УДК 81’33

ПРОСТРАНСТВЕННЫЕ, ВРЕМЕННЫЕ И ДИСЦИПЛИНАРНЫЕ АСПЕКТЫ РАСПРОСТРАНЕНИЯ ПОНЯТИЯ «ИНФОРМАЦИОННЫЙ РЕСУРС» [1]

Дмитрий Александрович Баранов
преподаватель кафедры математического обеспечения информационных систем
Оренбургский государственный университет
460018, г. Оренбург, пр-т Победы, д. 13. baranov@semograph.com

Константин Игоревич Белоусов
д. филол. н., профессор кафедры теоретического и прикладного языкознания
Пермский государственный национальный исследовательский университет
614990, г. Пермь, Букирева, 15. belousovki@gmail.com

Дилара Ахметовна Ичкинеева
к. филол. н., доцент кафедры иностранных языков естественных факультетов
Башкирский государственный университет
450076, г. Уфа, З. Валиди, 32. dilaraichkineeva@gmail.com

Описывается графосемантический подход к исследованию процессов освоения базовых понятий современной науки ее отдельными областями. На материале зарубежных научных публикаций, размещенных на платформе издательства Springer, анализируется распространение понятия «информационный ресурс» в дисциплинарных областях научного знания с учетом временного и пространственного (географического) параметров, характеризующих научные публикации. Одним из результатов исследования является прогнозная модель состояния предметных областей науки, построенная с помощью имитационного моделирования на основе марковского процесса первого порядка.

Ключевые слова: информационный ресурс; научная предметная область; понятие; корпус текстов; статистические методы; прогнозирование; имитационное моделирование; графосемантическое моделирование.

1. Введение

Понятие «информационный ресурс» (далее – ИР)[2] является одним из основных в сетецентристском информационном пространстве постиндустриального общества, поэтому используется в самых разных эпистемологических контекстах: от теоретико-методологических и организационно-технологических аспектов, возникающих при создании сложных сетевых инфраструктур для совместного использования ресурсов [Kogalovsky 2000; Lee et al. 2007; Mastroianni, Talia, Trunfio 2004 и др.] до отраслевой проблематики, актуальной при переходе от управления информацией к управлению знаниями [Bryson 2001; García, Cuello 2010; O’Leary 2002], в частности в области организации образовательного процесса, функционирования сетевой науки, в сферах управления предприятиями и отраслями экономики [Gunasekaran, Khalil, Mahbubur 2002; Senge 2001; Skyrme 2001 и др.]

Понятие ИР является родовым для существующих типов информационных ресурсов, среди которых, в частности, электронные библиотеки, информационные системы, базы данных, базы знаний, каталоги, патенты и мн. др. В данном контексте понятие ИР является репрезентантом всей предметной области ИР в научных текстах по нескольким причинам.

1. Типы ИР многообразны и напрямую зависят от развития технологическй базы. Поскольку с появлением новых форматов, созданных для хранения, обработки, анализа и трансляции информации, возникают новые типы ИР, которые могут вытеснять из сферы использования рас-

пространенные типы ИР, постольку квантитативный анализ развития всей предметной области (далее – ПрО) ИР сложно осуществим.

2. В то же время понятие ИР является понятием онтологии верхнего уровня данной ПрО, ее наиболее абстрактной сущностью, соответственно, может использоваться в более широком спектре контекстов, нежели отдельные разновидности ИР (каталоги, книги, патенты и мн. др.);

3) Кроме того, использование родового понятия свидетельствует о рефлексии исследователей над данной ПрО, а уровень отрефлектированности отражает состояние концептосферы ПрО, порождаемой в процессе осмысления понятий, составляющих онтологию ПрО[3].

Таким образом, несмотря на то что анализ распространения понятия ИР в научном контенте не оперирует количественными показателями развития всей ПрО ИР (т. е. не учитывает обращения к множеству конкретных типов ИР), он все же позволяет моделировать **динамику** развития данной ПрО, показывает уровень интереса к данной проблематике.

При этом представляется актуальным анализ использования родового понятия ПрО ИР не только во временной динамике, но также в дисциплинарном (в рамках разных наук) и территориальном аспектах.

2. Методология и технология

Исследование распространения понятия ИР в научных публикациях осуществляется в рамках разработанной исследовательской программы, представленной в ряде работ [Белоусов и др. 2015; Belousov et al. 2014; Belousov, Baranov, Zelyanskaya 2014]. В то же время данное исследование отличается от предшествующих по нескольким параметрам: а) существенному увеличению объема анализируемого научного контента; б) обращению к авторитетным зарубежным базам данных, позволяющим изучать информационное пространство мировой науки; в) изменению принципа формирования корпуса рефератов научных публикаций.

Исследование осуществляется на платформе ИС «Семограф» (http://semograph.com). Для построения графов применяется программное средство Gephi (http://gephi.org). В качестве основного инструмента исследования графосемантической модели используется R – язык программирования высокого уровня, предназначенный для статистической обработки данных (http://www.r-project.org).

2.1. Основные этапы исследования

1. Создание корпуса рефератов научных публикаций в ИС «Семограф», использующих понятие “Information Resource” («Информационный ресурс»).

В качестве источника качественного научного контента нами была взята база данных издательства Springer (http://link.springer.com), на платформе которой представлены журналы а) всего спектра научных дисциплин и направлений; б) выходящие как под издательской маркой Springer, так и под марками других авторитетных издательств. Кроме того, согласно данным Elsevier [Elsevier: Content: электр. ресурс], только на журналы издательства Springer приходится 8% качественного контента, индексируемого Scopus. Объем же всего научного контента, размещенного на платформе Springer.com, составляет более 8,5 млн документов.

В качестве запроса к БД Springer было взято точное соответствие “Information Resource”. Весь собранный **материал** насчитывает 5 871 публикацию, размещенную на платформе 1 164 журналов.

Описание документов (публикаций) на платформе Springer включает в себя следующие метаданные: аннотацию, ключевые слова, название статьи, автора(-ов), аффилиацию авторов, выходные данные публикации, DOI, название журнала, издательство, дату, топики, индустриальные сектора.

Сбор контента в ИС «Семограф» осуществлялся автоматизировано в поисково-аналитическом модуле, состоящем из следующих инструментов:

- поискового робота, осуществляющего полнотекстовый поиск необходимого контента на общедоступных веб-страницах выбранных ресурсов;
- модуля разбора (парсера) и библиотеки скриптов на языке программирования Python (http://python.org), осуществляющих автоматический разбор содержимого (контента) веб-ресурсов и выделения структурированных данных (в качестве основы для поискового робота и парсера используется Python-фреймворк Scrapy (http://scrapy.org));
- сервера полнотекстового поиска Apache Solr (http://lucene.apache.org/solr), на который отправляются структурированные данные с веб-ресурсов после их извлечения;
- специального интерфейса, встроенного в ИС «Семограф» и соотносящего поля разметки документа в Solr с полями, создаваемого для импорта данных проекта в ИС «Семограф» (этот интерфейс дает возможность непосредственного импорта данных из Solr в проект «Семографа»).

Фрейм Проекта в ИС «Семограф» в рамках данного исследования включает следующие эле-

менты: контекст, система компонентов, система метаполей, система семантических полей (далее – С-полей), инструменты полевого анализа, инструменты работы с С-полями и метаполями.

Контекст – в нашем случае, страница описания одной научной статьи. Контекст включает в себя:

- собственно **поле контекста** (в нем размещается аннотация статьи);
- **систему метаполей** (Автор(-ы), Название статьи, Дата (год) и др. данные, взятые из описания документов на платформе Springer);
- **систему компонентов** (операциональных единиц, характеризующих каждую публикацию и использующихся для экспертной классификации).

В данном исследовании в качестве компонентов были использованы **топики** (частнонаучные ПрО), описывающие научные статьи. Если ключевые слова, использовавшиеся нами ранее [Belousov et al. 2014] в качестве операциональных единиц, создаются авторами и репрезентируют основные теоретические конструкты проведенных исследований, то топики как частнонаучные ПрО, часто задаются редакторами журналов и маркируют публикацию в целом, поэтому в отличие от ключевых слов топики а) не обладают высоким уровнем вариативности (для данного объема материала можно прогнозировать количество ключевых слов ~10^4, тогда как топиков всего 672 единицы), б) имеют внутреннюю форму, позволяющую с большей определенностью относить топик к тем или иным дисциплинарным ПрО, что удобно осуществлять на полидисциплинарном материале.

Инструментарий полевого анализа позволяет осуществлять экспертную классификацию компонентов (в нашем случае – топиков). Результатом полевого анализа является **система С-полей**.

На данном этапе все множество компонентов – топиков к каждой статье – распределяется по дисциплинарным ПрО в силу большой размерности модели, которая основывалась бы только на компонентах. Так, в нашем случае общее количество топиков составило 672 единицы: от высокочастотных типа Plant Sciences (482), Educational Technology (377) до низкочастотных, таких как, например, Environmental Law (1), Historical Linguistics (1) и др. Недостатком такой модели является ее размер: в нашем случае, структурная модель состояла бы из 672 узлов.

