OSTIS-2015
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

СЕМАНТИЧЕСКАЯ ТЕХНОЛОГИЯ КАРТИРОВАНИЯ СЕМАНТИЧЕСКИХ ТЕХНОЛОГИЙ (Наукометрический анализ конференций OSTIS)

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

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

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

Обсуждаются вопросы картирования научных направлений на основе методов наукометрии с использованием семантических технологий. Исходными данными для проведения исследования является корпус статей, опубликованных в трудах конференций серии OSTIS «Open Semantic Technologies for Intelligent Systems». Дается краткое описание основных понятий наукометрии и обосновывается важность применения семантических технологий в решении возникающих здесь задач. Предлагаются модели, методы и средства анализа научных направлений и выявления центров компетенций и центров превосходства на базе семантизации методов наукометрии. Приводятся результаты семантического картирования предметной области семантических технологий, специфицированной в трудах конференций серии OSTIS.
Ключевые слова: наукометрия; онтологическое моделирование; карта науки; скрытые коллективы; центры превосходства и компетенций; семантические технологии; извлечение информации из текстов

Введение

Исследования и разработки по картированию научных направлений на основе методов и средств наукометрии активно развиваются во всем мире [Small, 2010; Borner et al., 2012; Boyack et al., 2014]. Для оценки ситуации в той или иной области, как правило, используются статистические методы библиометрического анализа публикаций [Boyack et al., 2005; Klavans et al., 2006; Shibata, et al., 2008], которые далеко не всегда адекватны задачам построения действительно информативных карт науки. В связи с этим в последнее время в проектах по наукометрии все большее внимание уделяется поиску новых методов картирования научных направлений и способов выявления центров компетенции и превосходства в различных областях науки и техники [Boyack, 2009; Upham et al., 2010; Klavans et al., 2010; Erdi et al., 2013].

При этом «горячими точками» исследований и разработок является применение методов интеллектуального анализа данных [Паклин и др., 2009; Witten et al., 2011], компьютерной лингвистики и, в частности, извлечения информации из текстов [Li, et al., 2011; Boyack et al., 2013; Efimenko et al., 2014], а также специальных средств визуализации результатов библиометрии [Borner et al., 2003; Klavans et al., 2011; Van Eck et al., 2014**]**.

С учетом вышесказанного, целью настоящей работы является обсуждение вопросов построения семантических карт научных направлений на базе совместного использования методов наукометрии и семантических технологий, а в качестве исходных данных для проведения исследования выступает корпус статей, опубликованных в трудах конференций серии OSTIS.

1. Основные понятия наукометрии

1.1. Предварительные замечания

В рамках общего направления «Информатика» одной из востребованных в настоящее время научных дисциплин, связанной с количественными измерениями хранимой и используемой информации является **инфометрия**[1], как правило, включающая:

- **библиометрию (Bibliometrics)**, которая занимается изучением документов на основе

[1] http://en.m.wikipedia.org/wiki/Informetrics

количественного анализа первичных и вторичных источников информации с помощью формализованных методов с целью получения данных об эффективности, динамике, структуре и закономерностях развития исследуемых областей;

- **наукометрию (Scientometrics)** – научную дисциплину, связанную с изучением количественных методов развития науки как информационного процесса;
- **вебометрику (Webometrics)** – раздел информатики, где исследуются количественные аспекты конструирования и использования информационных ресурсов, структур и технологий применительно к Интернет;
- **альтметрику (Altmetrics)** – раздел инфометрии, где исследуются различные аспекты использования открытых информационных ресурсов, включая социальные сети и другие источники информации.

В настоящей работе авторы, в основном, концентрируются на вопросах библиометрии и наукометрии. Поэтому в оставшейся части данного раздела обсуждаются понятия и проблемы именно этих направлений инфометрии.

1.2. Основные понятия и индикаторы

Как известно, сам термин «наукометрия» был впервые введен в монографии [Налимов и др., 1969], где данное понятие связывалось с изучением эволюции науки через многочисленные измерения и статистическую обработку информации. Наукометрию часто применяют как абсолютную основу оценки выполнения и финансирования различных организационных единиц (институтов, команд, индивидуумов, а также проектов и разработок).

В западной науке данное направление (Scientometrics) трактуют шире, включая в него изучение, измерения и аналитику в области науки, технологий и инноваций для использования в контексте научной политики и управления.

Основным понятием библиометрии и наукометрии является **индекс цитирования** научных статей (**ИЦ**) — реферативная база данных научных публикаций, индексирующая ссылки, указанные в пристатейных списках этих публикаций и предоставляющая количественные показатели этих ссылок. Заметим, что в России распространена особая интерпретация данного понятия, как показателя, указывающего на значимость статьи и вычисляющегося на основе последующих публикаций, ссылающихся на данную работу.

История создания индексов (или указателей) научного цитирования начинается с 70-х годов XIX века, когда практически одновременно появляются индекс юридических документов Shepard's Citations (1873 г.) и индекс научных публикаций по медицине Index Medicus (1879 г.). Последний просуществовал вплоть до 2004 г.

В 1960 году Институт научной информации (ISI), основанный Юджином Гарфилдом, ввёл индекс цитирования для статей, опубликованных в научных журналах, положив начало «Science Citation Index (SCI)», а затем включив в него индексы цитирования по общественным наукам («Social Sciences Citation Index», SSCI) и искусствам («Arts and Humanities Citation Index», AHCI). Начиная с 2006 г. появились другие источники подобных данных, например, Google Scholar. Наиболее известными ИЦ в настоящее время являются библиометрические базы Web of Science и SCOPUS.

В 1974 году в ВИНИТИ были предприняты попытки создания отечественного указателя научного цитирования (УНЦ), который в технологическом плане должен был стать «аналогом» SCI [Егоров и др., 2006]. Работы по библиометрии и наукометрии активно велись и в других организациях России [Налимов и др., 1971; Евстигнеев, 1987; Маршакова, 1988; Хайтун, 1989], но с приходом на отечественный рынок БД и сервисов Web of Science, SCOPUS, PubMed и некоторых других отечественные исследования и разработки в данной области «сходят на нет». И только в последнее время здесь намечается некоторое «оживление» ситуации [Кулинич, 2011; Крюков и др., 2013; Хорошевский, 2012a; Хорошевский и др., 2014; Efimenko et al., 2014].

В 1987 Китай запустил проект по созданию Китайского индекса научного цитирования Chinese Science Citation Index, а в следующем, 1988 появился его конкурент — China Scientific and Technical Papers and Citations. В 1997 начата разработка китайского индекса цитирования по общественным наукам Chinese Social Sciences Citation Index.

В 1995 году Япония приступила к созданию национального индекса цитирования Citation Database for Japanese Papers, разработчиком которого является Национальный институт информатики Японии.

Разработки национальных индексов ведутся на Тайване (Taiwan Humanities Citation Index), а также в ряде европейских стран (Польша, Испания).

Базовые индикаторы и показатели библиометрии и наукометрии[2], как правило, включают:

- **g-индекс** (2006 г., Leo Eggh) – индекс для измерения научной продуктивности, который рассчитывается на основе распределения цитирований, полученных публикациями ученого следующим образом: для данного множества статей, отсортированного в порядке убывания количества цитирований, которые получили эти статьи, g-индекс это наибольшее число, такое что g самых цитируемых статей получили (суммарно) не менее g^2 цитирований.
- **h-индекс** (индекс Хирша) – показатель, предложенный в 2005 г. аргентино-американским физиком Хорхе Хиршем из Калифорнийского

[2] https://ru.wikipedia.org/wiki/Категория:Наукометрия

университета в Сан-Диего, который является количественной характеристикой продуктивности ученого, группы ученых, научной организации или страны в целом, основанной на количестве публикаций и количестве цитирований этих публикаций. Согласно Хиршу, ученый имеет индекс h, если h из его Np статей цитируются как минимум h раз каждая, в то время как оставшиеся (Np — h) статей цитируются не более, чем h раз каждая.

- **i-индекс** (Предложен в 2006 г. независимо М. Космульским и Г. Пратхапом) – индекс публикационной активности научной организации, рассчитываемый на основе библиометрических показателей с использованием распределения индекса Хирша ученых из данной научной организации. Научная организация имеет индекс i, если не менее i ученых из этой организации имеют h-индекс не менее i.
- **Исследовательские фронты**, основанные на использовании методов библиографических связей (ретроспективный метод) и/или коцитирования (проспективный метод).
- **Импакт-фактор** (**ИФ** или **IF**) - численный показатель важности научного журнала. С 60-х годов прошлого века он ежегодно рассчитывается Институтом научной информации (ISI), который в 1992 г. был приобретен корпорацией Thomson, ныне называется Thomson Scientific и публикуется в журнале «Journal Citation Report». Расчёт импакт-фактора основан на 3-х летнем периоде: так, например, импакт-фактор журнала в 2011 году – $IF_{2011} = A/B$, где: A - число цитирований в течение 2011 г. в журналах, отслеживаемых ISI, статей, опубликованных в данном журнале в 2009—2010 гг.; B — число статей, опубликованных в данном журнале в 2009-2010 гг.

