УДК 51-77

Афанасьев В.Н., Баранов Д.А., Влацкая И.В., Ичкинеева Д.А.
Оренбургский государственный университет
E-mail: baranov@semograph.com

ПРИМЕНЕНИЕ ГРАФОСЕМАНТИЧЕСКОГО МОДЕЛИРОВАНИЯ ДЛЯ АНАЛИЗА ПРЕДМЕТНЫХ ОБЛАСТЕЙ АГЕНТОВ НАУЧНОГО ПРОИЗВОДСТВА (НА ПРИМЕРЕ ЖУРНАЛА «ВОПРОСЫ ЭКОНОМИКИ»)

В данной работе рассматриваются возможности метода графосемантического моделирования применительно к задаче анализа и оценки предметных областей агентов научного производства. В качестве примера исследуется материал журнала «Вопросы экономики» за 2010-2012 годы (36 номеров). Описаны основные этапы построения графосемантической модели, рассмотрены проблемы, возникающие при моделировании предметных областей агентов научного производства. Показаны некоторые виды анализа, применимые к графосемантической модели.

Ключевые слова: графосемантическое моделирование, моделирование предметных областей, наукометрия, кластерный анализ.

Оценка результативности научной деятельности отдельных учёных, научных коллективов, организаций, является объектом множества исследований. На данный момент наиболее распространены статистические методы оценки по формальным показателям, в частности библиометрическим. К таким методам относятся различные индексы цитирования, импакт-факторы и т.д. Исходными данными для них являются списки цитирований работ отдельных учёных. Сбором и анализом подобных данных занимаются такие международные институты, как Institute for Scientific Information (ISI), International Mathematical Union (IMU), International Council for Industrial and Applied Mathematics (ICAM), а также проекты по созданию реферативных и библиографических баз научных публикаций, среди которых Web of Science и Scopus. Так же можно отметить российский проект «Научная электронная библиотека», реализующий собственный метод оценки «российский индекс научного цитирования» (РИНЦ). Общим подходом данной группы методов является агрегирование различных показателей результативности научной деятельности в единую числовую величину, пригодную для сравнения. Этот подход часто подвергается критике, поскольку в его результате происходит потеря значительной части информации об оцениваемом объекте [1, 7].

Реже применяются методы, основанные на экспертных оценках, так как их применение в глобальных масштабах затруднено необходимостью привлечения авторитетных экспертов и проблемой субъективности мнения экспертов. Следует отметить, что многие статистические подходы так же не являются глобальными, так как статистическая выборка ограничена определённой базой журналов.

Ещё одним подходом к оценке науки являются научные карты. Понятие научной карты применяют к достаточно обширному списку методов агрегирования, анализа и визуализации данных о состоянии различных агентов научного производства. Обычно подобные карты отображают взаимосвязи между различными элементами моделируемой системы, например между научными направлениями. Наиболее распространённым методом построения научных карт является совместное цитирование (коцитирования, co-citation), предложенный ещё в 70-х годах XX века. Ключевой идеей данного метода является кластерный анализ публикаций на основе их совместного цитирования последующими работами [4, 10]. Карты, построенные на основе данного метода могут использоваться для оценки состояния научных направлений и выявления связей между ними. Так же можно выделить карты научных компетенций, позволяющие оценить роль отдельных агентов научного производства разных уровней, и карты научных парадигм. Существуют и другие типы научных карт. Наиболее крупным и успешным проектом построения глобальной научной карты, является Map Of Science от SciTech Strategies.

Предлагаемый в данной работе подход можно рассматривать как тип научной карты, однако принцип её построения значительно отличается от большинства существующих методов. В его основе лежит метод графосемантического моделирования Описание метода гра-

фосемантического моделирования и его приложения были опубликованы в работах [2, 3, 5]. Применение данного метода позволяет перейти от отдельных публикаций к анализу структуры предметной области моделируемых объектов. Предлагаемый подход может быть применён для оценки, мониторинга и прогнозирования результативности любого агента научного производства (учёного, научного коллектива, организации, журнала и т.д.).

Методику графосемантического моделирования предметных областей агентов научного производства можно разделить на несколько этапов. На первом этапе производится сбор исходных данных, описывающих результаты научной деятельности. Исходными данными, в данной задаче являются объекты публикационной активности (точнее, их библиографические данные): статьи, тезисы, монографии и т.д. Такой выбор исходных данных обусловлен их высокой доступностью, применимостью к различным агентам научного производства и высокой доступностью, по сравнению с данными анкетирования или экспертными оценками.