За основу классификации (списка С-полей) был взят перечень дисциплин Springer (http://link.springer.com).

2.2 Принципы классификации компонентов (топиков) в системе С-полей

1. Вводится понятие С-поля, являющегося, в нашем случае, дисциплинарной ПрО, под которым понимается множество топиков, объединенных общей семантикой.

2. При отнесении топика к тому или иному С-полю рассматривается наличие эксплицитной выраженности в семантике топика компонентов, связанных с семантикой дисциплинарной ПрО. Например, топик Medicine/Public Health относится и к MEDICINE, и к PUBLIC HEALTH.

3. В случае отсутствия в семантике топика эксплицитных семантических компонентов, непосредственно входящих в те или иные С-поля (научные дисциплины), отнесение термина к полю осуществляется с опорой на научную традицию, в том числе для этого используются классификации Subdiscipline.

4. В качестве одного из критериев ограничения содержания поля была взята частотность употребления в выборке текстов топиков, входящих в данное С-поле. В том случае, если частотность топиков, входящих в С-поле, достигала уровня отдельной дисциплины, данное множество обретало самостоятельность. В частности, в отдельную дисциплину была выделена ПрО PLANT SCIENCES, в которую вошли самые высокочастотные топики в корпусе: Plant Sciences (482), Plant Genetics & Genomics (400) и некоторые другие. И наоборот, низкочастотные С-поля объединялись при наличии общей семантики. Например, С-поля MATHEMATICS и STATISTICS были объединены в одно С-поле MATHEMATICS AND STATISTICS для того, чтобы повысить значимость С-поля и создать саму возможность вхождения С-поля в создаваемую математическую модель. В то же время ряд низкочастотных С-полей FOOD SCIENCE AND NUTRITION, ASTRONOMY, LAW, MATERIALS, ENERGY остался в пределах границ своих дисциплинарных ПрО, для того чтобы не размывать содержание тех дисциплинарных ПрО, с которыми можно было бы объединить данные дисциплины.

Классификация топиков по С-полям (научным дисциплинам) позволила снизить размерность графа с 672 узлов (топиков) до 24 узлов (дисциплин).

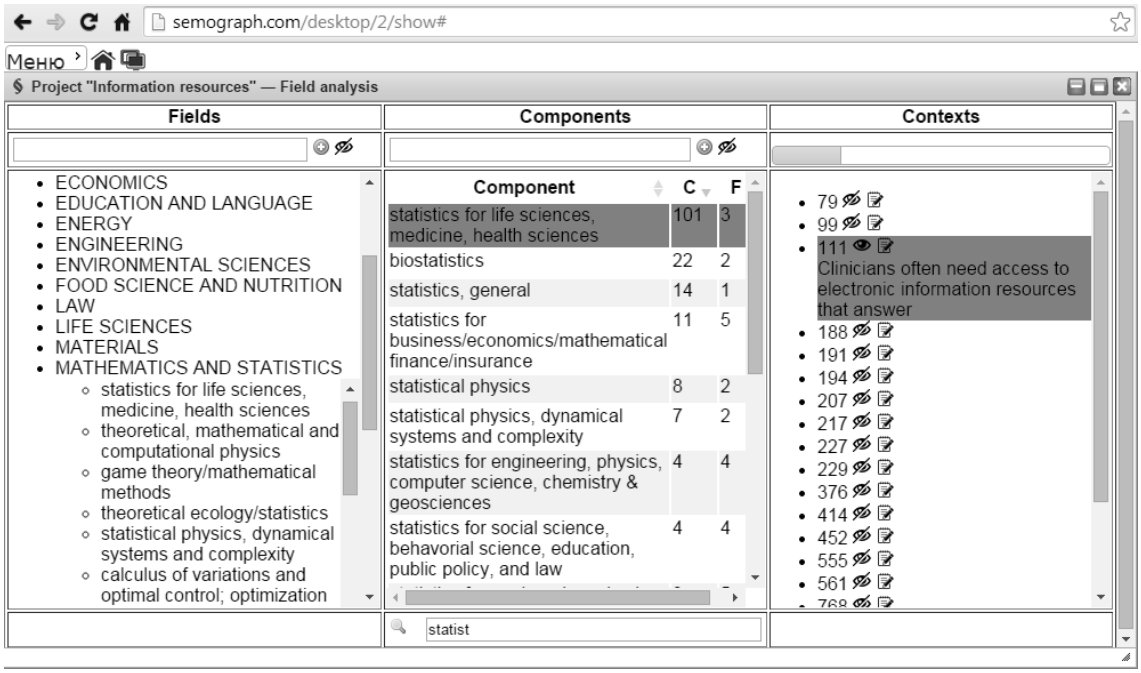


Рис. 1. Окно полевого анализа (классификация топиков по научным дисциплинам)

На рисунке 1 показано окно полевого анализа: топики отображаются в поле Components; научные дисциплины (дисциплинарные ПрО) представлены в поле Fields; в поле Context отображаются номера контекстов, в которых используется выделенный топик, и их содержание (аннотация). Столбец С-поля Components репрезентирует показатели встречаемости топика в корпусе, столбец F – показатели количества полей, к которым приписан топик.

2.3. Инструменты работы с С-полями и метаполями

Так как основные единицы модели – это метаполя и С-поля, то моделируемые отношения между ними могут быть следующие:

1. Отношения типа «Метаполя: Метаполя». Данный тип отношений позволяет выявлять распределение значений одних типов метаданных по значениям других типов метаданных. Например, рассматривать распределение публикаций в контексте «Географическая локализация (страны) – Временная динамика (года)».

2. Отношения типа «С-поля: С-поля». Этот тип отношений позволяет генерировать семантическую карту (С-карту) и семантический граф (С-граф). С-карта отражает совместное присутствие двух С-полей в одном и том же контексте с учетом подобной встречаемости во всех контекстах выборки или корпуса. Полагается, что если два компонента даются в описании одной и той же статьи, то они становятся связанными между собой через отнесение их к одному контексту. Соответствующим образом мы делаем вывод о связи между С-полями, в которые входят указанные компоненты. С-карта автоматически генерируется на основе подсчета количества связей между С-полями в пределах всей выборки/корпуса. С-граф представляет собой графическую экспликацию связей между выделенными С-полями в С-карте.

3. Отношения типа «С-поля: Метаполя». Результатом анализа данных типов отношений становится распределение значений С-полей по значениям тех или иных метаполей, например, годам, странам, дисциплинам и др.

4. Описанный метод структурирования С-графа, основывается на связности всех пар С-полей друг с другом. Однако в конкретных исследованиях количество актуализованных С-полей может быть больше двух, что делает актуальным разработку такого **метода структурирования информационного пространства, который учитывал бы композиции С-полей, присутствующие в каждой отдельной публикации**. Для этого вводится понятие единичной ПрО – композиции С-полей, формирующейся в одном научном тексте. Единичные ПрО могут быть уникальными (встречающимися только в одном тексте) и повторяющимися. Кластеризация совокупности единичных ПрО, реализованных во всем корпусе публикаций, позволяет выделить в информационном пространстве отдельные научные направления. Кроме того, на основе полученных данных строится прогноз состояния выявленных кластеров-направлений.

3. Результаты исследования: писание и интерпретация

3.1. Общие характеристики анализируемого научного контента

На рисунке 2 представлены количественные данные использования понятия ИР за период 1965–2014 гг. На графике видно, что появление интереса к проблематике ИР приходится на 1991 г., когда Всемирная паутина (World Wide Web) стала доступна в Интернете, что создало условия и перспективы для новой организации рынка ИР, а начиная с 2001 г. (создание «Википедии» со свободно распространяемым контентом) наблюдается постоянный рост интереса к ИР.

География распределения публикаций, использующих понятие ИР, представлена на рисунке 3. На диаграмме видно, что количественно доминируют публикации исследователей из США (2104 статьи, использующие понятие ИР); объем контента китайских исследователей заметно меньше. Однако если сопоставить временну́ю динамику количества публикаций американских и китайских ученых (см. рис. 4), то можно отметить, что в последние годы рост интереса к теме ИР у китайцев происходит быстрее, чем у американцев.