В соответствии с ИФ (в основном в других странах, но в последнее время и в России) оценивают уровень журналов, качество статей, опубликованных в них, дают финансовую поддержку исследователям и даже принимают сотрудников на работу.

К основным библиометрическим базам и сервисам относятся:

- Web of Knowledge (Web of Science)[3] – поисковая платформа, объединяющая реферативные базы данных публикаций в научных журналах и патентов, в том числе базы, учитывающие взаимное цитирование публикаций, разрабатываемая и предоставляемая компанией Thomson Reuters. Web of Knowledge охватывает материалы по естественным, техническим, биологическим, общественным, гуманитарным наукам и искусству. Платформа обладает встроенными возможностями поиска, анализа и управления библиографической информацией.
- SCOPUS[4] – библиографическая и реферативная база данных и инструмент для отслеживания цитируемости статей, опубликованных в научных изданиях. Индексирует 18 тыс. названий научных изданий по техническим, медицинским и гуманитарным наукам 5 тыс. издателей, включая научные журналы, материалы конференций и серийные книжные издания. Разработчиком и владельцем Scopus является издательская корпорация Elsevier. В отличие от Web of Knowledge, в Scopus не используется понятие импакт-факторов, но очень широко применяется индекс Хирша.
- Академия Google (Google Scholar)[5] – свободно доступная поисковая система, обеспечивающая полнотекстовый поиск научных публикаций всех форматов и дисциплин. Система работает с ноября 2004 года, первоначально в статусе бета-версии. Индекс Google Scholar включает в себя большинство рецензируемых онлайн журналов Европы и Америки крупнейших научных издательств. По функциям он похож на свободно доступные системы Scirus от Elsevier, CiteSeerX и getCITED, а также на платные сервисы Scopus и Web of Science.
- РИНЦ[6] – библиографическая база данных научных публикаций российских ученых. Для получения необходимых пользователю данных о публикациях и цитируемости статей на основе базы данных РИНЦ разработан аналитический инструментарий ScienceIndex. Проект РИНЦ разрабатывается с 2005 года компанией «Научная электронная библиотека» (ELIBRARY.ru)

Базис библиометрических баз и сервисов составляет инструментарий наукометрии. При этом типология таких инструментов включает

- средства загрузки данных;
- средства аналитики на данных и
- средства визуализации результатов.

Известными примерами инструментов наукометрии являются, в частности:

- Аналитические сервисы Web of Science и SCOPUS.
- Системы CitNet Explorer и VOS Viewer из Лейденского университета.
- Коммерческая платформа VantagePoint.
- Аналитические инструменты компании SciTech Strategies и др.

Мощными инструментами визуализации результатов наукометрии являются сети цитирования и коцитирования, диаграммы взаимовлияния научных направлений (Win-Win партнерства), сети динамики развития научных направлений, а также геоландшафты, обеспечивающие поиск партнеров, выявление лидерства стран, распространения научных идей, научных школ и виртуальных международных коллективов.

В качестве примеров, на Рис. 1-4 приводятся скриншоты, иллюстрирующие перечисленные выше

[3] http://thomsonreuters.com/about-us/
[4] http://www.scopus.com/
[5] http://scholar.google.ru/
[6] http://elibrary.ru/

сервисы наукометрии.

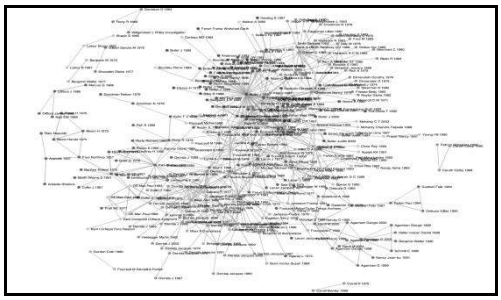

Рисунок 1 – Фрагмент сети цитирования и коцитирования

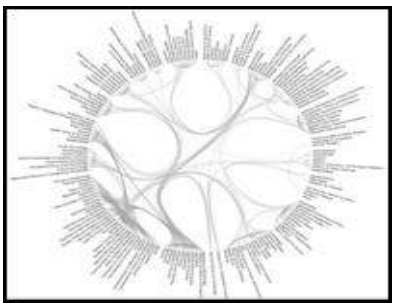

Рисунок 2 – Пример Win-Win партнерства

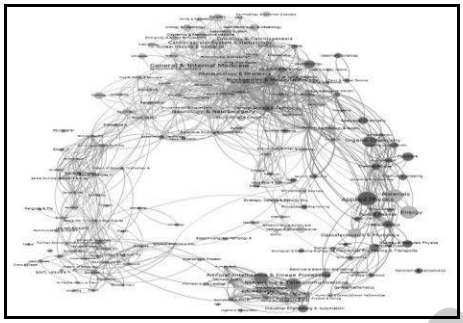

Рисунок 3 – Пример динамики развития научных направлений

Рисунок 4 – Пример геоландшафта научных направлений

1.3. Наукометрические БД и сервисы

В настоящее время работы по теории и практике наукометрии активно развиваются во всем мире, а особое внимание уделяется инструментальным средствам. Не имея возможности в настоящей работе дать даже краткий обзор исследовательской активности в данной области, мы представим ниже лишь несколько интересных систем и сервисов наукометрии.

1.3.1. *Модели и инструменты SciTech Strategies*

Модели, методы и инструменты картирования науки и технологий от компании SciTech Strategies[7] из США известны во всем мире. И даже само понятие карт науки было введено основателями этой компании К. Бояком (K.Boyack) и Р. Клэвансом (R. Klavans).

Для примера, на Рис. 5 приведена экранная форма библиотеки карт науки от компании SciTech Strategies.

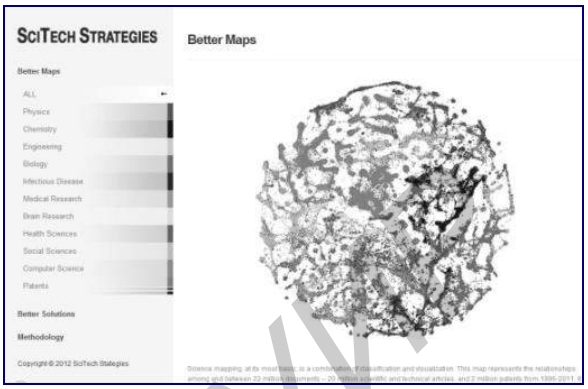

Рисунок 5 – Библиотека карт науки SciTech Strategies

В настоящее время К. Бояк, Р. Клэванс и их сотрудники активно работают над моделями наукометрии нового поколения [Boyack et al., 2014; Klavans et al., 2014a; Klavans et al., 2014b].

1.3.2. *Наукометрические инструменты Лейденского университета*

Центр исследований науки и технологий Лейденского университета в Нидерландах является одной из известных европейских точек превосходства и компетенций в данной области, а наукометрические инструменты CitNet Explorer и VOS Viewer [Van Eck et al., 2014], разработанные здесь, активно используются в разных странах.

Основные функционалы базового инструмента CitNet Explorer – следующие:

- В части данных поддерживаются
  - импорт из БД Web of Science,
  - экспорт в файлы Pajek-формата,
  - работа с «большими данными» (сети 10↑6 публикаций и 10↑7 ссылок).
- При визуализации поддерживается
  - Масштабирование и скроллинг сетей в стиле Google Maps,
  - отображение сетей косвенного цитирования,
  - экспорт скриншотов с результатами в Word или PowerPoint.
- Аналитика с помощью библиотеки алгоритмов для идентификации связанных компонент, кластеров, ядерных публикаций, кратчайших (максимальных) путей цитирования поддерживает:
  - фильтрование публикаций по времени,
  - функции drill down и expand для навигации по сетям цитирования.

Примеры экранных форм CitNet Explorer и VOS Viewer представлены на Рис. 6-7.

[7] http://www.mapofscience.com/

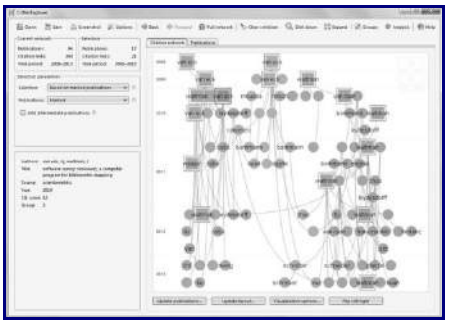

Рисунок 6 –Пример экранной формы CitNet Explorer

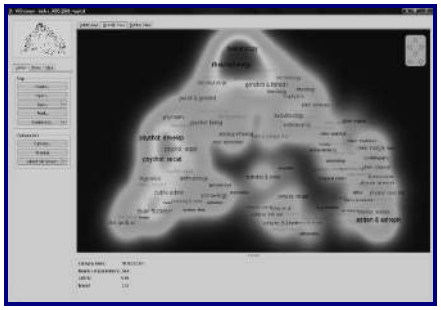

Рисунок 7 – Пример экранной формы VOS Viewver

Достоинством инструментов CitNet Explorer и VOS Viewer является и то, что, в отличие от многих других, они поставляются как программное обеспечение с открытым кодом, а недостатком – невозможность работы с русскоязычными текстами.

