OSTIS-2016
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

РОССИЯ В КОНТЕКСТЕ МИРОВЫХ ЦЕНТРОВ КОМПЕТЕНЦИЙ И ПРЕВОСХОДСТВА

Хорошевский В.Ф.[*], Ефименко И.В.[**]

[*]*Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН, г. Москва, Россия*
khor@ccas.ru

[**]*Факультет гуманитарных наук НИУ ВШЭ, г. Москва, Россия*
[**] *SemanticHub, г. Москва, Россия*
iefimenko@hse.ru

В работе представлен наукометрический анализ ландшафта исследований и разработок, выполняемых российскими специалистами в соавторстве со специалистами из других стран, с использованием семантических технологий. Целью исследования является идентификация российских и мировых центров компетенций и превосходства в различных предметных областях на основе специализированных инструментов, разработанных и реализованных авторами.
Ключевые слова: наукометрия; центры превосходства и компетенций; семантические технологии; извлечение информации из текстов

Введение

Задачавыявленияцентров компетенцийи превосходства в прорывных научно-технологических направлениях на основе мониторинга разнородных информационных ресурсовпредставляет интерес как с научно-исследовательской, так и с прикладной точки зрения[Акоев и др., 2014].При этом классический подход к решению этой задачи, основанный на использовании статистических методов библиометрического анализа наиболее значимых публикаций, как показывает целый ряд последних исследований, целесообразно развить за счет использования семантических технологий[Boyacketal., 2013; Efimenkoetal., 2014; Хорошевский и др., 2015].

Актуальность и практическая значимость соответствующих исследований определяется, с одной стороны, потребностьюлюбого государства в достоверной информации для определения своей научно-технической политики, а с другой – желанием бизнеса знать, инвестиции в какие направления и коллективы наиболее перспективны.

Особую роль результаты выявления центров компетенций и превосходства в прорывных научно-технологических направлениях играют для позиционирования любой страны в современном постиндустриальном обществе.

С учетом вышесказанного, в настоящей работе представлен наукометрический анализ ландшафта исследований и разработок, выполняемых российскими специалистами в соавторстве со специалистами из других стран, с использованием семантических технологий, что, по нашему мнению, может обеспечить идентификацию российских и мировых центров компетенций и превосходства в различных предметных областях.

Дополнительноеисследование было проведено для выявления технаучно-технических направлений, где Российская Федерация занимает лидирующие позиции, а также для идентификации стран-партнеров России по исследованиям и разработкам на основе сведений о соавторстве.

1. Методы, средства и данные для наукометрии ландшафтов исследований и разработок

1.1. Предварительные замечания

Как известно, современные библиометрические базы данных (WebofScience, Scopus, GoogleScholarи др.) предоставляют пользователям не только поисковые функции, но и целый ряд полезных аналитических инструментов, позволяющих изучать развитие научных направлений, приоритеты стран в

области исследований и разработок, интересы отдельных научно-исследовательских организаций, авторов и многое другое. При этом важную роль в данной области играют не только общие, но и специализированные инструменты, впечатляющий обзор которых дан в работе [Coboetal., 2011].

Вместе с тем, практически все инструменты характеризуются рядом ограничений, связанных как с особенностями самих библиометрических баз, так и с объективной организацией научно-технической информации.

Так, например, в широко распространенных инструментах библиометрии до сих пор до конца не решена проблема объединения синонимов даже для таких простых случаев, как наименования государств (Russia иRussianFederation; China и Peoples Republic China; USA и US и т.п.) и, что существенно сложнее, наименований организаций (Moscow State University и Lomonosov Moscow State University, а также MSU и т.п.).

Другой известной проблемой является сложность интерпретации содержания научных работ на основе ключевых слов автора, на использовании которых, прежде всего, основан содержательный анализ исследований и разработок в рамках библиометрии. Это связано, в частности, с тем, что во многих случаях авторы используют слишком общие ключевые слова. Возможным решением является семантический анализ аннотаций и, при наличии доступа, полных текстов научных публикаций. Необходимость использования полнотекстового анализа научно-технической информации отмечается в настоящий момент многими ведущими исследователями в данной области [Uphametal., 2010; Wangetal., 2010; Li, etal., 2011; Borneretal., 2012; Boyacketal., 2013; Efimenkoetal., 2014; Кулинич, 2011;Хорошевский и др., 2015], а развитие семантических технологий для нужд библиометрии и наукометрии, по-видимому, уже можно считать важным зарождающимся трендом в мире.

С учетом вышесказанного, целью настоящей работы является разработка методологии и интеллектуального инструментария для выявления центров компетенций и превосходства с применением полнотекстового анализа, а также проверка предлагаемых решений на статистически значимом объеме данных из библиометрической БД Web of Science.

1.2. Методология исследования

В основурешения задачи выявления центров компетенций и превосходства были положеныследующие гипотезы:

- для позиционирования специалистов и/или организаций в научно-технических сообществах важнейшим информационным каналом являются публикации полученных результатов;
- тематическое картирование исследований и разработок по публикациям может быть базисом для выявления прорывных направлений;
- цитирование опубликованных работ может использоваться в качестве одного из критериев оценки значимости результатов;
- соавторство в публикуемых работах может быть свидетельством признания компетенций членов авторских коллективов;
- аффилиации авторов с организациями могут использоваться для позиционирования последних в качестве центров компетенций и превосходства;
- геоландшафты авторских коллективов дают представление не только о кооперации в науке и технике, но и о научно-технической политике государств;
- факты финансирования исследований и разработок разными фондами (государственными и/или коммерческими) могут использоваться для выявления центров компетенций и превосходства.

Следует сразу отметить, что перечисленными выше гипотезами не исчерпывается перечень индикаторов, которые могут быть положены в основу разработки полномасштабных моделей центров компетенций и превосходства. Вместе с тем, по нашему мнению, для проверки именно этих гипотезв библиографических базах данных имеется достаточно информации. Поэтому в настоящем исследовании именно они иположены в основу методологии выявления центров компетенций и превосходства.