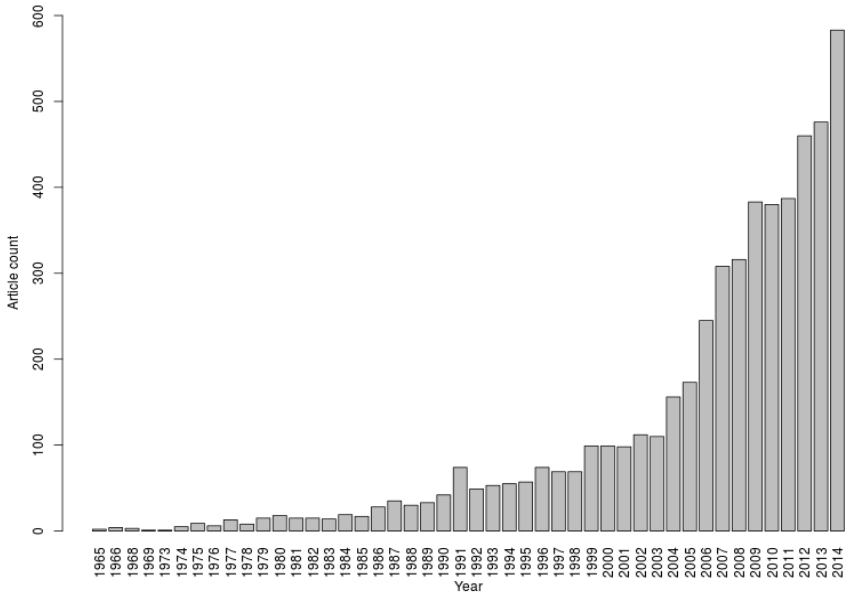


Рис. 2. Временная динамика количества публикаций, использующих понятие ИР (за период 1965–2014 гг.)

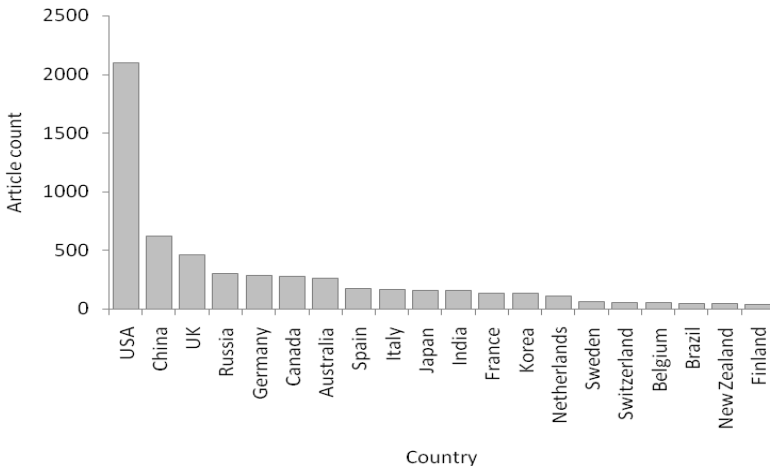


Рис. 3. География распределения публикаций, использующих понятие ИР (за период 1965–2014 гг.)

На рисунке 4 также видно, что российский научный сегмент значительно уступает американскому и китайскому, при том что стартовые условия с китайским сегментом были одинаковые; траектории «разошлись» в 2002 г.: российский сегмент демонстрирует небольшой рост, начиная с 2005 г.

Отдельно нужно остановиться на анализе значимости понятия ИР для каждой научной дисциплины.

На рисунке 5 представлено временно́е распределение объема публикаций, относящихся к тем или иным дисциплинарным ПрО за пятилетний период 2010–2014 гг. График дает представление об общем интересе каждой дисциплинарной ПрО к использованию понятия ИР: почти по всем направлениям заметен устойчивый рост распространения данного понятия. В наибольшей мере рост проявился в COMPUTER SCIENCE; заметная динамика отмечается также в областях BUSINESS AND MANAGEMENT, ENGINEERING, MEDICINE, EDUCATION AND LANGUAGE. COMPUTER SCIENCE в данном случае является основной точкой роста отрасли ИР, т. к. в настоящее время достижения в этой сфере способствуют развитию остальных областей науки, техники и социально-экономическому состоянию.

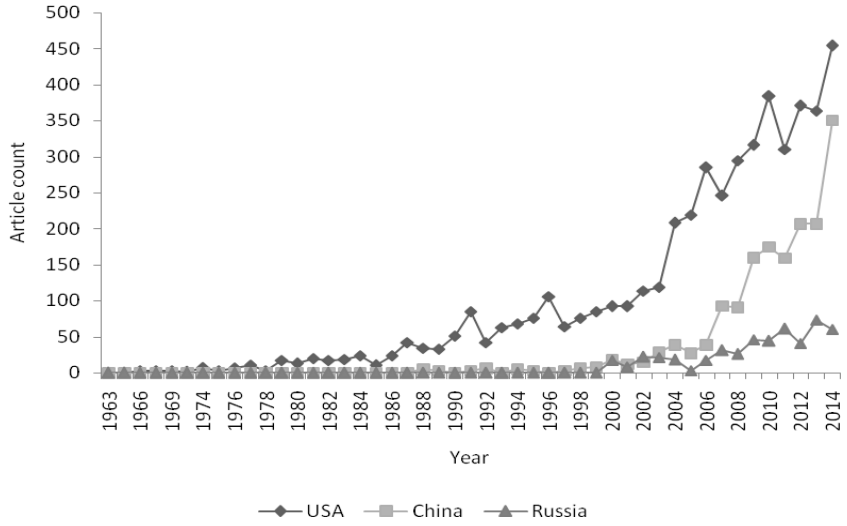


Рис. 4. Временна́я динамика количества публикаций, использующих понятие ИР, американских, китайских и российских исследователей

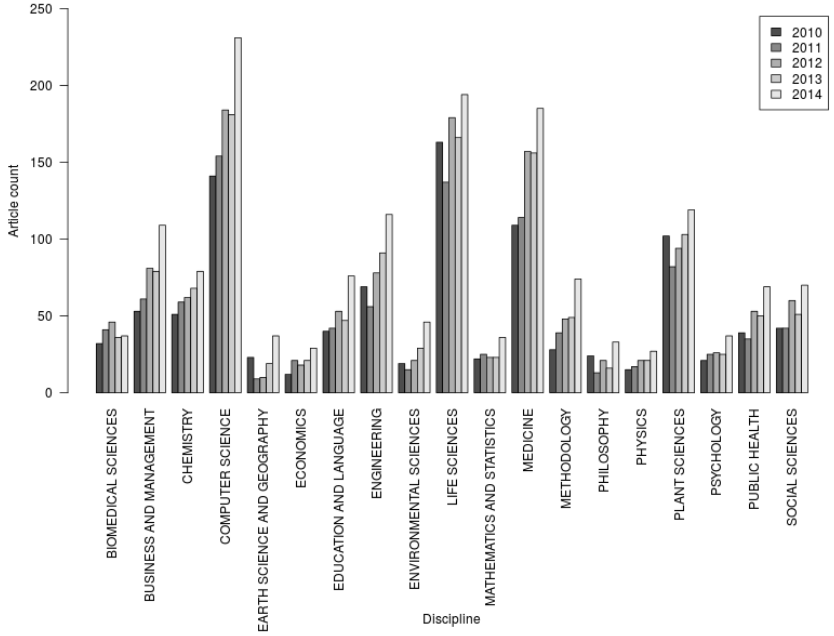


Рис. 5. Временно́е распределение частотности С-полей (дисциплин) за период 2010–2014 гг.

3.2. Анализ отношений типа «Метаполя: С-поля»

Одним из результатов использования фреймворка ИС «Семограф» является генерация таблиц сопряженности, в качестве признаков которых берутся «Метаполя : С-поля»[4], например, Годы: С-поля (дисциплины). Таблица сопряженности может использоваться в рамках метода снижения размерности признакового пространства с помощью анализа соответствий (Correspondence Analysis). Данный метод позволят визуально представить исследуемые С-поля и Годы в координатном пространстве переменных малой размерности (в частности, на плоскости) (подробнее см. [Боровиков 2003: 561–576]).

На рисунке 6 визуализированы результаты применения метода анализа соответствий к полученным данным использования понятия ИР в дисциплинарных ПрО за последний пятилетний период 2010–2014 гг.