1.3.3. *Коммерческая система VantagePoint*

Флагманским продуктом компании VantagePoint[8] является одноименная библиометрическая платформа, которая (по утверждению разработчиков) является мощным инструментом извлечения знаний из результатов поиска патентов и БД литературы, а визуальные VantagePoint-перспективы, скриншоты которых представлены на Рис. 8, позволяют быстро найти ответы на вопросы WHO, WHAT, WHEN и WHERE, обеспечивая поиск критических паттернов.

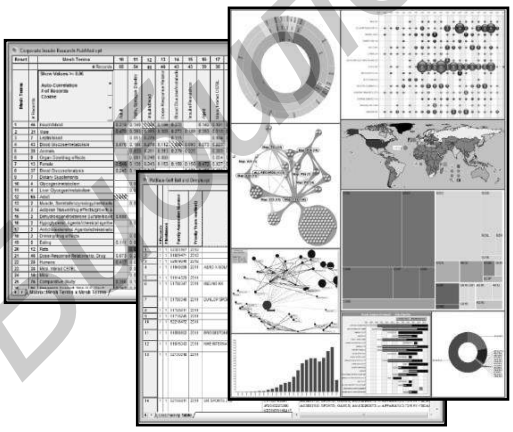

Рисунок 8 – Примеры VantagePoint-перспектив

1.3.4. *Аналитический инструментарий РИНЦ ScienceIndex*

Как отмечалось выше, российский индекс научного цитирования (РИНЦ) в настоящее время формируется на платформе e-Library, где основным аналитическим инструментом является ScienceIndex [Юрков, 2015], основные сервисы которого представлены на Рис. 9.

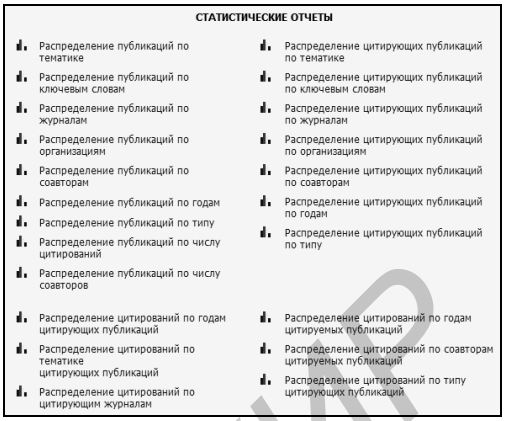

Рисунок 9 – Основные сервисы ScienceIndex РИНЦ

К сожалению, в настоящее время в БД РИНЦ проиндексировано мало (по сравнению с мировыми библиометрическими БД) документов, чир существенно затрудняет проведение аналитических исследований.

1.3.5. *Заключительные замечания*

Представленный выше краткий обзор основных понятий, методов и инструментов наукометрии позволяет сформулировать несколько положений.

Во-первых, можно констатировать, что важность наукометрии как одного из инструментов для оценки состояния исследований и разработок в научно-технической сфере и поддержки принятия решений в данной области уже практически осознана во всем мире. При этом для России основными задачами наукометрии являются:

- Выявление новых направлений научно-технического прогресса.
- Выявление центров компетенции и превосходства в прорывных направлениях научно-технического прогресса.
- Отображение научно-технологических ресурсов страны на мировые тренды научно-технического прогресса.
- Отображение научно-технологических ресурсов страны и мировых трендов на цели страны и доступные средства.
- Формирование целевых программ научно-технического прогресса.
- Планирование процессов поддержки целевых программ научно-технического прогресса.
- Оценка результатов выполнения целевых программ и формирование новых целей.

Во-вторых, можно отметить, что в рамках наукометрии активно развиваются следующие научно-технологические направления:

- Библиометрия и наукометрия научно-технической сферы, включая
  - индикативные методы библиометрии;
  - анализ исследовательских фронтов;

[8] https://www.thevantagepoint.com/

- выявление новых научно-технологических трендов.
- Форсайт в научно-технической сфере, в том числе
  - экспертные панели;
  - дорожное картирование.
- Системный анализ и исследование операций, включая
  - модели и методы прогнозирования;
  - методы экспертных оценок и согласования мнений;
  - методы планирования и оптимизации.

Основными проблемами в данной области, по нашему мнению, являются:

- Неполнота и «зашумленность» исходных данных.
- Преимущественная ориентация на статистические методы анализа процессов наукометрии.
- Отсутствие новых моделей и методов анализа процессов наукометрии.
- Недостаточно активное использование в научно-технической сфере уже существующих математических моделей и методов поддержки принятия решений.
- Слабое использование математических методов управления научно-технической сферой.

Представляется, что в свете указанных целей, задач и проблем семантические технологии могут стать драйвером создания новых методов мониторинга исследований и разработок в научно-технической сфере и драйвером создания новых методов поддержки принятия решений.

В заключение хотелось бы отметить, что научно-технологическое направление наукометрии нового поколения находится в настоящее время лишь в начале пути и требует серьезных междисциплинарных исследований и разработок.

2. Модели, методы и средства анализа научных направлений и выявления центров компетенций и превосходства

2.1. Предварительные замечания

Как отмечалось в работах [Wang, et al., 2010; Li, et al., 2011; Erdi, et al., 2013; Хорошевский и др., 2014], для построения карт науки целесообразно использовать гибридный подход, где методы кластеризации и классификации работают на данных, сформированных с помощью методов извлечения информации из текстов под управлением онтологий.

В основе предлагаемого в настоящей работе подхода к выделению из публикаций семантически значимой системы терминов предметной области, которые специфицируют характеристические вектора публикаций, лежит принцип «черного ящика» [Efimenko, et al., 2014], а интеграция результатов обработки коллекций отдельных жанров осуществляется на основе результатов статистической обработки коллекций публикаций. Спецификой предлагаемого подхода является и то, что после лингвистической и статистической обработки текстов происходит автоматическая генерация семантических представлений результатов, которые затем отражаются в экземплярную часть OWL-онтологии предметной области, в XML-файлы для системы кластеризации Carrot, а также в спецификации графического представления карты научного направления и выявления временных рядов терминов.

2.2. Онтологическая модель направления

Для автоматизации процессов построения карты научного направления OSTIS была использована система онтологических моделей, описанных в работе [Хорошевский и др., 2014]. Для примера, на Рис. 10 представлен фрагмент используемой в настоящем исследовании онтологической модели. Заметим, что эта система моделей используется как на этапе извлечения информации из текстов, так и в процессе генерации результатов обработки.

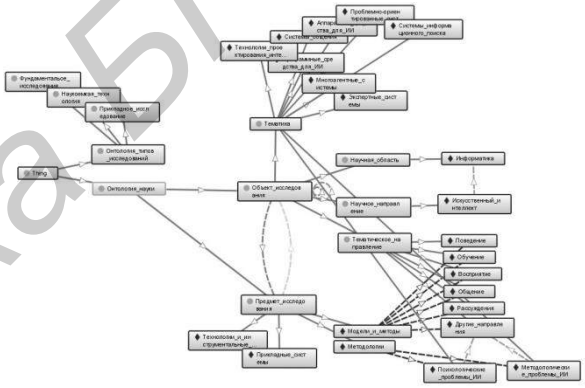

Рисунок 10 – Фрагмент онтологической модели

2.3. Технология построения карт научных направлений

Как показывает анализ литературы по средствам автоматизированного построения карт науки, в общей схеме обработки информации целесообразно выделять этапы

- выявления центров компетенции (включая организации и авторские коллективы) в предметной области,
- формирования репрезентативной коллекции документов с учетом выявленных на предыдущем этапе центров превосходства и компетенции и
- собственно обработки сформированных коллекций документов.

Технология построения семантических карт научного направления представлена на Рис. 11.

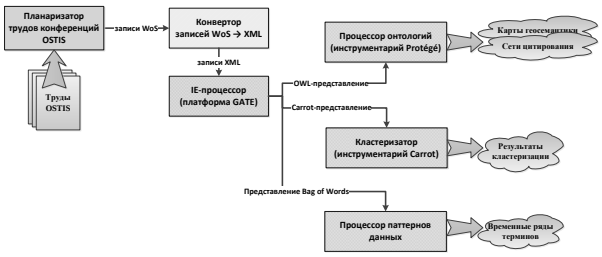

Рисунок 11 – Технология построения семантических карт научного направления

В качестве инструментария для извлечения информации из текстов использована платформа GATE, расширенная плагином русской морфологии, а также специально разработанными модулями обработки именных групп и генерации статистических портретов документов.