На втором этапе строится графосемантическая модель предметной области, описываемой исходными данными. Для этого каждому объекту публикационной активности ставится в соответствие контекст и заполняются его метаданные (название публикации, год опубликования, автор, организация, журнал и т.д.). Метаданные имеют большое значение, т.к. они позволяют детализировать графосемантическую модель и строить срезы. Например, с помощью метаданных можно выделить предметную область отдельного автора или журнала. Далее, в каждом объекте публикационной активности выделяются ключевые слова и добавляются в соответствующие контексты в качестве семантических компонентов. Результатом данного этапа является частично сформированная графосемантическая модель, содержащая отдельные контексты, не связанные между собой. Такая модель может быть использована для решения некоторых задач, например для оценки близости публикаций на основе представленных в них ключевых слов. Однако, из-за большого разнообразия ключевых слов, результативность данного подхода незначительна. Кроме того, при наличии перекрёстных ссылок в метаданных (извлечённых из списков цитирований), данная модель может использоваться для построения карт совместных цитирований публикаций.

Для выполнения третьего этапа требуется привлечение экспертов исследуемой научной отрасли. В задачи экспертов входит определении списка семантических полей, описывающих научную отрасль (например: «микроэкономика»), и установление связей этих полей с доступными семантическими компонентами (извлечёнными на втором этапе ключевыми словами). Отметим, что связи между полями и компонентами относятся к типу «многие ко многим». Количество семантических полей зависит от выбранного уровня детализации модели. Модель может включать несколько уровней детализации (иерархическая графосемантическая модель), при этом поля нижестоящих уровней включаются в вышестоящие поля, образуя иерархию. Такой подход позволяет точнее описывать моделируемые предметные области и производить анализ модели на любом из доступных уровней детализации. Полученная на третьем этапе графосемантическая модель является завершенной и может быть использована для построения семантических карт, графов и полевого анализа. Кроме того, модель может дополняться новыми контекстами и допускает изменение семантических полей и их связей с компонентами.

Основные возможности анализа представлены на примере графосемантической модели предметной области журнала «Вопросы экономики». «Вопросы экономики» является ведущим в России теоретическим и научно-практическим журналом общеэкономического содержания. Данный журнал находится в списке наиболее рейтинговых журналов РФ. С 2007 г. «Вопросы экономики» был включен в список российских научных журналов ВАК Минобрнауки России. Импакт-фактор издания – 3,831, журнал включен в международные базы цитирования. По мнению ряда ведущих экономистов РФ, журнал «Вопросы экономики» играет системообразующую роль в российской экономике. «Вопросы экономики» называют журналом академическим и специализированным, на протяжении многих лет поддерживающим высокие стандарты качества (Г. Фетисов, М. Ершов и др.). Аудиторией журнала являются экономисты-исследователи, преподаватели и студенты вузов, руководители

федеральных и региональных органов власти, отвечающие за разработку экономической политики, аналитические подразделения крупных предприятий, корпораций и банков.

В качестве исходных данных в данной работе использовались библиографически данные о научных статьях, опубликованных в журнале «Вопросы экономики» в 2010–2012 гг. При этом были выделены следующие типы метаданных:

1. Название – название контекста.
2. Аннотация – тело контекста.
3. Автор – список авторов публикации, включающий их ФИО и место работы (организацию).
4. Год – год опубликования работы.
5. Номер – номер журнала.
6. Ключевые слова – ключевые слова, описывающие публикацию с точки зрения её авторов.

В сформированную графосемантическую модель были включены 308 контекстов и 933 уникальных семантических компонента (ключевых слов). В качестве эксперта выступил один из авторов данной статьи, заведующий кафедрой статистики и эконометрики Оренбургского государственного университета, доктор экономических наук, профессор В. Н. Афанасьев. В данном исследовании изучался верхний уровень детализации предметной области, охватываемой журналом «Вопросы экономики», поэтому экспертом были выделены 7 семантических полей этого уровня:

1. Микроэкономика (МИКРО);
2. Международная экономика (МЭ);
3. Экономическое положение (ЭП);
4. Качество жизни населения (КЖН);
5. Методология исследований социально-экономических процессов (МИСЭП);
6. Институциональное состояние (ИС);
7. Инновации. Научно-технический аспект экономики (ИНТА).

Подробно изучить сформированную модель, включая связи семантических полей и компонентов, можно в системе графосемантического моделирования «Семограф», где была выполнена данная работа [9, 8].