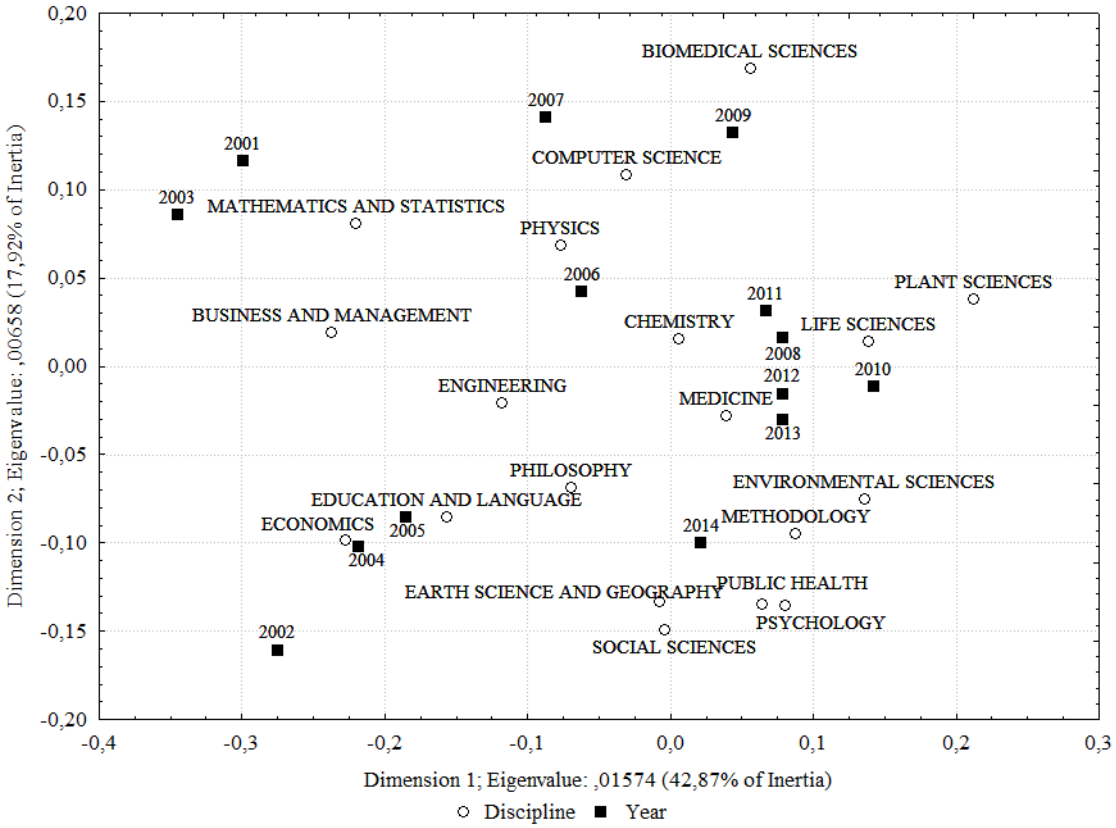


Рис. 6. Временна́я спецификация использования понятия ИР в дисциплинарных ПрО

Для понимания графика нужно учитывать следующее.

- Расстояние между данными одного типа свидетельствует о качестве связи между ними: чем меньше расстояние, тем связь сильнее, чем больше – тем слабее. В нашем случае речь может идти 1) о совместном использовании топиков в исследованиях, относящихся к разным дисциплинарным ПрО, и 2) о большем интересе отдельных наук к тематике ИР в определенные годы.
- Сила связи между данными разных типов устанавливается на основе размера угла между двумя точками (относящимися к разным типам данных), вершина которого расположена в центре тяжести графика (точке пересечения осей). Острый угол свидетельствует о положительной корреляции (чем меньше угол, тем выше корреляция); тупой угол – об отрицательной корреляции; прямой угол – об отсутствии корреляции [там же: 570–571].

На рисунке 6 видно, что основной интерес к ПрО в последние 2–3 года связан с науками «социального сектора», изучающими общество, человека, окружающую среду, в том числе, в контексте MEDICINE и PUBLIC HEALTH.

Кроме временно́й спецификации исследований, проводимых в разных научных областях, представляет интерес анализ распределения данного контента по странам (см. рис. 7). На рисунке 7 отражена дисциплинарная «специализация» стран в области научного контента, использующего понятие ИР.

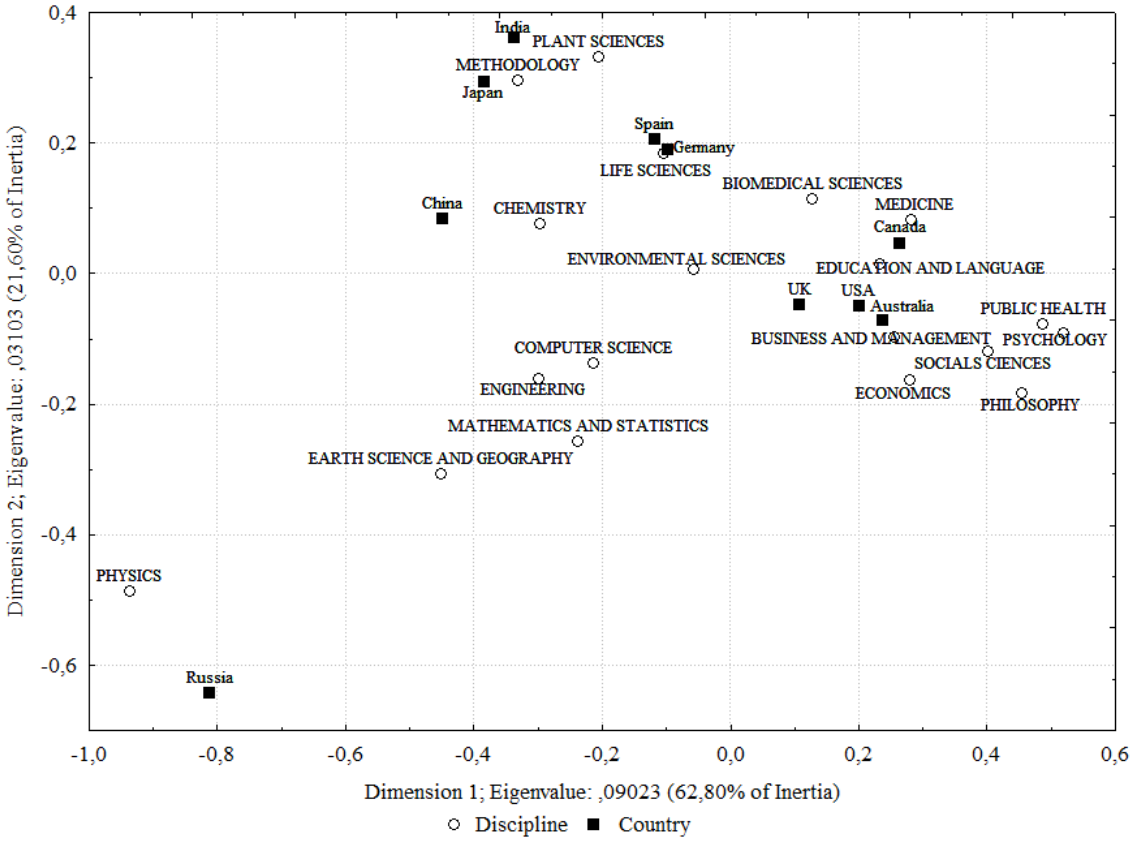


Рис. 7. Географическая (топ-10 стран) спецификация использования понятия ИР в дисциплинарных ПрО

Англоязычный мир, сгруппировавшийся в одном секторе поля, занимается в большей степени проблемами BUSINESS AND MANAGEMENT, SOCIAL SCIENCES, PSYCHOLOGY, ECONOMICS, EDUCATION AND LANGUAGE и PUBLIC HEALTH. Видно, что проблематика исследований связана, в основном, с социально-гуманитарными проблемами и сферой управления. Для испанских и немецких исследователей наиболее актуальны ИР в области LIFE SCIENCES, а для индийских ученых – в области PLANT SCIENCES. Для китайских исследователей особую значимость имеют разработки в областях ENVIRONMENTAL SCIENCES и CHEMISTRY, что вполне закономерно, если принять во внимание масштабы реального сектора китайской экономики и ее угрозы окружающей среде. Российские ученые работают в традиционных для себя областях PHYSICS, EARTH SCIENCE AND GEOGRAPHY, MATHEMATICS AND STATISTICS, COMPUTER SCIENCE и ENGINEERING. Следует отдельно отметить, что установленная «специализация» не свидетельствует о том, что ученые, относящиеся к странам с другой «специализацией», не занимаются теми же проблемами. Данная «специализация» стран указывает на особую значимость выявленных дисциплинарных ПрО в изучаемом секторе научного контента.

3.3. Анализ отношений типа «С-поля: С-поля»

Важным аспектом представленности публикации в информационном пространстве является количество топиков к статье, определяющих принадлежность ее к тем или иным ПрО. В данном случае количество топиков, описывающих публикацию, варьируется от 1 (статья однозначно закрепляется за одним частнонаучным направлением) до 12 (максимальное количество топиков к статье). При этом среднее количество топиков к статье в корпусе составляет 3,37.

Топики к одной статье могут относиться к нескольким дисциплинарным ПрО – в этом случае у публикаций появляется более высокий междисциплинарный потенциал. В то же время, частотны случаи отнесения нескольких топиков статьи к одной дисциплине. В таблице 1 представлено общее количество С-полей (дисциплин верхнего уровня) и количество С-полей на одну публикацию для каждой страны.