Для удобства анализа результатов обработки коллекций документов был разработан специальный модуль генерации OWL-представлений, которые загружались в систему Protégé[9] и использовались для дальнейшего анализа. При построении карты научного направления OSTIS в технологической цепочке использовался кластеризатор Carrot[10].

2.4. Методика обработки информации и анализа результатов

Методика обработки коллекции публикаций OSTIS состояла в следующем:

- Труды конференций за каждый год конвертировались в XML-представление, принятое в качестве стандарта для публикаций Web of Science.
- Сконвертированные труды обрабатывались гибридным модулем извлечения информации из текстов, на выходе которого формировались статистические и семантические портреты публикаций данного тома.
- Результаты загружались в систему
  - Carrot для обработки с помощью разных методов кластеризации и визуализации состава отдельных кластеров и взаимосвязей между ними;
  - Protégé для построения карт геосемантики направления и выявления в нем «скрытых коллективов».
- На статистических портретах отдельных конференций строились временные ряды терминов направления OSTIS.

Анализ полученных результатов выполнялся экспертами, которые формировали заключение по картированию научного направления и центрах компетенций и превосходства в данной области.

[9] http://www.protege.com
[10] http://www.Carrot.com

3. Наукометрический анализ конференций OSTIS

3.1. Корпус публикаций OSTIS

Как отмечалось выше, для наукометрического анализа направления «Открытые семантические технологии проектирования интеллектуальных систем» (OSTIS) были выбраны труды профильных конференций этой серии за 2012-2014 гг. Общий объем OSTIS-корпуса составил 273 статьи.

Все опубликованные статьи были обработаны с помощью специально разработанного конвертора, на вход которого подавались исходные тексты, а на выходе формировались их XML-представления. Для примера, на Рис. 12 приведен фрагмент одного из результатов таких XML-представлений.

```
<?xml version="1.0" encoding="UTF-8"?>
<Document>
  <meta>
    <prop name="uid">OSTIS_2013.txt</prop> …
  </meta>
<record>
      <PT>J</PT>
      ……
      <AF>Грибова В.В., Клещев А.С.</AF>
      <TI>ОБЛАЧНЫЕ …</TI>
      <LA>Russian</LA>
      <AB>В работе описана …</AB>
      <C1>[Грибова В.В., Клещев А.С.] Федеральное
          государственное …, г. Владивосток, Россия.</C1>
      <CR>
          [Bachant et al., 1984] Bachant, J. McDermott, J.
              R1 revisited: four years …
          ……..
          [Грибова и др., 2011] Грибова В.В., Клещев А.С.,
              Крылов и др. Проект IACPaaS – развиваемый …
      </CR>
      <PY>2013</PY>
      <WC>интеллектуальные системы; облачные технологии;
семантические технологии; онтологии; базы знаний</WC>
</record>
……
</Document>
```

Рисунок 12 – Фрагмент XML-представления статьи

В соответствии с представленной выше технологией полученные XML-представления обрабатывались гибридными IE-системами с целью выявления семантически значимых объектов и связей между ними, что позволило построить статистические и семантические портреты OSTIS, а также провести его наукометрический анализ.

3.2. Статистические портреты OSTIS

3.2.1. *Общий портрет направления*

Структура публикаций по исследованиям и разработкам в области открытых семантических технологий для проектирования интеллектуальных систем, как она представлена в трудах конференций серии OSTIS (после авторского обобщения тематики и направлений секций), зафиксирована в Табл. 1.

Понятно, что использованные для заполнения таблицы исходные рубрикаторы секций, не претендует на полноту охвата данной предметной области, но дают основу для оценки ситуации. Заметим также, что таблица формировалась на основе оглавлений трудов конференций и не всегда

соответствует общему числу действительно обработанных статей.

Таблица 1 – Структура публикаций конференций OSTIS

	2012	2013	2014
Семантические модели представления и обработки знаний и их программная и аппаратная реализация	13	16	10
Семантические технологии проектирования баз знаний, программ и пакетов программ.	13	16	
Семантические модели поиска, классификации/кластеризации и решения задач.	9		9
Компьютерная лингвистика и семантические технологии проектирования мультимодальных и ЕЯ-интерфейсов.	16	20	16
Онтологическое моделирование и онтологический инжиниринг.	6	10	15
Прикладные интеллектуальные системы.	12		31
Логико-семантические модели.		20	10
Научное наследие Мартынова.			4
Когнитивное моделирование.			6

3.2.2. *Статистика авторов и геостатистика*

Статистические портреты, представленные в Табл. 2, сформированы на основании авторских указателей в трудах конференций OSTIS, а остальные параметры получены в результате автоматической обработки OSTIS-корпуса.

Таблица 2 – Статистика авторов и геостатистика

	2012	2013	2014
Авторы публикаций	130	169	171
Страны	6	6	6
Города	20	27	30
Организации[11]	53	57	77

Заметим, что в каждой из конференций серии OSTIS участвовало 6 стран, но лишь 4 из них (Россия, Беларусь, Украина и Казахстан) были участниками всех конференций. Остальные две страны – следующие: Болгария, Испания (2012); Китай, Таджикистан (2013); США, Латвия (2014).

3.3. Семантические портреты OSTIS

В соответствии с методикой картирования научных направлений, представленной выше, для построения семантических портретов OSTIS все публикации одного года обрабатывались совместно под управлением онтологии предметной области (Рис. 10). Полученные результаты обсуждаются в следующих подразделах.

3.3.1. *Геосемантика*

Для выявления геосемантики OSTIS из общей онтологии направления с помощью специальных запросов были выделены все объекты типа Location, связанные отношениями locatedIn, в результате чего были сформированы общие карты, представленные на Рис. 13.

Для более детального обсуждения полученных результатов из общих карт были сформированы карты странового участия в конференциях серии OSTIS. Для примера на Рис. 14 представлена карта геосемантики Беларуси, которая демонстрирует явно выраженное увеличение представительства организаций из этой страны, что, в свою очередь, демонстрирует развитие белорусского кластера исследований и разработок в данной области

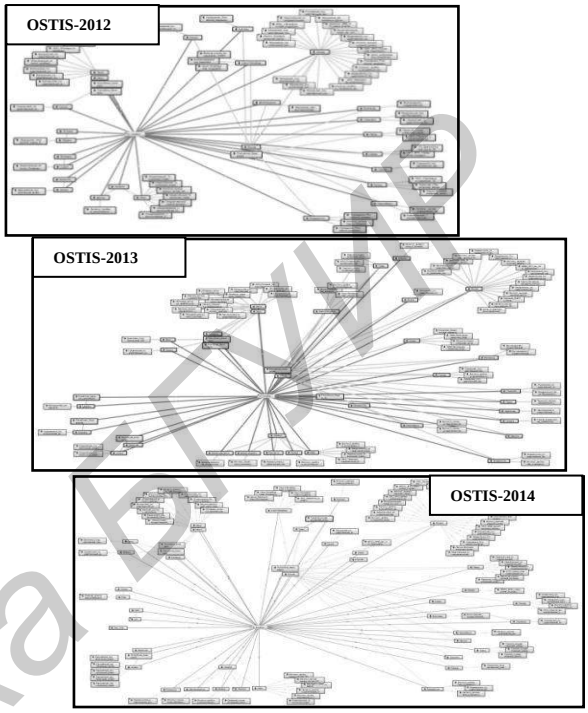

Рисунок 13 – Семантическая карта OSTIS: «Страны-Города-Организации»

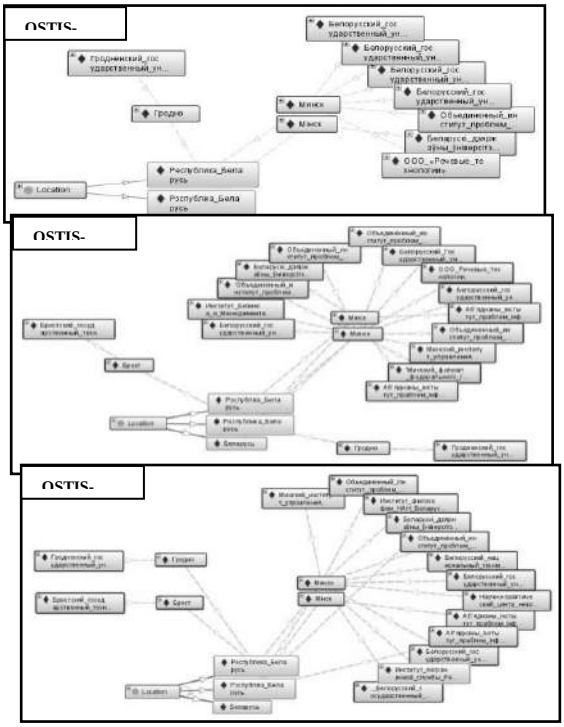

Рисунок 14 – Семантическая карта OSTIS: «Беларусь-Города-Организации»

[11] При подсчете числа синонимы названий организаций не учитывались.