1.3. Платформа наукометрии Semantic Hub

1.3.1. *Архитектура платформы*

Для интеллектуального наукометрического анализа корпусов текстов была разработана платформа Semantic Hub, общая архитектура которой представлена на Рис. 1.

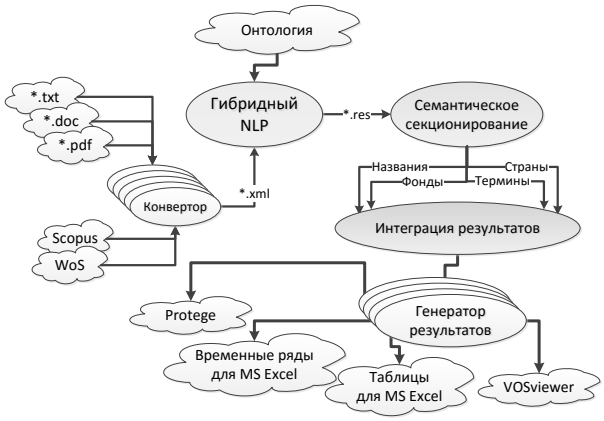

Рисунок1 – Общая архитектура платформы Semantic Hub

Компоненты основных стадий этой платформы кратко обсуждаются ниже.

1.3.2. *Стадия препроцессии*

Для эффективной обработки корпусов текстов в рамках платформы Semantic Hub используется

несколько конверторов, реализованных на языке Java. Результатом работы этих конверторов является XML-представление исходных данных в формате, описанном в работе [Хорошевский и др., 2015]. Для целей настоящего исследования использовался конвертор WoS->XML.

1.3.3. *Стадия извлечения информации из текстов*

Ядром платформы Semantic Hub является гибридный лингвистический процессор [Efimenkoetal., 2016], реализованный авторами в среде GATE, которая была разработана в Шеффилдском университете Великобритании[1]. Спецификойэтогопроцессораявляетсявозможностьо бработкиразличныхсемантически значимых фрагментов XML-представления с использованием лингвистических и статистических методов, а также интеграция результатов обработки отдельных текстов в результат обработки коллекции документов и множества коллекций документов.

Ключевую роль в гибридном лингвистическом процессоре играют модули извлечения из текстов терминов (в т.ч., многословных), поскольку именно они определяют качество дальнейшего наукометрического анализа и саму возможность семантически значимой интерпретации результатов. При этом следует отметить, что методы на основе извлечения биграмм и/или триграмм плохо работают в данном случае, поскольку научно-технические термины часто характеризуются существенно более сложными языковыми конструкциями, чем те, которые выделяются на уровне биграмм и/или триграмм. Использование обычных NP-чанкеров также не может рассматриваться как решение проблемы, поскольку, как правило, они не используют знаний о предметной области и уже в силу этого выделяют значительное число нерелевантных именных групп (например, таких как «thispaper» или «asanexample», «эта статья», «проведенный эксперимент» и др.).

Алгоритмы, реализованные в рамках гибридного лингвистического процессора платформы Semantic Hub, базируются на синтаксическом и семантическом анализе ЕЯ-текста, используют 100+ правил-шаблонов и обеспечивают извлечение из текстов сложных именных групп, например, таких как «high-angleannulardark-fieldscanningtransmissionelectronmicroscopy (haadf-stem)», «methanol-tolerantoxygenreductionreactionofpolymerelectrolytefuelcell», а также именных групп с формулами и аббревиатурами. Дополнительно к лингвистическим правилам в данном процессоре задействованы статистические алгоритмы TF-IDF ранжирования выделенных терминов и отсечения статистически незначимых терминов по заданному порогу. Кроме того, в гибридном лингвистическом процессоре платформы Semantic Hub используются специальные методы формирования словарей стоп-слов и выражений на базе концепции «черного ящика» [Efimenkoetal., 2014].

Реализованные алгоритмы обеспечивают формирование значительно более «чистых» списков терминов, извлекаемых из научно-технических текстов. В рамках данной работы извлекались термины 3-х типов: многословные термины (с включением аббревиатур), собственно аббревиатурыи формулы.

1.3.4. *Стадия генерации результатов*

На выходе стадии извлечения информации из текстов формируются множества специальных аннотаций, которые поступают на вход модулей генерации выходных представлений. Для целей настоящего исследования использовались специализированные генераторы статистических данных, полученных в процессе обработки ЕЯ-текстов, таблицы временных рядов, матрицы коллокаций, а также специальные матрицы для визуализации результатов в инструментарии VOSviewer[Ecketal., 2010]. Все генераторы реализованы на языке Java.

Таким образом, в рамках платформы Semantic Hub реализован полный цикл получения результатов для дальнейшего наукометрического анализа с использованием внешних аналитических инструментов.

1.4. Инструменты визуализации результатов

Для интерпретации результатов, полученных автоматически, безусловно, должны привлекаться экспертные сообщества, При этом, как правило, используются достаточно сложные аналитические инструменты. Поэтому до начала обсуждения полученных результатов, по нашему мнению, целесообразно хотя бы кратко обсудить те методы и средства визуализации, которые лежат в основе использованных в настоящем исследовании аналитических инструментов.

1.4.1. *Гистограммы*

Для визуализации распределений терминов и стран в настоящем исследовании использовались гистограммы, которые строились с использованием соответствующих функционалов MSExcel. Возможности этих функционалов хорошо известны и потому в данной работе не обсуждаются.

1.4.2. *Временные ряды*

Построение временных рядов терминов и стран в рамках настоящего исследования осуществлялась на результатах генерации нормированных распределений терминов и/или стран по отдельным временным периодам. При этом для собственно визуализации использовался MSExcel, соответствующие функционалы которого также хорошо известны и потому в данной работе не обсуждаются.

[1] GATE Official site: https://gate.ac.uk/

1.4.3. *Ландшафты*

В настоящем исследовании для визуализации ландшафтов использовались соответствующие задаче функционалы MSExcel, а также средства системы VOSviewer[2].

Как отмечалось выше, функционалы MSExcel хорошо известны и в дополнительных пояснениях, по-видимому, не нуждаются. Иная ситуация с использованием системы VOS viewer, в которой реализованы достаточно сложные алгоритмы кластеризации и визуализации. Поэтому ниже остановимся кратко на функционалах этой системы.