Таблица 1

Статистика по странам: количество публикаций, использующих понятие ИР, количество С-полей и количество С-полей на одну публикацию

Страна	Кол-во статей	Кол-во С-полей (дисциплинарных ПрО)	Кол-во С-полей на одну публикацию
USA	2104	5478	2,60
China	625	1660	2,66
UK	461	1299	2,82
Russia	302	511	1,69
Germany	287	746	2,60
Canada	280	762	2,72
Australia	261	707	2,71
Spain	172	444	2,58
Italy	167	436	2,61
Japan	163	445	2,73
India	158	484	3,06
France	137	367	2,68
Korea	135	357	2,64
Netherlands	114	353	3,10
Brazil	51	149	2,92

Заметим, что низкий показатель количества С-полей (дисциплинарных ПрО), описывающих статью (характеризующий, в частности, труды отечественных исследователей), свидетельствует о слабо выраженной междисциплинарной природе публикации. Кроме того, недостаточная маркированность статьи топиками ставит публикацию в «проигрышную позицию» при организации информационного поиска (вероятность появления такой публикации при запросе к БД почти в два раза ниже, чем у публикаций индийских и нидерландских исследователей). На рисунке 8 представлено распределение количества С-полей (дисциплин), актуализированных в публикациях, использующих понятие ИР. Видно, что минимальное количество С-полей, актуальных для каждой научной статьи 1, а максимальное – 9. Данное распределение показывает а) возможные границы междисциплинарного варьирования и б) значимость исследования структурной связности С-полей, формирующих концептуальное пространство отдельной публикации, журнала и научной дисциплины в целом, так как структурные связи отражают существующие в реальности комбинации (композиции) научных проблем и направлений.

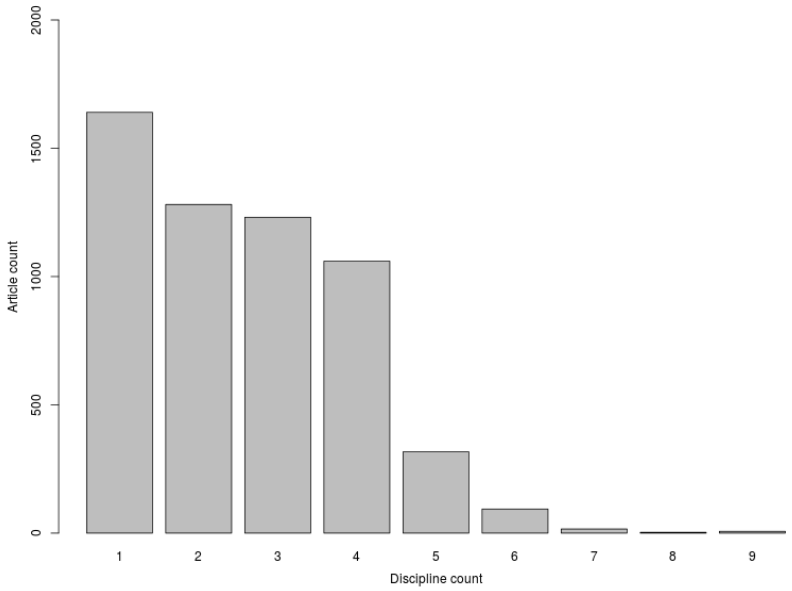


Рис. 8. Распределение количества С-полей (дисциплинарных ПрО) по научным статьям за период 2010-2014 гг.

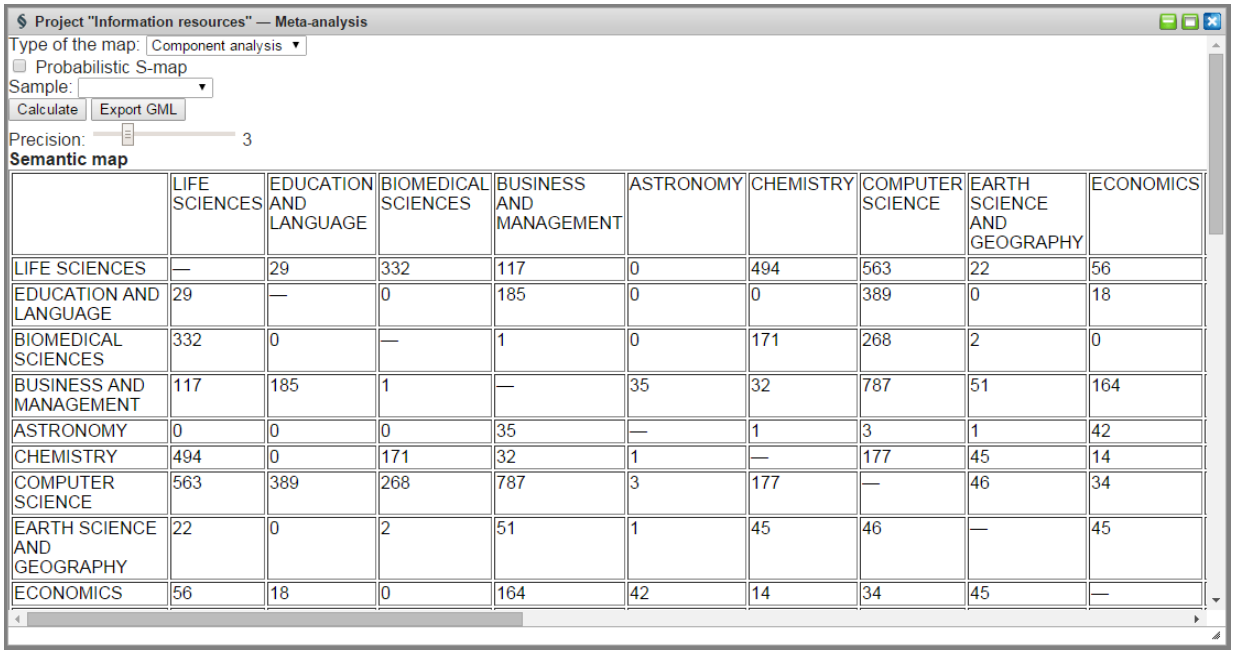
Моделирование структуры С-полей (дисциплинарных ПрО) осуществляется с помощью метода графосемантического моделирования. Генерация С-карты и С-графа может производиться на базе «сырого» материала: узлами структуры в таком случае будут являться частнонаучные ПрО (топики), а ребрами – показатели частоты их совместного использования при характеристике публикации. Недостатком такой модели, как уже говорилось, является ее размер. Поэтому представляет интерес обобщенная до дисциплинарных ПрО графосемантическая модель, состоящая из 24 узлов (дисциплин).

На рисунке 9 представлен скриншот С-карты, генерируемой в ИС «Семограф» по результатам полевого анализа. С-карта представляет собой квадратную матрицу, в строках и столбцах которой располагаются С-поля (дисциплины), а на пересечении строки и столбца – ячейка с показателем встречаемости обоих полей в контекстах корпуса. Например, на пересечении строки LIFE SCIENCES и столбца BUSINESS AND MANAGEMENT находится ячейка со значением 117. Это значение свидетельствует о том, что в корпусе имеется 117 публикаций, маркированных топиками, относящимися и к LIFE SCIENCES, и к BUSINESS AND MANAGEMENT.

С-карта может быть представлена в виде С-графа (см. рис. 10). В С-графе вершины соответствуют С-полям, а ребра – связям между ними. Размер вершины пропорционален частотности С-поля, а толщина ребра – силе связи между полями (частоте совместного присутствия двух С-полей в контекстах). Вершины располагаются произвольно; основной принцип расположения вершин состоит в достижении минимального количества пересечений между ребрами графа, т. е. служит целям понятности чтения графа.

Структура информационного пространства, представленная в С-графе, показывает значимость (вес) и связность научных дисциплин друг с другом, т. е. актуальные междисциплинарные связи. Например, видно, что магистральное направление COMPUTER SCIENCE в большей мере связано с BUSINESS AND MANAGEMENT, т. е. обусловлено созданием систем управления в бизнес-среде. MEDICINE имеет традиционно сильные связи с LIFE SCIENCES, но также и с COMPUTER SCIENCE (актуальные цифровые сетевые медицинские ресурсы и порталы в рамках HEALTH-COMMUNICATION). Поэтому значимость приобретает связь MEDICINE с PUBLIC HEALTH.

Значимые для отечественных исследователей области, такие как PHYSICS, EARTH SCIENCE AND GEOGRAPHY, MATHEMATICS AND STATISTICS, COMPUTER SCIENCE и ENGINEERING, кроме последних двух находятся либо на периферии (PHYSICS, MATHEMATICS AND STATISTICS), либо оказались низкочастотными ($\nu<0.01$), что не позволило их включить в С-граф (EARTH SCIENCE AND GEOGRAPHY).