Аналогичные тренды наблюдаются и для представительства организаций, участвовавших в конференциях серии OSTIS из других стран.

3.3.2. *«Скрытые коллективы»*

Для поиска в корпусе авторов, представленных в рудах конференций серии OSTIS, из общей онтологии направления с помощью специальных запросов были выделены все объекты типа Author и Reference, связанные отношениями beCoathor и referencedBy. locatedIn, в результате чего были сформированы общие карты цитирования. Для примера на Рис. 15 показана такая карта для конференции OSTIS-2012.

Как показывает анализ данной карты цитирования, «хорошей» визуализации результатов выявления «скрытых коллективов» с помощью стандартных средств системы онтологического инжиниринга Protégé (в силу значительного числа присутствующих в онтологии объектов и отношений) достичь практически невозможно.

Поэтому в данном случае авторы пошли по пути уменьшения структурной сложности визуализации за счет выделения с помощью специальных запросов фрагментов сетей цитирования для отдельных авторов, представленных в трудах конференций серии OSTIS.

Для примера, на Рис. 16 показана сети цитирования для автора Голенков, а на Рис. 17 и Рис. 18 – для авторов Гаврилова и Хорошевский.

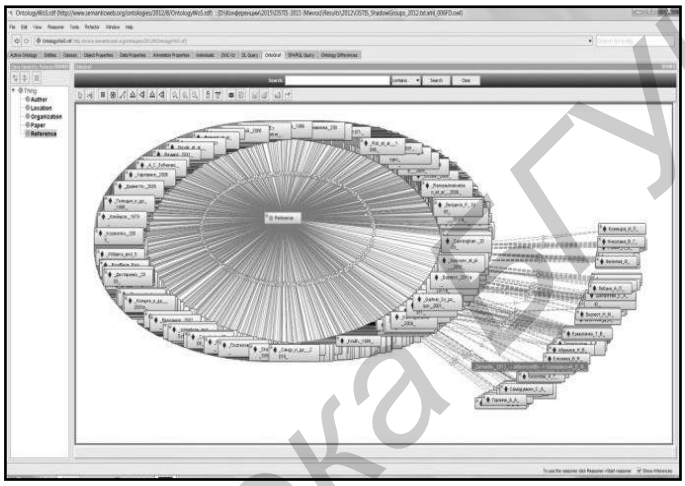

Рисунок 15 – Семантическая сеть цитирования (OSTIS-2012)