Ключевыми особенностями метода графосемантического моделирования являются семантическая карта и семантический граф. Семантическая карта для полученной модели приведена в таблице 1, а семантический граф на рисунке 1. В таблице 1 строки и столбцы соответствуют семантическим полям, а ячейки на их пересечении – числу совместных появлений в контекстах. В графе вершины соответствуют семантическим полям (цифра внутри вершины соответствует номеру семантического поля), а рёбра – связям между ними. Размер вершины пропорционален частотности поля, а толщина ребра – силе связи между полями.

Наиболее простым типом полевого анализа графосемантической модели является получение её статистических характеристик. Первой рассматриваемой характеристикой является распределения частотностей семантических полей для 2010 – 2012 годов, изображённое на рисунке 2. В данном исследовании, это распределение показывает трёхгодичную динамику полей. Из графика, приведённого на рисунке 2, можно сделать вывод об относительной стабильности предметной области на данном уровне детализации модели. Лишь поле «Качество жизни населения» показывает постоянный рост.

Таблица 1. Семантическая карта

	1	2	3	4	5	6	7
1	–	51	58	30	33	18	60
2	51	–	158	111	108	62	181
3	58	158	–	101	105	61	182
4	30	111	101	–	69	35	117
5	33	108	105	69	–	48	119
6	18	62	61	35	48	–	75
7	60	181	182	117	119	75	–

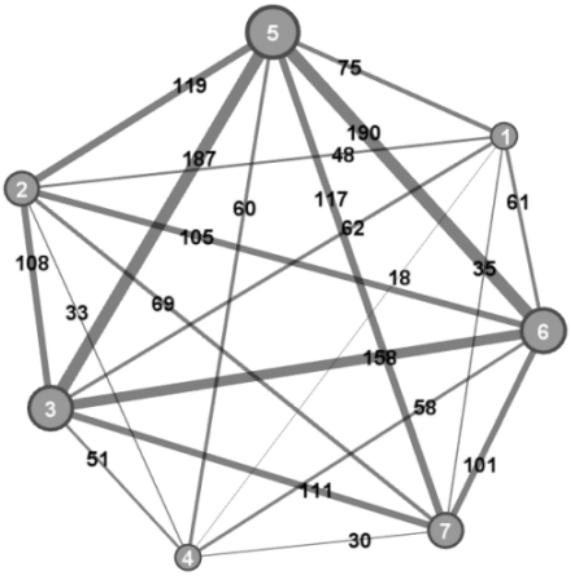

Рисунок 1. Семантический граф

Фактически, статистика по отдельным семантическим полям помогает обобщить информацию о ключевых словах, однако не позволяет оценить предметную область в целом. Для этого необходим переход к частным предметным областям, составляющим общую. Такие предметные области строятся для отдельных статей и включают их семантические поля. Другими словами, частная предметная область является подмножеством множества семантических полей (т. н. «набором полей» [5]). Таким образом, возможно существование 2^n различных частных предметных областей, где n – число семантических полей. В данном исследовании n=7, следовательно, возможно существование 128 различных частных предметных областей.

Предметные области могут быть перенумерованы следующим образом: частная предметная область представляется в виде бинарного вектора *vf* размерности n, где каждому компоненту вектора соответствует семантическое поле и элемент равен 1, если соответствующее поле представлено в статье и 0 в противном случае. Далее, данный вектор может быть представлен как запись целого числа в двоичной системе счисления и соответствующая запись этого числа в десятичной системе счисления присваевается в качестве номера частной предметной области. Очевидно, можно выполнить обратное преобразование переводом номера частной предметной области в двоичную систему счисления. На рисунке 3 представлено распределение частотностей различных частных предметных областей по статьям.

На основе рисунка 3 были отмечены пики, соответствующие частным предметным областям под номерами 18, 50, 83, 90. Эти частные предметные области состоят из следующих наборов семантических полей:

• «Институциональное состояние», «Методология исследований социально-экономических процессов» (18);

• «Институциональное состояние», «Методология исследований социально-экономических процессов», «Микроэкономика» (53);

• «Инновации. Научно-технический аспект экономики», «Институциональное состояние», «Методология исследований социально-экономических процессов», «Экономическое положение» (83);

• «Институциональное состояние», «Международная экономика», «Методология исследований социально-экономических процессов», «Экономическое положение» (90).