Project "Information resources" — Meta-analysis

Type of the map: Component analysis

Probabilistic S-map

Sample:

Calculate | Export GML

Precision: 3

Semantic map

	LIFE SCIENCES	EDUCATION AND LANGUAGE	BIOMEDICAL SCIENCES	BUSINESS AND MANAGEMENT	ASTRONOMY	CHEMISTRY	COMPUTER SCIENCE	EARTH SCIENCE AND GEOGRAPHY	ECONOMICS
LIFE SCIENCES	—	29	332	117	0	494	563	22	56
EDUCATION AND LANGUAGE	29	—	0	185	0	0	389	0	18
BIOMEDICAL SCIENCES	332	0	—	1	0	171	268	2	0
BUSINESS AND MANAGEMENT	117	185	1	—	35	32	787	51	164
ASTRONOMY	0	0	0	35	—	1	3	1	42
CHEMISTRY	494	0	171	32	1	—	177	45	14
COMPUTER SCIENCE	563	389	268	787	3	177	—	46	34
EARTH SCIENCE AND GEOGRAPHY	22	0	2	51	1	45	46	—	45
ECONOMICS	56	18	0	164	42	14	34	45	—

Рис. 9. Окно вычисления С-карты (фрагмент)

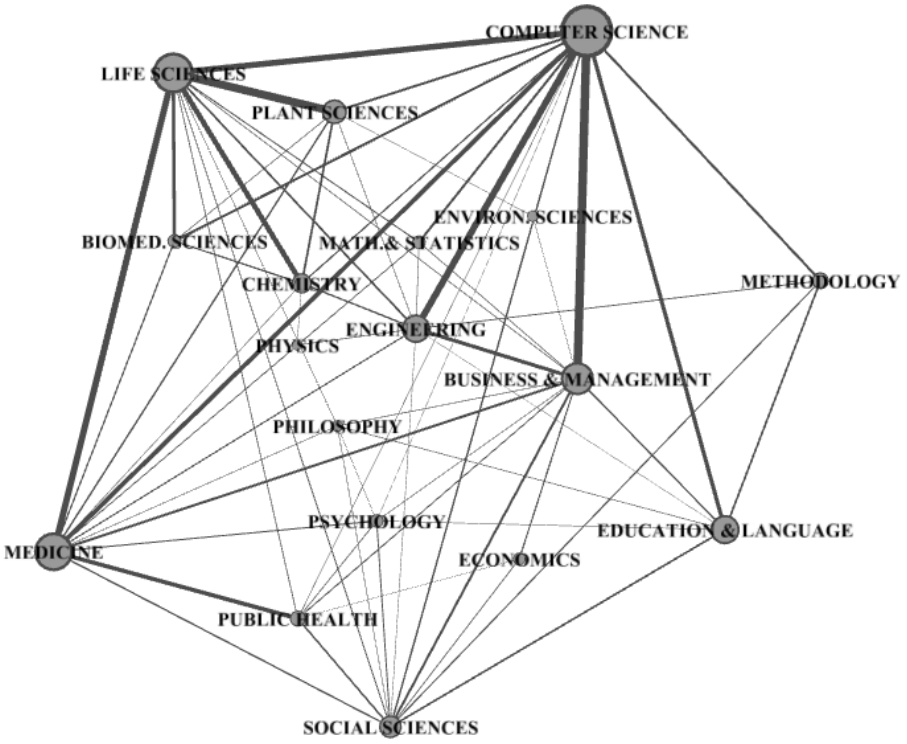


Рис. 10. С-граф дисциплинарных ПрО всех публикаций корпуса (на графе отображены связи, имеющие частоту больше 0.01)

3.4. Композиции С-полей и кластеризация контента

Как отмечалось, С-граф, строящийся на основе исчисления парных связей С-полей, не позволяет передать более сложные композиции С-полей, которые могут наблюдаться в отдельных исследованиях (как уже говорилось выше, актуализация в исследовании от 3 до 9 С-полей). Поэтому представляет интерес разработка такого **метода структурирования информационного пространства**, который учитывал бы композиции С-полей, присутствующие в каждой отдельной публикации.

В качестве метода, направленного на выделение частнонаучных ПрО[5] используется кластерный анализ, осуществляющийся над множеством контекстов, параметрами которых являются бинарные векторы, содержащие наборы полей соответствующих контекстов. Набором полей контекста называется множество С-полей, связанных с данным контекстом посредством произвольного числа терминов (в нашем случае топиков). Поскольку каждому контексту соответствует частнонаучная ПрО, можно использовать результат кластеризации в качестве оценки подобия публикаций и соответствующих им частнонаучных предметных областей.

Выделение частнонаучных ПрО осуществляется с помощью метода нечеткой кластеризации C-means [Kaymak, Setnes 2000: электр. ресурс]. В отличие от других распространенных методов кластерного анализа, таких как K-means или самоорганизующихся карт Кохонена, данный метод допускает принадлежность одного элемента двум и более кластерам, что в данной задаче позволяет отнести одну публикацию к нескольким направлениям в случае высокой неопределенности в ее описании. Кроме того, алгоритм метода C-means допускает возможность влияния на результат с помощью специального числового параметра m ($m \in R$, $m \geq 1$). В качестве исходного значения обычно выбирается $m=2$, затем в случае неудовлетворительного результата выполняется его подстройка. Результаты данного исследования получены с базовым значением параметра $m=2$.

Алгоритм нечеткой кластеризации C-means основывается на минимизации целевой функции $J(x, c, u, m)$:

$$J(x,c,u,m) = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2, m \in R, m > 1,$$

где m – нечуткий параметр (выбирается экспертом), u – степень принадлежности кластеру, $||*||$ – норма, характеризующая близость элементов анализируемого пространства. Процесс оптимизации целевой функции заключается в итератив-

ном пересчете степеней $u_{ij}, i = \overline{1, N}$ и центров кластеров $c_j, j = \overline{1, C}$:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}},$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m}.$$

Условие окончания итерации:

$$\max_{ij}\left(u_{ij}^{(k+1)} - u_{ij}^{(k)}\right) < \varepsilon,$$

где ε – заданный критерий, $\varepsilon > 0$.

В результате кластеризации определяются: 1) частнонаучные ПрО, соотносимые с выделенными кластерами; 2) публикации, в той или иной мере соответствующие частнонаучным ПрО. **Выделенные кластеры (междисциплинарные композиции С-полей) отражают магистральные направления исследований, использующих понятие ИР.**

На основе полученных данных публикации были отсортированы по номеру кластера и расстоянию от его центра в порядке возрастания значений. В таблице 2 представлена кластеризация контекстов в виде 10-тикластерной модели, которая была отобрана из спектра других моделей (состоящих из 6, 7, 8, 9 и 10 кластеров) на основании критерия наибольшего покрытия дисциплин и полученных ранее результатов.

Таблица 2

Частнонаучные предметные области (10-тикластерная модель)

№ кластера	Композиции С-полей	Количество статей	Репрезентативные статьи
1	UNDEFINED SPECIFICATION	1445	[Kolesnikova et al. 2010]
2	BUSINESS AND MANAGEMENT COMPUTER SCIENCE ENGINEERING	668	[Giaglis 2001]
3	ECONOMICS PUBLIC HEALTH SOCIAL SCIENCES	102	[Croucher 2011]
4	MEDICINE PUBLIC HEALTH	398	[Sills et al. 2001: электр. ресурс]
5	BUSINESS AND MANAGEMENT	256	[Griffin et al. 2013]
6	EDUCATION AND LANGUAGE	886	[Holley, Caldwell 2012]
7	LIFE SCIENCES PLANT SCIENCES	1282	[Meskauskiene et al. 2013: электр. ресурс]
8	BUSINESS AND MANAGEMENT ECONOMICS	214	[Gao, Zhang, Liu 2007]
9	BUSINESS AND MANAGEMENT COMPUTER SCIENCE EDUCATION AND LANGUAGE SOCIAL SCIENCES	157	[Rayward, Twidale 1999]
10	COMPUTER SCIENCE MATHEMATICS AND STATISTICS	241	[Dobrev et al. 2007]

В таблице видно, что большинство кластеров имеет междисциплинарную специфику. Отдельно нужно прокомментировать содержание первого кластера, который составили публикации с неопределенной спецификацией. Речь идет о следующем маркировании публикаций топиками: данным публикациям часто присваивается только один топик, который относится к дисциплинам периферийным (с точки зрения их веса в системе С-полей); то же самое наблюдается и в том случае, когда публикацию описывают несколько топиков, относящихся к одной периферийной дисциплинарной ПрО.

3.5. Прогнозирование состояния предметной области на основе имитационного моделирования

Полученное распределение публикаций по кластерам можно рассмотреть во временной динамике с прогнозированием состояния. Для исследования процесса изменения состояния научной предметной области нами была разработана методика на основе имитационного моделирования [Лукьянов, Слесарев 2001]. В качестве имитационной модели используется дискретный марковский процесс первого порядка. Состояниями этого процесса являются графосемантические модели предметной области на заданном шаге. Переход между состояниями осуществляется посредством появления новых работ (добавления новых контекстов в модель). Ниже приведено краткое описание процесса имитации.

