Зубарев, И.А. Тихомиров, О.Г. Григорьев Анализ качественных и количественных характеристик журнала «Искусственный интеллект и принятие решений» // Искусственный интеллект и принятие решений. 2015. – № 4. – с. 89-100.
9. Wu Y. et al. Robust local community detection: on free rider effect and its elimination //Proceedings of the VLDB Endowment. – 2015. – Vol. 8(7). – p. 798-809.
10. Osipov G. et al. Exactus Expert—Search and Analytical Engine for Research and Development Support //Novel Applications of Intelligent Systems. – Springer International Publishing, 2016. – p. 269-285.