Как известно, одним из основных инструментов визуализации результатов в системе VOSviewer являются библиометрические карты [Noyons, 2004; Bornеretal., 2012]. Для построения таких карт в качестве основных выступаютэтапы

- Построения матриц коллокаций концептов,
- вычислениякоэффициентов «силы связи» концептов,
- позиционированияконцептовнаосновеих «силы связи» впространственьшейразмерности, а также
- собственно визуализациивпространственьшейразмерностивв идетепловыхкарти/иликартраспределенияплотности концептов.

При этом элементами матриц коллокаций являются частоты совместной встречаемости концептов в разных документах. Так, например, для матрицы коллокацийгеоимен: если в авторских коллективах 13 статей одновременно присутствуют специалисты из России (R) и Германии (G), в 21 статье – одновременно специалисты из Франции (F) и Голландии (N), а специалистов из Бельгии (B) и Индии (I), одновременно присутствующих в авторских коллективах, нет N_{RG}=13;N_{FN}=21; N_{BI}=0.

Вычисление коэффициентов «силы связи» между концептамибазируется на нормализации данных из матрицы коллокаций в соответствии со следующей формулой:

$$c_{ij} = \frac{mc_{ij}}{c_{ii}c_{jj}} \quad \text{для } i \neq j. \qquad (1)$$

Где $\boldsymbol{c_{ij}}$ соответствует количеству документов, в которых концепты $\boldsymbol{i}$ и $\boldsymbol{j}$ встречаются вместе, $\boldsymbol{c_{ii}}$ соответствует количеству документов, в которых встречается концепт $\boldsymbol{i}$, а$\boldsymbol{m}$фиксирует общее число документов. Такой подход, по мнению авторов VOSviewer, дает лучшие результаты по сравнению, например, с косинусной мерой или мерой Джакарда. Для примера, при условиях, указанных выше для коэффициентов матрицы коллокаций геоимен и общего числа документов в корпусе $\boldsymbol{m}$=100; N_{RR}=100;N_{GG}=50; N_{FF}=30; N_{NN}=20; N_{BB}=10; N_{II}=3 значенияC_{RG}=2.6; C_{FN}=3.5; C_{BI}=0.

[2]http://www.vosviewer.com/relatedsoftware/

Позиционированиеконцептоввтепловыхкартахи/илив картах распределения плотности концептов осуществляетсянаосновеспециальногометодаучета« силы связи» между концептами [Ecketal., 2009; Ecketal., 2014].

СутьVOS-методазаключаетсявследующем. Необходимо расположить $\boldsymbol{n}$ концептов в 2-хмерномпространстветак, чтобы расстояние между любой парой концептов $\boldsymbol{i}$и$\boldsymbol{j}$отражало их «силу связи» $\boldsymbol{c_{ij}}$ максимально аккуратно. Приэтомконцептысбольшимизначениями «силысвязи» должнырасполагатьсяблизко, аконцептысмалымизначениями «силысвязи» – далекодруг от друга. ИдеяVOS-методасостоитвминимизациивзвешеннойсуммыквадратов эвклидовых расстояний между всеми парами концептов. Дляисключенияслучаев, вкоторыхвсеконцепты «стянутся» воднуточку, вводитсяограничениена сумму всех расстояний, которая должна быть равна заданной константе. Вматематическомсмыслепроизводитсяминимизация функции

$$E(x_1, \dots x_n) = \sum_{i<j} c_{ij} \|x_i - x_j\|^2 \qquad (2)$$

Где вектор $\boldsymbol{x_i}$ = ($\boldsymbol{x_{i1}}$, $\boldsymbol{x_{i2}}$) фиксирует расположение концепта $\boldsymbol{i}$ в 2-х мерном пространстве, $\| \cdot \|$ задает эвклидову норму, а минимизацияфункцииосуществляетсясучетомследующегоограничения

$$\frac{1}{n(n-1)} \sum_{i<j} \|x_i - x_j\| = 1. \qquad (3)$$

При этом численное решение задачи оптимизации функции с учетом ограниченийосуществляется за 2 шага. Сначалазадачаоптимизациисограничениямитрансформируетсявзадачу оптимизации без ограничений, которая решается с помощью мажоритарного алгоритма[Borgetal., 2005]. Дляисключенияэффектаполучениялокальногоминимумамажоритарныйалгоритм запускается 10 раз со случайными начальными данными.

Собственно визуализация результатов выполнения предыдущих этапов происходит в виде тепловых карт и/или карт распределения плотности концептов.

Визуализациятепловыхкартосуществляетсяспомощью Java-апплета, где место концепта фиксируется его наименованием-меткой. Приэтомважностьконцептаспецифицируетсяразмеромшрифтаметки, которыйсвязансчисломдокументовгдеконцептприсутствует. Распределение «интереса» между концептами специфицируется цветом кластера, в которых эти концепты находится.

Предполагается, что кластеризация концептов уже проведена на базе использования рассмотренных выше метрик. Пусть, для

определенности, в кластер 1 попали документы с авторами из России и Германии (всего в кластере 150 документов), в кластер 2 – из Франции и Голландии (всего в кластере 50 документов), а в кластер 3 – из Бельгии и Индии (всего в кластере 30 документов). Тогда область концепта ***i*** «окрашивается» в палитре RGB с помощью следующих формул:

$$R_i = \frac{p_i^1}{p_i^1 + p_i^2 + p_i^3} 180 + 75, \quad (4)$$

$$G_i = \frac{p_i^2}{p_i^1 + p_i^2 + p_i^3} 180 + 75, \quad (5)$$

$$B_i = \frac{p_i^3}{p_i^1 + p_i^2 + p_i^3} 180 + 75, \quad (6)$$

где p_i^1 задает отношение числа документов из кластера 1, в которых концепт ***i*** присутствует, к числу документов из кластера 1, p_i^2 – отношение числа документов из кластера 2, в которых концепт ***i*** присутствует, к числу документов из кластера 2, а p_i^3 – отношение числа документов из кластера 3, в которых концепт ***i*** присутствует к числу документов из кластера 3.