Пусть задана графосемантическая модель научной предметной области агента научного производства Ω и модель изменения её состояния в виде марковского процесса первого порядка с матрицей перехода $p_{N \times N}$, где N – количество всех допустимых состояний предметной области. Тогда, имитационное моделирование первых T-переходов может быть осуществлено следующим образом

1. В качестве начального состояние X_0 выбирается заданная модель Ω.

2. Для каждого шага $t = \overline{1, T}$:

(a) выбирается отрезок $x : \left[0; \sum_{i=1}^{N} p_{ti}\right]$;

(b) генерируется («разыгрывается») случайное число r, равномерно распределённое на отрезке x;

(c) в качестве состояния X_t выбирается такое состояние $\Omega^{(j)}$, для которого выполняется условие

$$\sum_{i=1}^{j-1} p_{ti} < r \le \sum_{i=1}^{j} p_{ti}.$$

(d) переход к шагу $t + 1$.

3. Полученная последовательность состояний $X_t, t = \overline{0, T}$ принимается в качестве результата имитационного моделирования.

Вероятности p_{ti} вычисляются на основе модели перехода:

$$P(F_\sigma, \overline{F \setminus F_\sigma}) = \mathrm{P}(F_\sigma) \prod_{i=1}^{n-k} \frac{P(\overline{f_{k+i}}) \prod_{j=1}^{k} P(f_j \mid \overline{f_{k+i}}) \prod_{j=1}^{i-1} P(\overline{f_{k+j}} \mid \overline{f_{k+i}})}{P(f_1, f_2, ..., f_k, \overline{f_{k+1}}, ..., \overline{f_{k+i-1}})},$$

где $P(F_\sigma, \overline{F \setminus F_\sigma})$ – вероятность присутствия в следующем контексте набора полей F_σ при отсутствии других семантических полей, $P(\overline{f_j})$ – вероятность присутствия поля $f_j \in F$ в новом контексте. Вывод данной формулы приведен в [Баранов 2013].

Принцип, используемый на шагах (b),(c) так же называется «принципом рулетки» [Осовский 2002].

Полученная последовательность $X_t, t = \overline{0, T}$ является прогнозом состояния исследуемой предметной области.

Для оценки ошибки прогноза был применён ретроспективный анализ. На основе контекстов за 1965–2014 гг. была построена исходная модель Ω^1_{rfh}, используемая в качестве исходной для прогноза состояния научной ПрО на 2015 г. (рис. 11). На рисунке 11 представлен рост мощности кластеров научных публикаций, использующих понятие ИР, во временной динамике. Часть графика (с 1965 по 2014 г.) построена по известным данным, а значения для 2015 г. получено на основе прогноза состояния модели по описанной методике. На графике видно, что большинство кластеров сохраняет тенденцию роста, однако кластеры 6 и 7 показывают чуть меньшую динамику, по сравнению с последними годами, а кластер 4, согласно прогнозу, будет развиваться значительно медленнее.

График на рисунке 12 построен на основе данных публикаций российских ученых. График показывает, что до 1995 г. публикации российских ученых по данной теме не представлены в журналах Springer.

На диаграмме также видно, что выборка по публикациям российских ученых сохраняет тенденцию быстрого роста 1 кластера, начиная с 2000 г. Это объясняется особенностями данного кластера: он агрегирует большое количество публикаций с отсутствием значительных признаков (например, публикации с ПрО из одного топика). В то же время остальные кластеры демонстрируют гораздо более скромный рост по сравнению с аналогичными кластерами на рисунке 11. Поэтому можно сформулировать рекомендацию редакторам журналов вводить большее количество спецификаций статьи (топиков), относящихся к разным дисциплинарным ПрО. Помимо того что публикации с одним топиком (не входящие в тренды), имеют слабые позиции в системе поиска и фильтрации материала, они теряют всякую спецификацию, что и показывают результаты кластеризации контента. Кроме того, полученные результаты свидетельствуют о необходимости разработки комплекса мер в области стимулирования развития в России отдельных частнонаучных ПрО, успешно развивающихся в мировой науке.

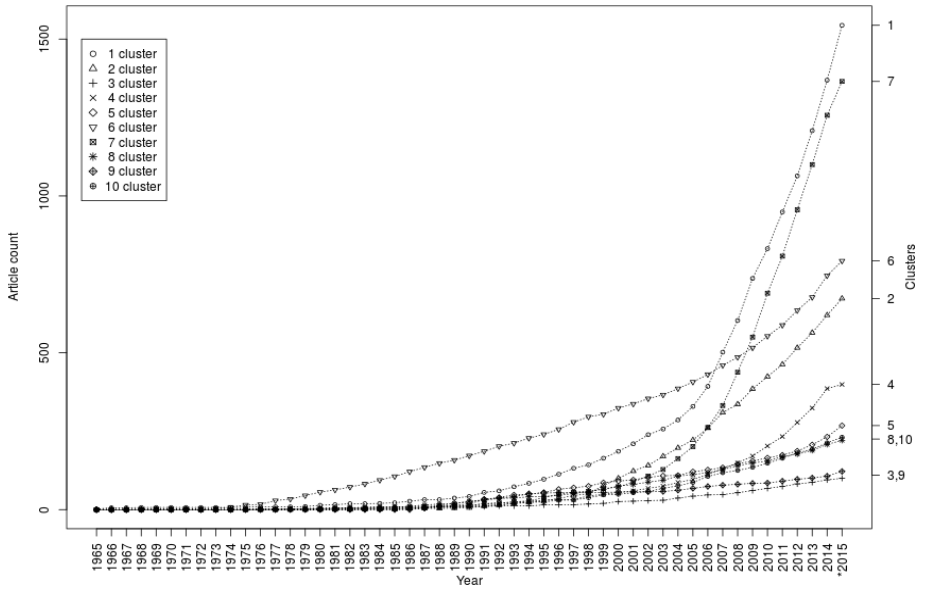


Рис. 11. Изменение мощности кластеров научных публикаций с 1965 г. по 2015 г.
Примечание. Для визуализации данных используется график с накоплении

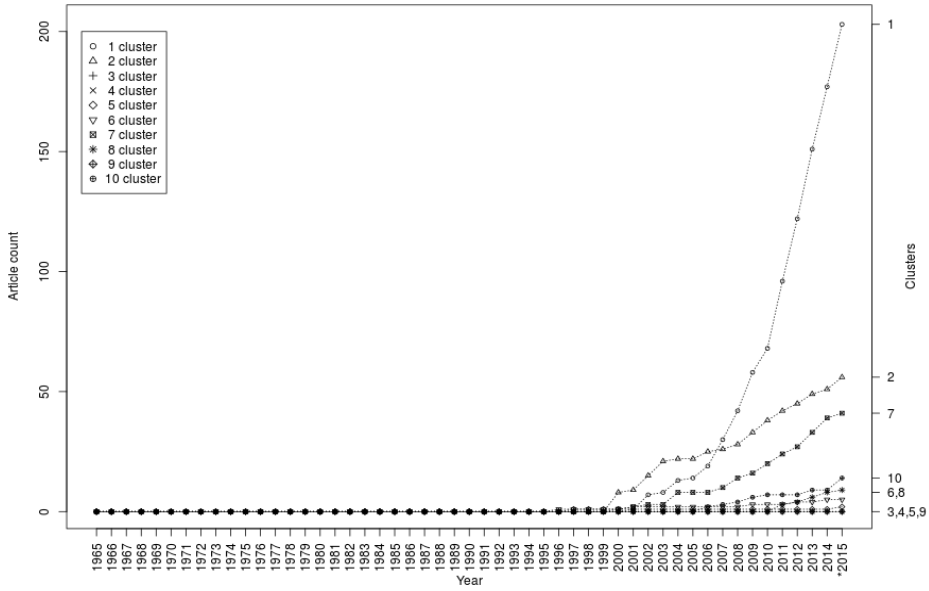


Рис. 12. Изменение мощности кластеров российских научных публикаций с 1965 г. по 2015 г.
Примечание. Для визуализации данных используется график с накоплением.

4. Заключение

Изучение процессов освоения базовых понятий современной науки ее отдельными областями представляет несомненный интерес для всего спектра дисциплин и направлений, поскольку позволяет оценить актуальность и потенциал того или иного теоретического конструкта в конкурентном пространстве научных идей. В данном контексте можно рассматривать и проведенное исследование, позволившее: а) выявить динамику интереса к понятию ИР, которую можно охарактеризовать как крайне актуальную, с большим приростом научного контента (в том числе и в рамках прогнозируемого состояния); б) определить междисциплинарные предметные области, возникающие благодаря обращению к понятию ИР; в) изучить временную и географическую спецификации выделенных предметных областей, их вес в информационном пространстве, создаваемом анализируемым контентом.

Графосемантический подход к исследованию процессов освоения базовых понятий современной науки может использоваться как методика обзора состояния научных исследований в разрабатываемой области, и более широко – служить средством мониторинга состояния частнонаучных предметных областей.

Примечания

[1] Исследование выполнялось при поддержке Российского фонда фундаментальных исследований (проекты № 15-06-06373 и № 14-06-31143).