Очевидно, описанный способ нумерации частных предметных областей не позволяет судить о их подобии или связи. Для этого необходим другой подход, учитывающий состав предметных областей. В качестве такого подхода в данном исследовании используется нечёткий

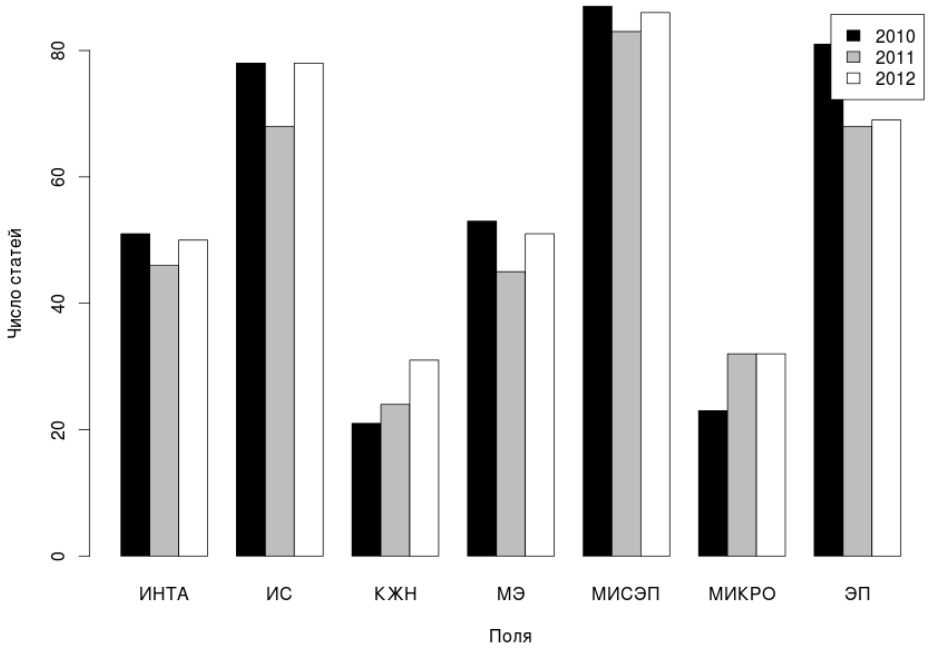

Рисунок 2. Распределение полей по статьям и годам

кластерный анализ по алгоритму C-means [6]. Кластерный анализ производился над множеством контекстов, параметрами были бинарные вектора *vf*, содержащие наборы полей соответствующих контекстов. Поскольку каждому контексту соответствует частная предметная область, можно использовать результат в качестве оценки подобия публикаций и соответствующих им частных предметных областей. Ниже приведено подробное описание алгоритма C-means.

Алгоритм нечёткой 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 – нечёткий параметр, u – степень принадлежности кластеру, $\|*\|$ – норма, характеризующая близость элементов анализируемого пространства. Процесс оптимизации целевой функции заключается в итеративном пересчёте степеней $u_{ij}, i = \overline{1,N}$ и центров кластеров $c_j, j = \overline{1,C}$:

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

Условие окончания итерации:

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

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

Число кластеров было выбрано равным 4, параметр m=2. Исходными центрами кластеров были выбраны описанные выше пики распределения. В результате были получены кластеры, описанные в таблице 2. Следует отметить, что в описании каждого кластера представлена одна репрезентативная статья, однако их может быть несколько (находящихся на одном расстоянии от центра кластера), но их частные предметные области совпадают. Фактически, кластеры описывают основные научные направления в данной отрасли, которые можно выделить на выбранном уровне детализации модели.

На основе полученных результатов, частные предметные области были отсортированы по номеру кластера и расстоянию от его центра и перенумерованы в порядке возрастания. Результаты представлены на графике, изображённом на рисунке 4. На данном графике координатные оси соответствуют частным предметным областям и номерам журнала «Вопросы экономики» соответственно. Точка ставится в случае, если по данным координатам существует публикация (т.е. работа, опубликованная в данном номере, охватывает данную предметную об-