Из наших предположений и формул (4-6) нетрудно получить цвета концептов. Так, например, RGB_R=(255, 75, 75) и RGB_G=(255, 75, 75), что понятно, поскольку авторы из этих стран находятся в одном кластере. Но RGB_F=(75, 255, 75) и RGB_N=(75, 255, 75), поскольку авторы из этих стран находятся в другом кластере.

Известным недостатком тепловых карт является частое перекрытие наименований концептов, что затрудняет понимание общей структуры карты концептов. Поэтому в систему VOSviewer включена опция визуализации результатов в виде карт распределения плотности концептов, где представлены не все концепты, а только те из них, которые встречаются часто, причем для индикации концептов с одинаковой плотностью в разных областях карты используются одинаковые цвета. При этом плотность концепта зависит от числа его «соседей» и от их важности следующим образом: увеличение числа «соседей» концепта и уменьшение расстояния между ними приводит к увеличению плотности, причем увеличение числа документов, в которых присутствует концепт, также увеличивает его плотность. Для вычисления значений плотности концептов авторы системы VOSviewer используют функциональные возможности пакета MATHLAB. Для цветового ранжирования концептов значения их плотностей упорядочиваются на шкале «голубой-красный» от меньших значений (голубой цвет) к большим (красный цвет).

1.5. Формирование корпуса для проведения исследования

Для решения поставленных выше задач авторами был проведен библиометрический анализ 255 000+ библиографических описаний научных публикаций с российским авторством (соавторством), представленных в WebofScience за 2009-2015 гг., на базе доступных инструментов из этой БД, а также углубленный семантический анализ 7 000 аннотаций наиболее цитируемых публикаций из этого корпуса с использованием собственного инструментария семантического наукометрического анализа [Efimenkoetal., 2014; [Хорошевский и др., 2015] и специализированного инструментария визуализации результатов VOSviewer, разработанного в Лейденском университете. При этом российскими считались те авторы, которые явно указали свою аффилиацию с организациями, расположенными в России.

Для формирования корпуса в качестве первого шага был сформирован запрос к БД Web of Science, позволяющий отобрать все публикации, где среди авторов присутствуют российские соавторы (или единственный автор аффилирован с российской организацией), за период 2009-2015 гг. Число публикаций по годам в результатах запроса представлено в Таблице 1.

Таблица 1 –Публикации с российскими авторами по данным Web of Science, 2009-2015 гг. (по состоянию на 25 января 2016 г.)

№	Год	Число публикаций
1.	2009	34 370
2.	2010	33 415
3.	2011	34 640
4.	2012	34 689
5.	2013	37 268
6.	2014	41 032
7.	2015	40 379

Необходимыми элементами моделей центров компетенций и превосходства являются индикаторы о результатах, влияющих на развитие последующих исследований и разработок, а также признание научным сообществом, в т.ч. международным. Таким образом, было принято решение выполнять детализированный анализ для публикаций с самыми высокими показателями цитируемости. Поскольку значения числа публикаций по годам близки для всех лет из рассматриваемого периода, для каждого года было выбрано по 1 000 самых цитируемых публикаций. Показатели цитируемости по годам представлены в Таблице 2.

Таблица 2 – Статистика цитирований для Топ-1000 публикаций с российскими авторами по данным Web of Science, 2009-2015 гг. (по состоянию на 25 января 2016 г.)

№	Год	Число цитирований для наиболее цитируемой публикации (1 место в рейтинге)	Число цитирований для последней публикации из Топ-1000 (1 000 место в рейтинге)
1.	2009	1 887	36
2.	2010	3 979	32
3.	2011	1 650	28

4.	2012	4 368	23
5.	2013	508	17
6.	2014	2 579	10
7.	2015	161	3

Таким образом, был сформирован корпус текстов для семантического анализа объемом 7 000 аннотаций.

2. Результаты и обсуждение

Как отмечалось выше, в рамках настоящего исследования был проведен библиометрический анализ достаточно полного корпуса публикаций с российским соавторством с помощью инструментов Web of Science и семантический наукометрический анализ Топ-1000 публикаций из этого корпуса по каждому из 2009-2015 гг. с использованием платформы SemanticHub и внешних аналитических инструментов.

Ниже представлены полученные результаты и приводится их обсуждение.

2.1. Результаты библиометрии

Подготовка статистических данныхдля анализарассмотренного выше корпуса осуществлялась с использованием соответствующих инструментов WebofScience, а для их интерпретации использовались подходящие функционалы MSExcel.

На Рис. 2-4 выборочно представлены сведения о топ-10 странах-партнерах, полученные на основе данных о соавторстве на всем множестве публикаций за каждый год (анализ был выполнен для всех стран мира, по доступным сведениям, не менее 100 стран для каждого года). Результаты анализа показывают, что ключевые партнеры (топ-5) не меняются с течением времени, при этом в целом по рейтингу наблюдается рост значимости партнерства с отдельными странами. В частности, за рассматриваемый период изменилось место Китая в рейтинге стран-партнеров, причем речь идет не о случайных выбросах, а о тренде: 14 место в 2009 году, 9 место в 2010 и 2011 годах, 7 место в 2012- 2014 годах, 6 место в 2015 году.