[2] Под ИР понимается «совокупность данных, организованных для эффективного получения достоверной информации» (ГОСТ 7.0 – 99).

[3] Понятие (или «узел») онтологии ПрО находится в процессе постоянного изменения в силу непрерывно меняющегося научного контекста, создаваемого каждым отдельным исследованием, что приводит к изменению содержания и объема понятия, появлению новых связей между «узлами» (т. е. приращением структуры). Этот процесс концептуализации (см., например [Павиленис 1983]) имеет следствием образование концептосферы над онтологией ПрО. Хорошей иллюстрацией процесса концептуализации понятия служит одновременное функционирование в границах ПрО десятков и сотен определений одного и того же термина (см. [Белоусов, Зелянская 2008]).

[4] Из отношений типа «Метаполя : Метаполя» наибольший интерес представляют данные, соотносящие время и географию публикаций (небольшой фрагмент данной статистики отображен на рис. 3).

[5] Частнонаучная предметная область – представленная в виде математической модели композиция С-полей, репрезентирующая отдельный научный сегмент в общей предметной области [Белоусов, Баранов, Зелянская 2014 и др.].

Список литературы

Баранов Д.А. Вероятностный подход к графосемантическому моделированию // Формирование основных направлений развития современной статистики и эконометрики: материалы I Междунар. научн. конф. Оренбург, 2013. С. 166–175.

Белоусов К.И., Зелянская Н.Л. Моделирование понятийного потенциала термина заглавие // Известия высших учебных заведений. Поволжский регион. Гуманитарные науки. 2008. № 4(8). С. 62–71.

Белоусов К.И. и др. Реализация концептуально-гипертекстовой структуры предметной области в журнале «Вопросы когнитивной лингвистики» / К.И. Белоусов, Д.А. Баранов, Е.В. Ерофеева, Н.Л. Зелянская, Д.А. Ичкинеева // Вопросы когнитивной лингвистики. 2015. № 2. С. 75–88.

Боровиков В. STATISTICA: Искусство анализа данных на компьютере: для профессионалов. СПб.: Питер, 2003. 688 с.

Лукьянов В.С., Слесарев Г.В. Проектирование компьютерных сетей методами имитационного моделирования. Волгоград, 2001. 72 с.

Осовский С. Нейронные сети для обработки информации. М.: Финансы и статистика, 2002. 344 с.

Павиленис Р.И. Проблема смысла: современный логико-философский анализ языка. М.: Мысль, 1983. 286 с.

Belousov K.I. et al. The Forecasting of a Scientific Domain (on the Basis of a Leading Subject Journal) / K.I. Belousov, D.A. Baranov, E.V. Erofeeva, N.L. Zelyanskaya, D.A. Ichkineeva // Automatic Documentation and Mathematical Linguistics. 2014. Vol. 48(5). P. 246–258.

Belousov K.I., Baranov D.A., Zelyanskaya N.L. A research team and its subject area: Towards the question of the effective planning of scientific activities // Scientific and Technical Information Processing. 2014. Vol. 41(2). P. 85–97.

Bryson J. Measuring the Performance of Libraries in the Knowledge Economy and Society // Australian Academic & Research Libraries. 2001. Vol. 32(4). P. 332–342.

Elsevier: Content [Электронный ресурс]. URL: http://www.elsevier.com/online-tools/scopus/content-overview (дата обращения: 09.06.2015).

Croucher S. The Nonchalant Migrants: Americans Living North of the 49th Parallel // Journal of International Migration and Integration = Revue de l'integrationet de la migration international. 2011. Vol. 2(12). P. 113–131.

Dobrev S. et al. Mobile Search for a Black Hole in an Anonymous Ring / S. Dobrev, P. Flocchini, G. Prencipe, N. Santoro // Algorithmica. 2007. Vol. 1(48). P. 67–90.

Gao X., Zhang P., Liu X. Competing with MNEs: developing manufacturing capabilities or innovation capabilities // The Journal of Technology Transfer. 2007. Vol. 1–2(32). P. 87–107.

García A.M.D., Cuello R.O. A model of equitable and sustainable redistribution of knowledge // Educational Technology Research and Development. 2010. Vol. 58(6). P. 781–790.

Giaglis G.M. A Taxonomy of Business Process Modeling and Information Systems Modeling Techniques // International Journal of Flexible Manufacturing Systems. 2001. Vol. 2(13). P. 209–228.

Griffin A. et al. Marketing's roles in innovation in business-to-business firms: status, issues, and research agenda / A. Griffin, B.W. Josephson, G. Lilien, F. Wiersema, B. Bayus, R. Chandy, E. Dahan, S. Gaskin, A. Kohli, C. Miller, R. Oliva, J. Spanjol // Marketing Letters. 2013. Vol. 4(24). P. 323–337.

Gunasekaran A., Khalil O., Mahbubur S. Knowledge and infirmation technology management: human and social perspectives. Idea Group Publishing. 2002. 464 p.

Holley K.A., Caldwell M.L. The Challenges of Designing and Implementing a Doctoral Student Mentoring Program // Innovative Higher Education. 2012. Vol. 3(37). P. 243–253.

Kaymak U., Setnes M. Extended fuzzy clustering algorithms // Erasmus Research Institute of Management. 2000. [Электронный ресурс]. URL: http://repub.eur.nl/pub/57/erimrs20001123094510.pdf (дата обращения: 09.09.2015).

Kogalovsky M.R. Systematization of information resources collections in digital libraries // Programming and Computer Software. 2000. Vol. 3(26). P. 140–155.

Kolesnikova V.M. et al. Soil attribute database of Russia / V.M. Kolesnikova, I.O. Alyabina, L.A. Vorobjeva, E.N. Molchanov, S.A. Shoba, V.A. Rozhkov // Eurasian Soil Science. 2010. Vol. 8(43). P. 839–847.

Lee R. et al. Controls as a Sharable Knowledge Commodity: An Architecture for Open Exchange / R. Lee, K. Dutta, R. Henry, V. Nguyen // Group Decision and Negotiation. 2007. Vol. 16, Iss. 2. P. 143–167.

Mastroianni C., Talia D., Trunfio P. Metadata for Managing Grid Resources in Data Mining Applications // Journal of Grid Computing. 2004. Vol. 2(1). P. 85–102.

Meskauskiene R. et al. Controlled vocabularies for plant anatomical parts optimized for use in data analysis tools and for cross-species studies / R. Meskauskiene, O. Laule, N.V. Ivanov, F. Martin, M. Wyss, W. Gruissem, P. Zimmermann // Plant Methods. 2013. Vol. 9(33). [Электронный ресурс]. URL: http://link.springer.com/article/10.1186/1746-4811-9-33 (дата обращения: 06.10.2015).

O'Leary D.E. Enterprise Knowlwdge Management // Computer. 2002. Vol. 31(3). P. 54–61.

Rayward W.B., Twidale M.B. From Docent to Cyberdocent: Education and Guidancein the Virtual Museum // Archives and Museum Informatics. 1999. Vol. 1(13). P. 23–53.

Senge P. Reflection on "A Leader's New work: Building Learning Organization" // Knowledge management: classic and contemporary works. 2001. P. 53–60.

Sills E.S. et al. Diagnostic and treatment characteristics of polycystic ovary syndrome: descriptive measurements of patient perception and awareness from 657 confidential self-reports / E.S. Sills, M. Perloe, M.J. Tucker, C.R. Kaplan, M.G. Genton, G.L. Schattman // BMC Women's Health. 2001. Vol. 1(3). [Электронный ресурс]. URL: http://link.springer.com/article/10.1186/1472-6874-1-3 (дата обращения: 09.06.2015).

Skyrme D.J. Developing a Knowledge Strategy: From Management to Leadership // Knowledge management: classic and contemporary works. 2001. P. 61–84.

SPACIAL, TEMPORAL AND DISCIPLINARY ASPECTS OF THE NOTION "INFORMATION RESOURCE" EXTENSION

Dmitry A. Baranov
Postgraduate of Computer Security and Sofware Information Systems Department
Orenburg State University

Konstantin I. Belousov
Professor of Theoretical and Applied Linguistics Department
Perm State University

Dilara A. Ichkineeva
Assistant Professor of Foreign Languages for Natural Faculties Department
Bashkir State University

Graph-semantic approach to studying the processes of acquiring contemporary science basic concepts by its separate fields is described. On the material of foreign scientific publications, placed on the platform of the publishing house Springer, the dissemination of the concept INFIRMATION RESOURCES in the disciplinary areas of scientific knowledge is analyzed, taking into account the temporal and spatial (geographical) parameters characterizing scientific publications. One result of the study is a forecasting model of the state of scientific subject areas built with the help of simulation based on Markov process of the first order.

Key words: information resources; scientific subject area; the concept; the corpus of texts; statistical methods; forecasting; simulation; graph-semantic modeling.