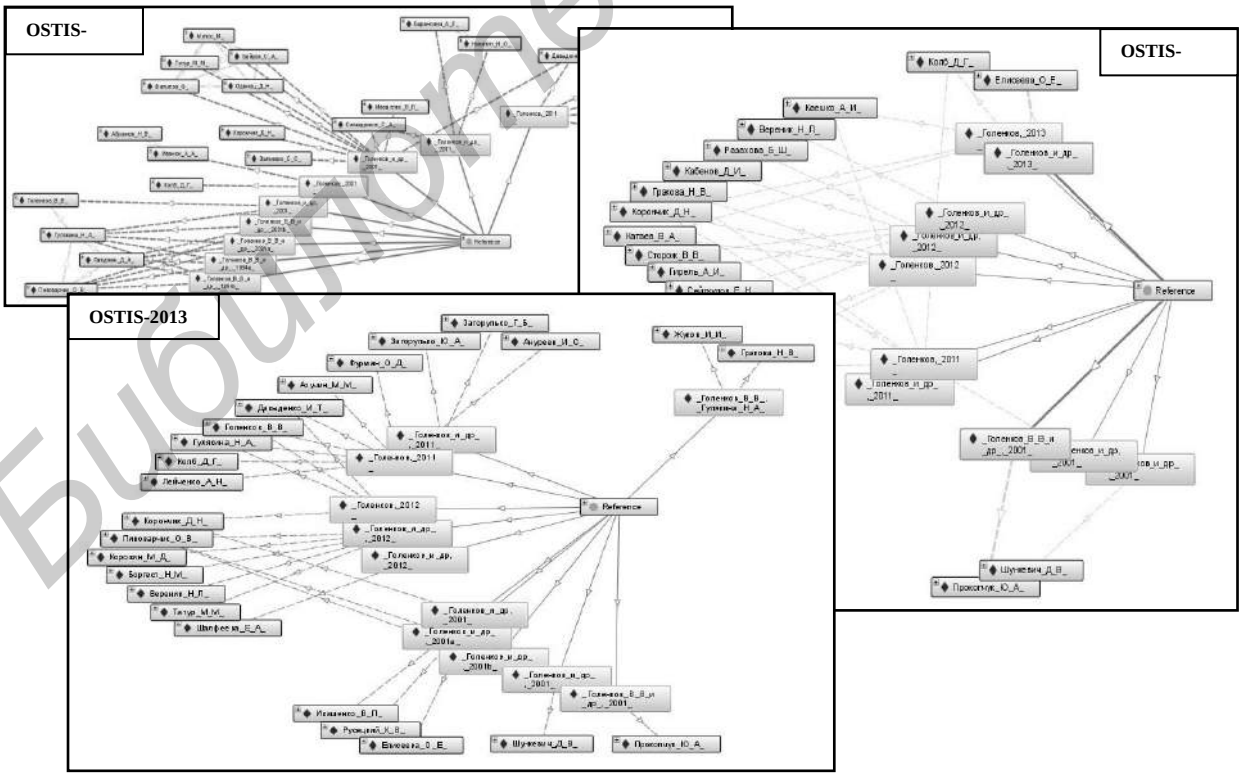

Рисунок 16 – Динамика развития сетей цитирования для автора Голенков

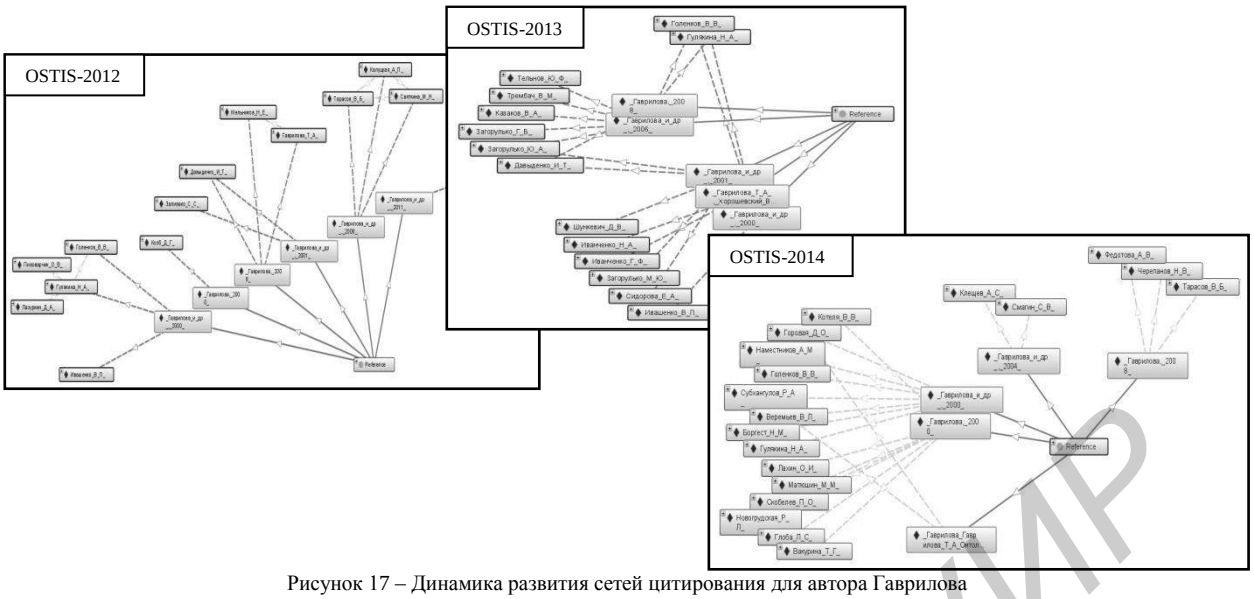

Рисунок 17 – Динамика развития сетей цитирования для автора Гаврилова

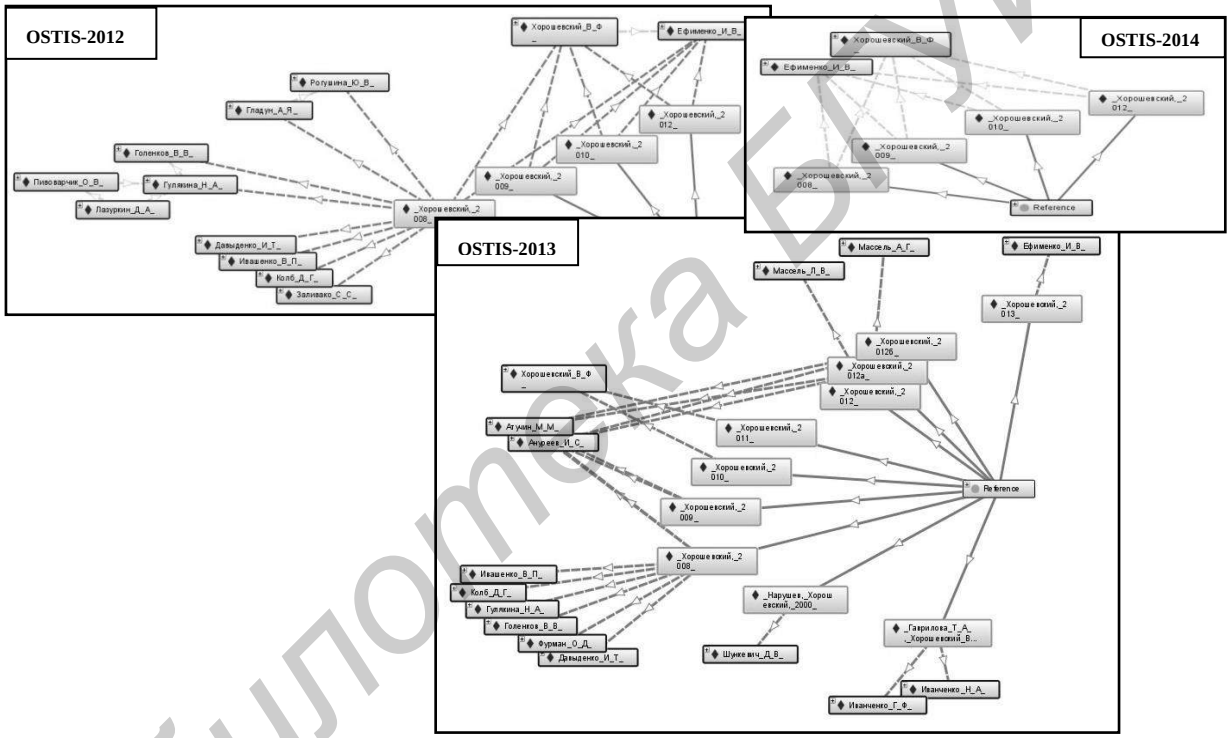

Рисунок 18 – Динамика развития сетей цитирования для автора Хорошевский

Как показывает анализ сетей цитирования, представленных на Рис. 16, для данного автора наблюдается хорошая динамика изменения состава цитируемых публикаций. Так, например, в 2012 г. основные цитирования датированы 1998г. и 2001г., в в 2013 и 2014 гг. наблюдается сдвиг датирования цитируемых работ на 2011г. и 2012г. В 2014г. появляются цитирования работ данного автора (и с соавторами), опубликованных в 2013г. Таким образом, можно констатировать, что авторы публикаций в трудах конференций серии OSTIS достаточно активно используют результаты работ В.В. Голенкова (и с соавторами) и отслеживают новые работы данного автора.

Как показывает анализ сетей цитирования, представленных на Рис. 17-18, для данных авторов наблюдается иная динамика изменения состава цитируемых публикаций.

Так, для Т.А. Гавриловой (одного из самых известных в России и за ее пределами специалиста в области онтологического инжиниринга) наиболее цитируемой публикацией в трудах всех конференций серии OSTIS является совместная с В.Ф. Хорошевским монография [Гаврилова и др., 2000] и переиздание этой монографии в 2001г., а из других публикаций цитируются только публикация по онтологическому инжинирингу на сайте[12] и работы [Гаврилова и др., 2006; Гаврилова и др., 2008]. Такая ситуация может объясняться, с одной стороны, важностью выше упомянутой монографии,

[12] http://www.big.spb.ru/publications/bigspb/km/

которая уже стала классической, а с другой – недостаточным доступом к другим публикациям данного автора.

Несколько другая, но, вместе с тем, тоже не лучшая ситуация наблюдается и для динамики цитирования в трудах конференций серии OSTIS публикаций В.Ф. Хорошевского. И здесь, в основном, цитируется уже упоминавшаяся монография по базам знаний интеллектуальных систем, а также обзорные работы [Хорошевский, 2008; Хорошевский, 2009; Хорошевский, 2012a]. При этом достаточно свежие работы данного автора в области семантических технологий [Хорошевский, 2012b; Хорошевский, 2013] имеют либо автоссылки, либо ссылки его соавторов по другим работам.

Вместе с тем, анализ сетей цитирования, полученных в результате обработки трудов конференций серии OSTIS, показывает, что таких специалистов, как В.В. Голенков, Т.А. Гаврилова и В.Ф. Хорошевский (с соавторами) можно идентифицировать в качестве центров «скрытых коллективов» в области семантических технологий.

Кластеры направлений

Кластеризация направлений, представленных в трудах конференций серии OSTIS, осуществлялась с использованием гибридного подхода. При этом целью обработки каждой статьи является формирование ее семантического портрета в виде «мешка слов» (Bag of Words). В нашем случае такие портреты формируются из авторских ключевых слов, а также из статистически значимых терминов, выделенных из названий статей, аннотаций и названий работ в ссылках. Для каждого года OSTIS семантические портреты отдельных статей объединяются в общий портрет, фрагмент которого приведен на Рис. 19.

```
<?xml version="1.0" encoding="UTF-8"?>
<searchresult>
  <query>OSTIS_2013.txt.xml_000E3</query>
  <document id="СИСТЕМА УПРАВЛЕНИЯ…">
    <title>СИСТЕМА УПРАВЛЕНИЯ…</title>
    <url>D:/ScienceMap/СИСТЕМА УПРАВЛЕНИЯ… СИСТ
    <snippet>система управление проектирование интеллект
  </document>
  ……
  <document id="ПОСТРОЕНИЕ…">
    <title>ПОСТРОЕНИЕ…</title>
    <url>D:/ScienceMap/ПОСТРОЕНИЕ… ПРОЦЕССОРА.txt
    <snippet>аппаратный реализация; интеллектуальный сис
  </document>
</searchresult>
```

Рисунок 19 – Фрагмент семантического портрета OSTIS-2013

Полученные портреты обрабатывались по годам с помощью кластеризатора Carrot. Ниже обсуждаются результаты кластеризации портретов OSTIS, представленные на Рис. 20-21.

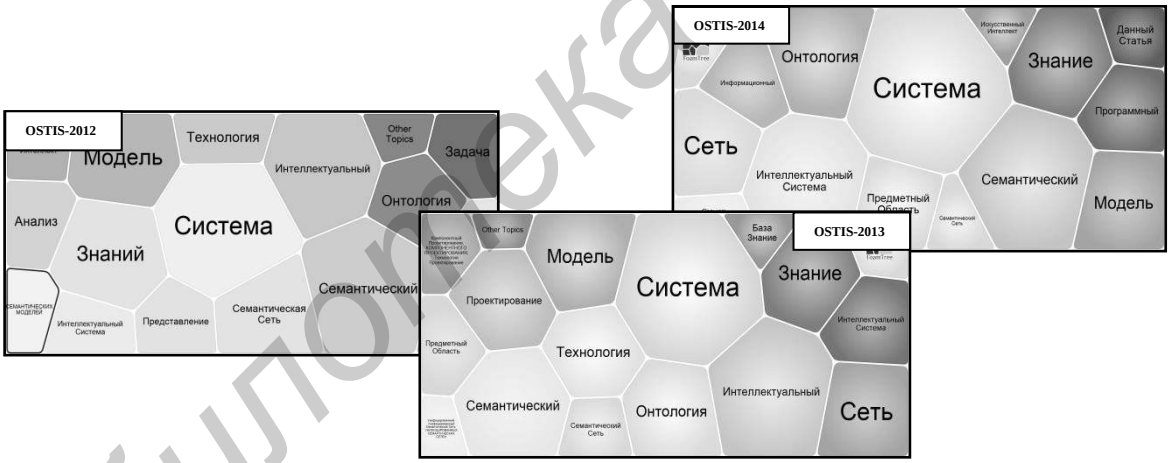

Рисунок 20 – FoamTree представление результатов кластеризации направлений OSTIS (метод STC)