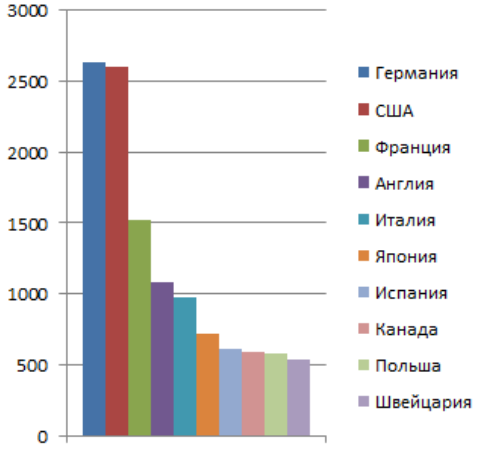

Рисунок2 – Топ-10 стран-партнеров РФ на основе сведений о соавторстве, по числу публикаций, 2009 год

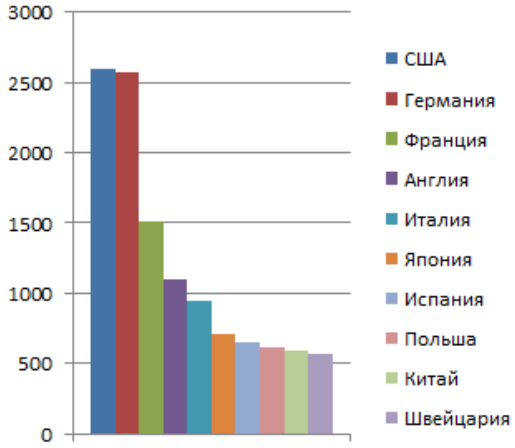

Рисунок3 – Топ-10 стран-партнеров РФ на основе сведений о соавторстве, по числу публикаций, 2010 год

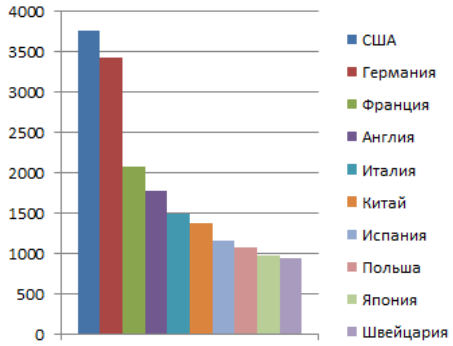

Рисунок4 – Топ-10 стран-партнеров РФ на основе сведений о соавторстве, по числу публикаций, 2015 год

Отдельно был выполнен анализ наиболее результативных коллабораций – стран-партнеров по наиболее цитируемым (топ-1000) публикациям. Пример результатов для 2009 года представлен на Рис. 5. Результаты анализа показывают, что наиболее результативные с точки зрения показателей цитируемости партнерства, в большинстве случаев, совпадают с партнерствами, наиболее продуктивными по числу публикаций.

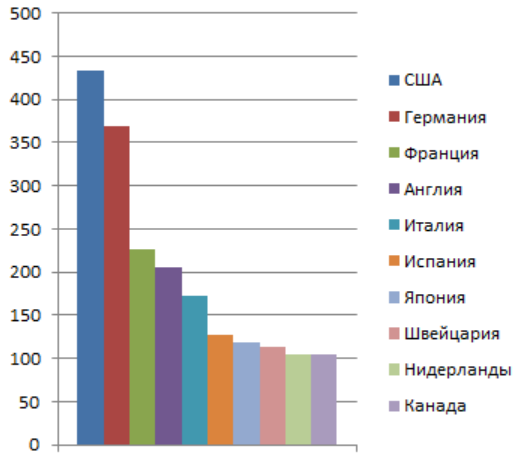

Рисунок5 – Топ-10 стран-партнеров РФ на основе сведений о соавторстве, по топ-1000 наиболее цитируемых публикаций, 2009 год

Дополнительно был выполнен анализ научных направлений, которым соответствуют публикации российских авторов (в т.ч. с зарубежными соавторами). Для этого были использованы сведения о категориях WebofScience (направления научных исследований по классификатору, используемому данной платформой), где представлены российские публикации, как на всем объеме публикаций, так и для наиболее цитируемых работ, для каждого года отдельно и за период 2009-2015 гг. в целом.

Пример результата для периода в целом (топ-10 категорий WebofScience по числу публикаций с российскими авторами) представлен на Рис. 6.

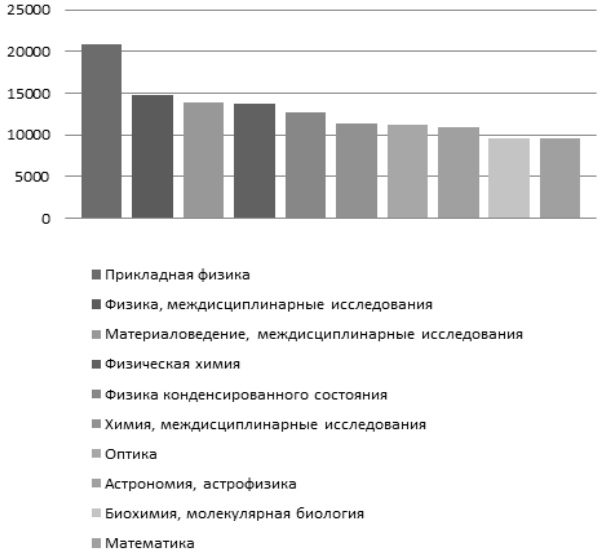

Рисунок6 – Топ-10 категорий Web of Science по числу публикаций с российскими авторами, 2009-2015 гг.

Результаты анализа показывают, что классические инструменты библиометрии дают общее представление о направлениях исследований и разработок, которые могут рассматриваться как научно-технологические области, где позиционируются центры компетенций и превосходства. Однако речь в данном случае идет только о верхнеуровневом анализе. Детальные сведения о ландшафте исследований и разработок не могут быть получены с помощью инструментов библиометрических баз данных без использования интеллектуального семантического анализа.

2.2. Семантический наукометрический анализ корпуса самых цитируемых публикаций

В результате работы модулей и компонент платформы SemanticHub из текстов научных публикаций и/или их аннотаций в рамках настоящего исследования извлекались значимые научно-технологические концепты в форме именных групп (однословных и, чаще, многословных терминов), позволяющие сделать вывод о содержании публикации и сути выполненной научной работы. Научно-технологические концепты извлекались также из названий публикаций.

Дополнительно к научно-технологическим концептам из аффилиаций авторов извлекались наименования государств, а из благодарностей – наименования фондов, финансирующихисследования и разработки.

2.2.1. *Гистограммы*

Как представляется, для дальнейшего анализа результатов наибольший интерес представляют распределения частот выделенных с помощью гибридного лингвистического процессора научно-технологических концептов, аффилиированных с авторами стран и фондов, финансирующих исследования и разработки. Общие гистограммы распределения этих индикаторов за 2009-20015 гг. представлены на Рис. 7-9.

Научно-технологическихтерминов выделено 101397. При этом на уровне отсечения 1 их 19173, а на уровне отсечения 10 – всего 1414. Ниже, на Рис. 7, представлено распределение частот встречаемости Топ-100 терминов.

Рисунок7 – Распределение Топ-100 терминов (2009-2015 гг.)

Всего выделено 166 стран, представленных в авторских коллективах. 98 стран представлены не менее чем в 10 случаях, 61 страна – не менее чем в 100 случаях и 40 стран – не менее чем в 400 случаях. На Рис. 8приведенораспределение Топ-40 стран из авторских коллективов.

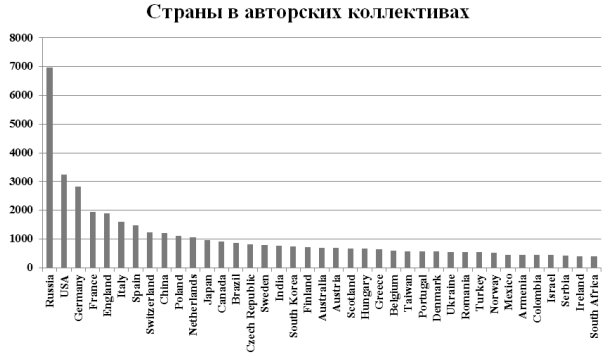

Рисунок 8 – Распределение Топ-40 стран (2009-2015 гг.)

Как показал анализ результатов, гипотеза о том, что для наукометрического анализа источников финансирования исследований и разработок можно использовать значения соответствующего тэга записей WebofScience, оказалась неверной. Такая ситуация, по нашему мнению, связана с тем, что в словарях фондов в БД WebofScience представлены

стандартизованные наименования, а авторы специфицируют одни и те же фонды в разделах благодарностях самыми разными способами. И более того, выходной формат тэга фондов в этой БД неоднозначен. Поэтому результаты обработки тэга фондов в данном исследовании были вручную отфильтрованы, а на Рис. 9 представлены результаты для Топ-100 финансирующих организаций.

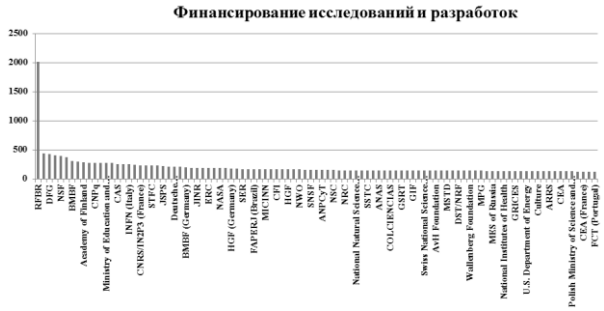

Рисунок9 – Распределение Топ-100 фондов (2009-2015 гг.)

Анализ представленных выше гистограмм частот показывает достаточно четкое соответствие закону Ципфа.

2.2.2. *Временные ряды*

В настоящей работе для анализа динамики исследований и разработок строились нормализованные временные ряды терминов и других индикаторов. Для примера на Рис. 10-11 представлены полученные результаты.

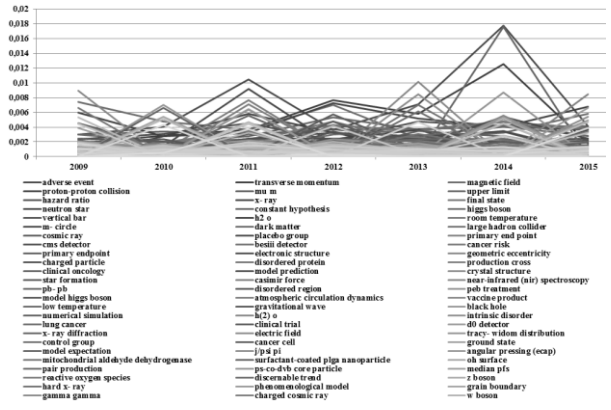

Рисунок10 – Временные ряды Топ-100 терминов (2009-2015 гг.)

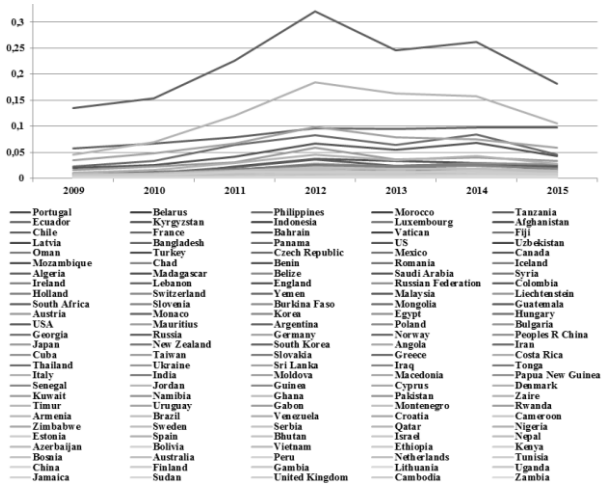

Рисунок11 – Временные ряды Топ-100 стран (2009-2015 гг.)

2.2.3. *Ландшафты*

Как известно, для аналитиков наибольший интерес представляют карты научных направлений. Поэтому в даннойработе строились тепловые карты и карты распределения плотности научно-технических терминов, странавторских коллективов и фондов, финансирующих исследования и разработки. Для примера, на Рис. 12 представленатепловая карта стран из авторских коллективов, а на Рис. 13 –спектр сетей кластеров научно-технических терминов. Детальное обсуждение полученных результатов представлено ниже.

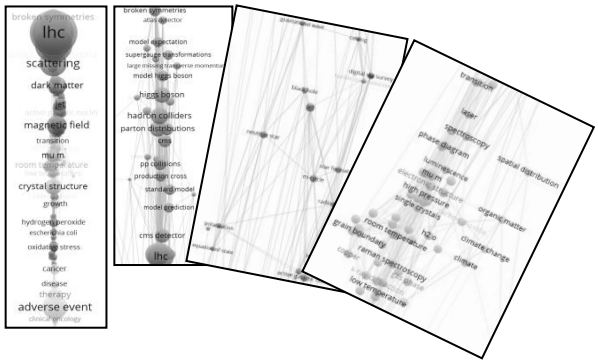

Рисунок12 – Тепловая карта стран из авторских коллективов по странам (2009-2015 гг.)