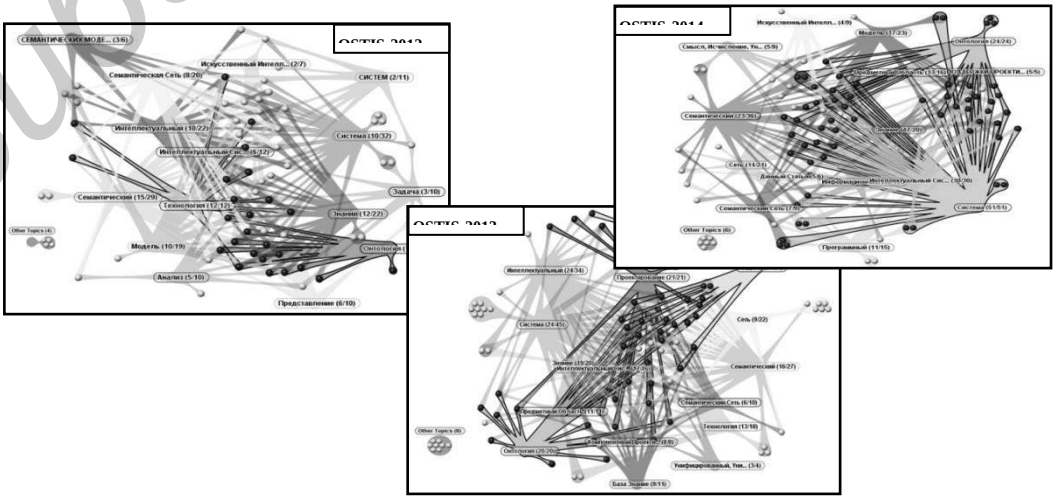

Рисунок 21 – AdunaMap представление результатов кластеризации направлений OSTIS (метод STC)

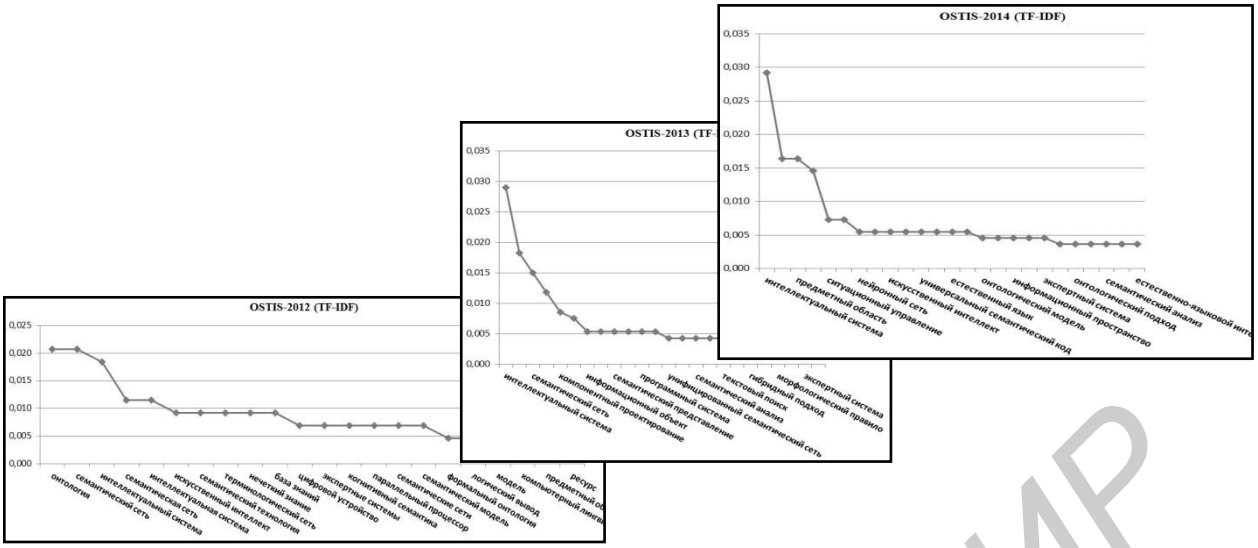

Рисунок 22 – Представление 25 наиболее частотных терминов OSTIS по параметру TF-IDF

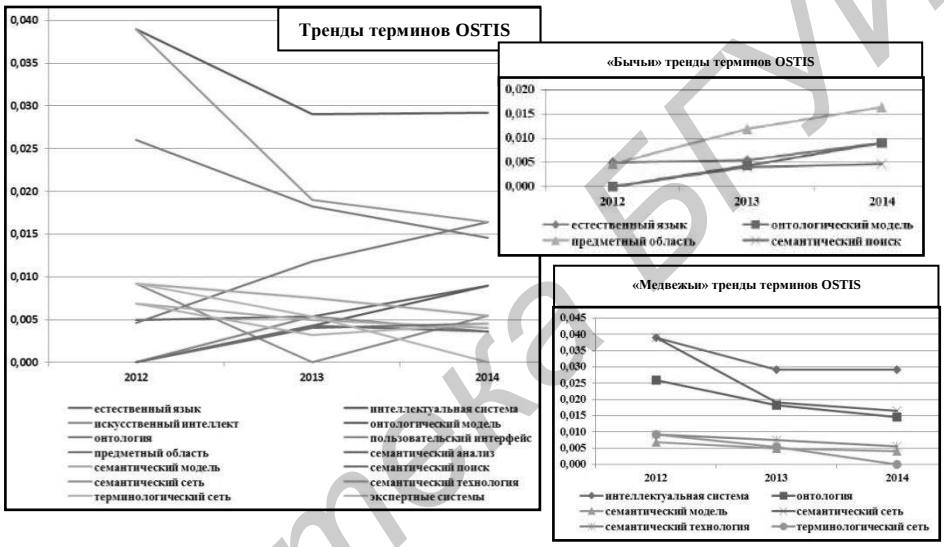

Рисунок 23 – Паттерны данных терминов OSTIS

Результаты кластеризации, представленные в виде, аналогичном картам Кохонена (Рис. 20), показывают, что наиболее значимыми для данного направления являются топики «Система», «Модель» и «Онтология», что, впрочем, является ожидаемым.

Гораздо более интересным является представление тех же результатов в виде карт взаимосвязи отдельных публикаций портретов OSTIS с различными кластерами (Рис. 21). Во всех случаях здесь наблюдается значительный «вклад» портретов отдельных публикаций в кластеры «Онтология», «Система» и «Модель». Отсюда следует, что в конференциях серии OSTIS значительное число работ представляет результаты на стыке направлений, что подтверждается экспертами в области всего направления OSTIS.

Паттерны данных терминов

Использование гибридных методов наукометрии позволило авторам провести еще одно исследование полученных результатов – анализ паттернов данных, построенных на статистически значимых множествах терминов, выявленных в портретах по отдельным годам OSTIS-конференций. Результаты этого исследования обсуждаются ниже и представлены на Рис. 22-23.

Представленные на Рис. 22 кривые показывают, что по параметру TF-IDF наиболее значимыми терминами в 2012г. были «Онтология», «Семантические сети» и «Интеллектуальные системы», в 2013г. – «Интеллектуальные системы», «Семантические сети» и «Компонентное проектирование», а в 2014г. – «Интеллектуальные системы», «Предметная область» и «Ситуационное управление».

Паттерны данных терминов OSTIS, представленные на Рис. 23, демонстрируют разное «поведение». Так, термины «Естественный язык», «Предметная область», «Онтологическая модель» и, частично, «Семантический поиск» относятся к «бычьим» трендам, а термины «Интеллектуальные системы», «Онтология», «Семантическая модель», «Семантическая сеть», «Семантические технологии» и «Терминологические сети» – к «медвежьим» трендам. Следует сразу отметить, что

для серьезных выводов и обобщений данных мало, но можно предположить, что отмеченные тенденции связаны с естественной диверсификацией терминов.

Проблемы и направления дальнейших исследований

В современой наукометрии существует много проблем, среди которых можно выделить:

- организационно-технические проблемы и
- проблемы чисто научного плана.

В первом классе проблем явно доминируют сложности получения представительных и качественных электронных корпусов исходных данных, поскольку публикации существенно отличаются и по форматам, и по качеству их оформления. Для примера можно отметить, что авторы статей, опубликованных в трудах конференций серии OSTIS, далеко не всегда следуют заданным организаторами стандартам. Особенно часто наблюдаются несоответствие и ошибки в оформлении списков литературы к статьям, а также «вольная» трактовка авторами названий организаций, в которых они работают, и геоимен. Следует также отметить недооценку авторами таких разделов статьи, как аннотация и списки ключевых терминов. К сожалению, эти проблемы в большей степени характерны для нашей действительности. В странах с устоявшимися научными традициями большинство из указанных проблем уже решены – как за счет автоматизации процессов приема авторских материалов для публикаций, так и более ответственным отношением самих авторов к процессам подготовки и оформления статей.

Вместе с тем, в современной наукометрии существуют и чисто научные проблемы, к числу которых, в первую очередь, относятся:

- разработка эффективных и надежных методов идентификации авторов и их аффиляций,
- разработка новых моделей семантизации публикаций,
- разработка новых моделей и методов оценки публикационной активности,
- разработка новых моделей и методов выявления и оценки перспективных направлений исследований и разработок, а также
- создание инструментов наукометрии нового поколения, где классические методы были бы интегрированы с методами искусственного интеллекта и компьютерной лингвистики.

С учетом вышесказанного и анализа полученных в рамках настоящего исследования результатов, авторы планируют в дальнейшем сосредоточиться на разработке и реализации более мощных и адекватных методов и средств использования семантических технологий в наукометрии.