Рисунок13 – Сети научно-технических терминов(2009-2015 гг.)

Анализ результатов показывает, что Топ-5 стран (по данным WebofScience) это Россия-США-Германия-Франция-Англия. Следующую по рангу «весов» системы VOSviewerгруппу составляют Италия и Испания, а замыкают «чертову дюжину» самых активных сран Швейцария, Китай, Польша, Голландия, Япония и Канада.

Понятно, что Россия входит в Топ-5, поскольку в исходном корпусе изначально присутствовали только те статьи, в которых были (со)авторы из России. Что же касается остальных стран, их присутствие в лидирующих группах, по оценке экспертов, вполне предсказуемо.

Для выявления центров компетенции и превосходства интересе не просто анализ научно-технической активности по странам, но и выявление тех тематик, где сраны-лидеры концентрируют свои исследования и разработки. С учетом этого, на Рис. 14 фрагмент карты наиболее частотных терминов из статей Топ-5 стран, а на Рис. 15 –

тепловая карта Топ-10 научно-технических терминов из статей Топ-5 стран.

Последняя наиболее интересна для понимания того, чем занимаются страны-лидеры. При этом важно, что теми же вопросами занимается и Россия, причем ее уровень компетентности в этих вопросах высок, поскольку в противном случае российских исследователей не приглашали бы в авторские коллективы самых цитируемых статей, опубликованных в высоко рейтинговых научных изданиях.

Анализ данных, представленных на Рис. 15, показывает, что страны-лидеры концентрируются (по данным WebofScience), в основном, в области наиболее значимых областей физики.

По нашему мнению, результаты анализа демонстрируют не только научно-техническую важность новых направлений физики, но и давнюю ориентацию ThomsonReuterна индексирование работ в области ядерной физики, космоса и некоторых других направлениях для мониторинга получаемых в других странах научно-технических решений.

3. Направления дальнейших исследований и разработок

С учетом полученных результатов в качестве направлений дальнейших исследований и разработок целесообразными представляются авторам следующие:

- Расширение корпуса для обработки за счет других библиометрических баз (Scopus, Pub Med, Google Schoolar и т.п.), а также за счет тематических архивов научных сообществ.
- Специализация запросов к базам данных библиометриис целью формирования тематических корпусов для дальнейшей обработки.
- Подключение уже разработанных авторами модулей детализации аффилиаций авторов к гибридному лингвистическому процессору платформы SemanticHub.

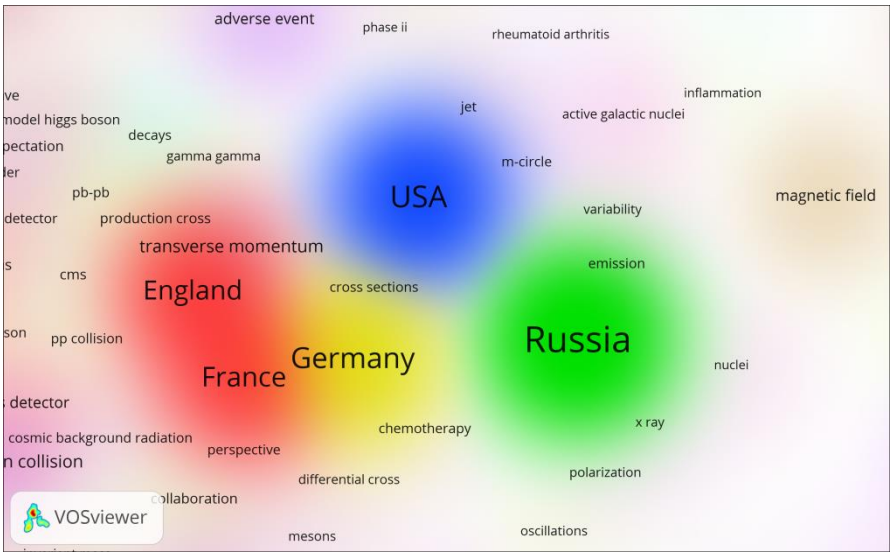

Рисунок 14 – Фрагмент карты наиболее частотных научно-технических терминов из статей Топ-5 стран (2009-2015 гг.)

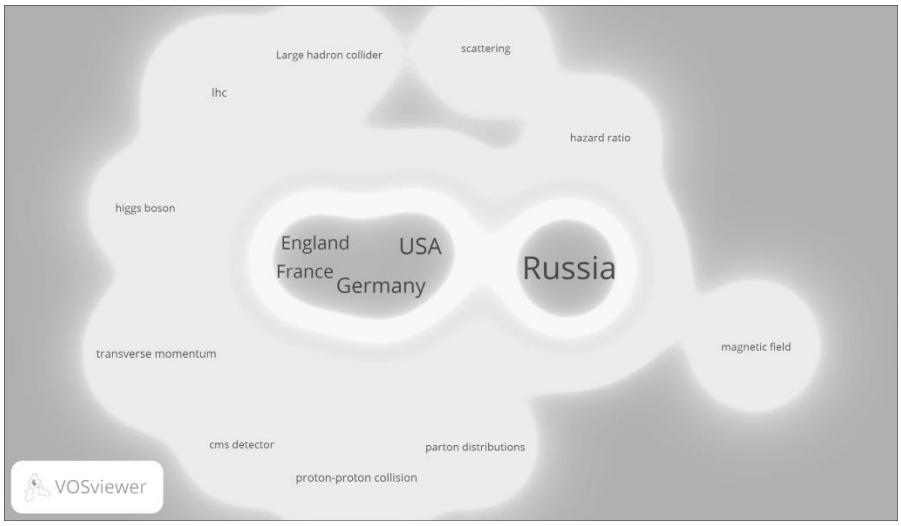

Рисунок 15 – Тепловая карта Топ-10 научно-технических терминов из статей Топ-5 стран (2009-2015 гг.)

- Доработка гибридного лингвистического процессора платформы SemanticHub в части наименований финансирующих исследования и разработки фондов.
- Использование в рамках анализа не только наиболее частотных терминов, но терминов, которые могут представлять «слабые сигналы».

Заключение

В работе представлен наукометрический анализ ландшафта исследований и разработок, выполняемых российскими специалистами в соавторстве со специалистами из других стран, с использованием семантических технологий сцелью выявления российских и мировых центров компетенций и превосходства в различных предметных областях на основе специализированных инструментов, разработанных и реализованных авторами.

Как показал проведенный анализ, платформа интеллектуальной наукометрии Semantic Hub может быть использована для решения поставленных задач и дает семантически интерпретируемые результаты.

Работа выполнена при частичной поддержке гранта РФФИ № 15-01-06819 «Исследование и разработка онтологических моделей центров компетенции/превосходства в прорывных научно-технологических направлениях на основе мониторинга разнородных информационных ресурсов», а также Фонда «Центр стратегических разработок» (в рамках разработки Стратегии научно-технологического развития РФ на долгосрочный период).

Библиографическийсписок

[Акоев и др., 2014]Акоев М.А. и др. Руководство по наукометрии: индикаторы развития науки и технологии. – Екатеринбург, Изд-во Урал. ун-та, 2014. -250с.

[Кулинич, 2011]Кулинич А.А. Компьютерные системы анализа ситуаций и поддержки принятия решений на основе когнитивных карт: подходы и методы./ Проблемы управления, 2011, № 4 С.31-45.

[Хорошевский и др., 2015] Хорошевский В.Ф., Ефименко И.В. Семантическая технология картирования семантических технологий (Наукометрический анализ конференций OSTIS). // Труды Y международной научно-технической конференции «Открытые семантические технологии проектирования интеллектуальных систем». Минск 19-21 февраля 2015 г., - Минск, БГУИР, 2015, с. 43-57.

[Borg et al., 2005] Borg I., Groenen P. J. F. Modern Multidimensional Scaling. Springer, second edition, 2005.

[Borner et al., 2012]Borner, K., Boyack, K. W., Milojevic, S., & Morris, S. (2012). An introduction to modeling science: Basic model types, key definitions, and a general framework for the comparison of process models. / Modeling Science Dynamics (Understanding Complex Systems), Springer-Verlag, p 3-22.

[Boyack et al., 2013]Boyack, K. W., Small, H., &Klavans, R.. (2013). Improving the accuracy of co-citation clustering using full text. Journal of the American Society for Information Science and Technology, 64(9), 1759-1767.

[Cobo et al., 2011] Cobo M.J., López-Herrera A.G., Herrera-Viedma E., and Herrera F. Science Mapping Software Tools: Review, Analysis, and Cooperative Study Among Tools. Journal of the American Society for Information Science and Technology, 62(7):1382–1402, 2011

[Eck et al., 2009] Van Eck, N.J., &Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. Journal of the American Society for Information Science and Technology, 60(8), 1635–1651.

[Eck et al., 2010] Van Eck N.J., Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523–538.

[Eck et al., 2014] Van Eck N.J., Waltman L. Visualizing bibliometric networks. /In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), Measuring scholarly impact: Methods and practice (pp. 285–320). Springer.

[Efimenko et al., 2014] Efimenko I., Khoroshevsky V. New Technology Trends Watch: an Approach and Case Study. // Proc. of AIMSA-2014. Springer-Verlag, 2014. P. 170-177.

[Efimenko et al., 2015] Efimenko I., Khoroshevsky V. Peaks, Slopes, Canyons, Plateaus: Identifying Technology Trends throughout the Life Cycle. // International Journal of Innovation and Technology Management (Special Issue “Bibliometrics and Social Network Analysis Methods for Technology and Innovation Management”, 2015) (In press)

[Efimenko et al., 2016] Efimenko I., Khoroshevsky V., Noyons E. Anticipating Future Pathways of Science, Technologies & Innovations: (Map of Science)^2 Approach. Chapter in Book: Anticipating Future Innovation Pathways through Large Data Analytics (eds.: T. Daim, A. Porter, D. Chiavetta, O. Saritas), Springer Verlag, 2016 (In press)

[Li, et al., 2011]Li H., Xu F., Uszkoreit H.: TechWatchTool: Innovation and Trend Monitoring. In: Proc. of the International Conference on Recent Advances in Natural Language Processing 2011 RANLP 2011, Tissar, Bulgaria, pp. 660-665 (2011).

[Noyons, 2004] Noyons E. C. M. Science maps within a science policy context. In H. F. Moed, W. Gl¨anzel, and U. Schmoch, editors, Handbook of Quantitative Science and Technology Research, pages 237–255. Kluwer Academic Publishers, 2004.

[Upham et al., 2010]Upham, S. P., & Small, H. (2010). Emerging research fronts in science and technology: Patterns of new knowledge development. Scientometrics, 83(1), 15-38.

[Wang et al., 2010] Wang et al. Identifying technology trends for RD planning using TRIZ and text mining, RD Management, vol. 40, N 5, 2010.

RUSSIA AMONG THE WORLD CENTERS OF EXCELLENCE

Khoroshevsky V.F.[*], Efimenko I.V. [**]

**Dorodnicyn Computing Centre, Federal Research Center «Computer Science and Control» RAS, Moscow, Russia*

khor@ccas.ru

***Faculty of Humanities, NRU HSE, Moscow, Russia*
Semantic Hub, Moscow, Russia

iefimenko@hse.ru

The paper presents a scientometric analysis of the landscape of R&D carried out by Russian scientists in cooperation with foreign co-authors. A novel approach is based on semantic technologies in Scientometrics. It makes use of tools for intelligent big data analysis which were developed by the authors of the paper within the SemanticHub platform. The research is aimed at identification of Russian and world-level centers of excellence in various R&D domains.

Keywords: Scientometrics; Centers of excellence; Semantic Technologies; Information Extraction