Заключение

В работе представлено обсуждение вопросов картирования научных направлений с использованием семантических технологий на примере анализа корпуса статей, опубликованных в трудах конференций серии OSTIS «Open Semantic Technologies for Intelligent Systems». Для понимания изложения дано краткое описание основных понятий наукометрии и обосновывана важность применения семантических технологий в решении возникающих здесь задач. Полученные результаты показывают, что предложенные модели, методы и средства анализа научных направлений и выявления центров компетенций и центров превосходства на базе семантизации методов наукометрии позволяют оценить и статику, и динамику развития исследований и разработок в определенных предметных областях, что должно обеспечить автоматизацию процессов поддержки принятия решений.

Авторы считают своим приятным долгом поблагодарить административную группу сайта OSTIS за оперативное предоставление электронных версий трудов всех конференций OSTIS без чего настоящее исследование было бы невозможным.

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».

Библиографический список

[Borner et al., 2003] Borner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. Annual Review of Information Science and Technology 37, 179-255.

[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., 2005] Boyack, K. W., Klavans, R., & Borner, K. (2005). Mapping the backbone of science. Scientometrics, 64(3), 351-374.

[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.

[Boyack et al., 2014] Boyack, K. W., & Klavans, R. (2014). Creation of a highly detailed, dynamic, global model and map of science. Journal of the Association for Information Science and Technology, 65(4), 670-685.

[Boyack, 2009] Boyack, K. W. (2009). Using detailed maps of science to identify potential collaborations. Scientometrics, 79(1), 27-44.

[**Efimenko et al., 2014]** Efimenko I., Khoroshevsky V. New Technology Trends Watch: an Approach and Case Study. In: Proc. of AIMSA-2014.

[Erdi et al., 2013] Erdi P., Makovi K., and et al. Prediction of emerging technologies based on analysis of the US patent citation network. Scientometrics, 2013, 95 (1), pp. 225-242 (2013).

[Klavans et al., 2006] Klavans, R., & Boyack, K. W. (2006). Quantitative evaluation of large maps of science. Scientometrics, 68(3), 475-499.

[Klavans et al., 2010] Klavans, R., & Boyack, K. W. (2010). Toward an objective, reliable, and accurate method for measuring research leadership. Scientometrics, 82(3), 539-553.

[Klavans et al., 2011] Klavans, R., & Boyack, K. W. (2011). Using global mapping to create more accurate document-level maps

of research fields. Journal of the American Society for Information Science and Technology, 62(1), 1-18.

[Klavans et al., 2014a] Klavans R., Boyack K. W. Scientific Superstars and their Effect on the Evolution of Science. // Proc. Of ENID STI Conference, Rome, Italy, 2011, p. 7-9.

[Klavans et al., 2014b] Klavans R., Boyack K.W., Small H. Indicators and precursors of hot science. // Proc. Of 17th International Conference on Science and Technology Indicators, Leiden, Netherlands, 2014, p.475-487.

[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).

[Shibata, et al., 2008] Shibata N., Kajikawa Yu., Takeda Y., Matsushima K.: Detecting emerging research fronts based on topological measures in citation networks of scientific publications, Technovation, Vol. 28, Issue 11, November 2008, pp. 758–775 (2008).

[Small, 2010] Small, H. (2010). Maps of science as interdisciplinary discourse: Co-citation contexts and the role of analogy. Scientometrics, 83(3), 835-849.

[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.

[Van Eck et al., 2014] Van Eck, N.J., & Waltman, L. (2014). CitNetExplorer: A new software tool for analyzing and visualizing citation networks. Journal of Informetrics, 8(4), 802-823.

[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.

[Witten et al., 2011] Witten Ian H., Frank Eibe and Hall Mark A. Data Mining: Practical Machine Learning Tools and Techniques. — 3rd Edition. — Morgan Kaufmann, 2011. — P. 664.

[Гаврилова и др., 2000] Гаврилова, Т.А. Базы знаний интеллектуальных систем/ Т.А. Гаврилова, В.Ф. Хорошевский //СПб – Питер, 2000 г., 384 с.

[Гаврилова и др., 2006] Гаврилова, Т.А. Модели и методы формирования онтологий / Т.А. Гаврилова, Д.В. Кудрявцев, В.А. Горовой // Научно-технические ведомости СПбГПУ, № 4, 2006. – С.21-28.

[Гаврилова и др., 2008] Гаврилова, Т. А. Визуальные методы работы со знаниями: попытка обзора / Т. А. Гаврилова, Н. А. Гулякина // Искусственный интеллект и принятие решений, 2008, № 1, С. 15-21.

[Евстигнеев, 1987] Евстигнеев В.А. Методы теории графов в наукометрии: исследование структуры пространства журналов и незримых коллективов в программировании. // Новосибирск, 1987. (Препр. АН СССР. Новосиб. филиал. ИТМ и ВТ им. С.А.Лебедева; № 4).

[Егоров и др., 2006] Егоров В. С., Пожидаев А. В., Чернобровская Т. Н. Систематизация и использование сведений о научных мероприятиях в автоматизированной технологии ВИНИТИ. // НТИ. Сер. 1. – 2006. – №4.– С.17-23.

[Крюков и др., 2013] Крюков К.В., Кузнецов О.П., Суховеров В.С. О понятии формальной компетентности научных сотрудников. // Труды международной конференции OSTIS-2013, Минск, Беларусь, 2013.

[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.

[Маршакова, 1988] Маршакова И.В. Система цитирования научной литературы как средство слежения за развитием науки. — М.: Наука, 1988. 288 с.

[Налимов и др., 1969] Налимов В.В. Мульченко З. М. Наукометрия. Изучение науки как информационного процесса / В. В. Налимов, З. М. Мульченко. — М.: Наука, 1969. — 192 с.

[Налимов и др., 1971] Налимов В.В., Кордон И.В., Корнеева А.Я. Географическое распределение научной информации // Информационные материалы Научного совета по комплексной проблеме "Кибернетика" АН СССР. 1971. № 2 (49). С. 3-37.

[Паклин и др., 2009] Паклин Н. Б., Орешков В. И. Бизнес-аналитика: от данных к знаниям (+ CD). — СПб.: Изд. Питер, 2009. — 624 с.

[**Хайтун, 1989]** Хайтун С.Д. Проблемы количественного анализа науки. - М.: Наука, 1989. 280 с.

[Хорошевский и др., 2014] Хорошевский В.Ф., Ефименко И.В. Искусственный интеллект: карта научного направления в трудах конференций РАИИ. // Труды 14-й национальной конференции по искусственному интеллекту с международным участием, КИИ-2014, Том 1. С. 160-168. Казань, Россия

[Хорошевский, 2008] Хорошевский, В.Ф. Пространства знаний в сети Интернет и Semantic Web (Часть 1) // Искусственный интеллект и принятие решений. - 2008. - № 1. - С.80-97.

[Хорошевский, 2009] Хорошевский В.Ф. Пространства знаний в сети Интернет и Semantic Web (Часть 2) // Искусственный Интеллект и Принятие решений, № 4, 2009, с.15-36.

[Хорошевский, 2012b] Хорошевский В.Ф. Выявление новых технологических трендов: проблемы и перспективы. // Труды 13-й Конференции по Искусственному Интеллекту с международным участием, КИИ-2012, Том 1, с. 252-259. Белгород, Россия, 2012. Изд-во БГТУ, 2012.

[Хорошевский, 2012a] Хорошевский В.Ф. Пространства знаний в сети Интернет и Semantic Web (Часть 3) // Искусственный Интеллект и Принятие решений, № 1, 2012, с.3-38.

[Хорошевский, 2013] Хорошевский В. Ф. Автоматизация процессов выявления технологических трендов в системе АРМ Тренд. // Материалы III международной научно- технической конференции «Открытые семантические технологии проектирования интеллектуальных систем (OSTIS-2012)». Минск, 2013. С. 233–241.

[Юрков, 2015] Юрков А. В. Интернет-аналитика для прикладной наукометрии. // Труды международной конференции SCIENCE INDEX 2015: аналитические инструменты и сервисы для оценки научной деятельности, 17-24 января 2015 г., Андорра. http://elibrary.ru/projects/conference/andorra2015/conf_2015_1_presentations.asp

SCIENTOMETRICS OF THE DOMAIN: OSTIS-CONFERENCES CASE STUDIES

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

**Dorodnicyn Computing Centre RAS, Moscow, Russia*

khor@ccas.ru

***Faculties of Humanities, NRU HSE, Moscow, Russia*

iefimenko@hse.ru

The paper discusses “maps of science” based on advanced scientometrics methods using semantic technologies. Text collections of articles published in the proceedings of OSTIS Conference («Open Semantic Technologies for Intelligent Systems», 2012-2014) are used as input data. The proposed models, methods and tools are oriented towards identification of centers of excellence based on “scientometrics semantization” techniques. The results obtained allow users to analyze static and dynamic aspects of research and development in specific S&T fields, which is a basis for the decision support.