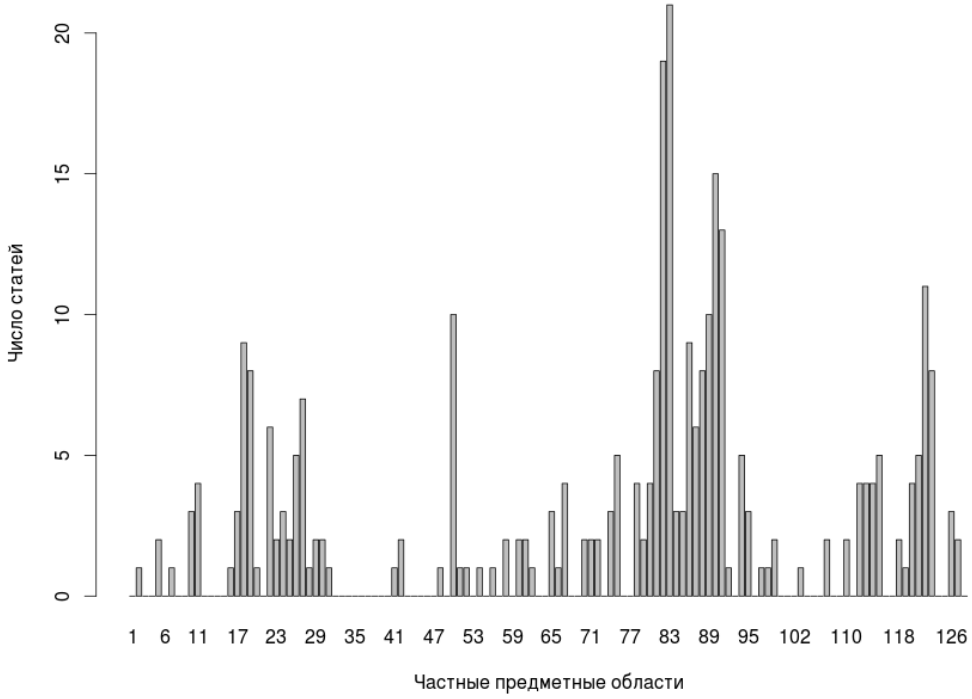

Рисунок 3. Распределение частотностей предметных областей

ласть). Плотность окрестности точки зависит от числа публикаций в данной точке. На горизонтальной оси указаны только номера кластеров, а на вертикальной оси только года. Можно отметить относительно равномерное заполнение всех кластеров, хотя можно выделить отдельные свободные и занятые направления. Помимо графика, отражающего изменение интереса к различным предметным областям с течением времени, на основе рассматриваемой методики могут быть острены графики, показывающие области, разрабатываемые отдельными учёными или организациями. На рисунке 5 изображено распределение статей по предметным областям и авторам. Так же, как и на графике 4, на горизонтальной оси расположены научные направления (пронумерованные кластеры и отдельные предметные области внутри них). На вертикальной оси отображены первые 30 авторов журнала «Вопросы экономики» по числу статей в данном журнале за рассматриваемый период.

Результаты, представленные на графиках 4 и 5 являются наглядным средством оценки состояния научной отрасли, согласно материалам журнала «Вопросы экономики» за 2010-2012 годы. Однако, такие графики не подходят для прогностических целей и планирования. Для этих целей, на основе рассмотренной модели, был построен график динамики выделенных научных направлений (кластеров), представленный на рисунке 6. На данном графике представлены кумулятивные суммы числа публикаций (вертикальная ось) в выделенных направлениях (кластерах) по всем 36 номерам 2010-2012 годов (горизонтальная ось).

Из графика 6 видно, что наиболее активно развивающимся направлением является направление 3 («Экономическое положение на макро– и микроэкономических уровнях и инновации»). Кроме того, направление 1 («Методология исследования институционального состояния») разрабатывалось практически такими же темпами, как направление 3, до второго квартала 2012 года. Можно отметить, что кумулятивные кривые динамики развития направлений 2 («Исследование микроэкономических процессов») и 4 практически параллельны, при этом направление 4 («Институциональное состояние в международной экономике») периодически делает незначительные ускорения. Так же отметим, что график 6 (либо аналогичные ему) может быть непосредственно использован при планировании научной деятельности, поскольку наглядно отражает состояние основных научных направлений на рассматриваемом уровне.

Предложенная методика позволяет проводить мониторинг, анализ и оценку актуального состояния предметных областей агентов научного производства. Результаты графосемантического моделирования позволяют решать и более сложные задачи, такие как прогнозиро-

Таблица 2. Описание кластеров частных предметных областей (направлений)

№	Набор полей	Статей	Наиболее репрезентативная статья	Название кластера
1	Институциональное состояние; Методология исследований социально-экономических процессов	103	«Возникновение институтов: методолого-индивидуалистический подход»	Методология исследования институционального состояния
2	Институциональное состояние; Методология исследований социально-экономических процессов; Микроэкономика	36	«Деньги и рынок»	Исследование микроэкономических процессов
3	Инновации. Научно-технический аспект экономики; Институциональное состояние; Методология исследований социально-экономических процессов; Экономическое положение	111	«В поисках новой методологии: сравнительный и исторический институциональный анализ Авнера Грейфа»	Экономическо положение на макро- и микроэкономических уровнях и инновации
4	Институциональное состояние; Международная экономика; Методология исследований социально-экономических процессов; Экономическое положение	58	«В защиту евро: подход австрийской школы (критика ошибок ЕЦБ и интервенционизма Брюсселя)»	Институциональное состояние в международной экономике

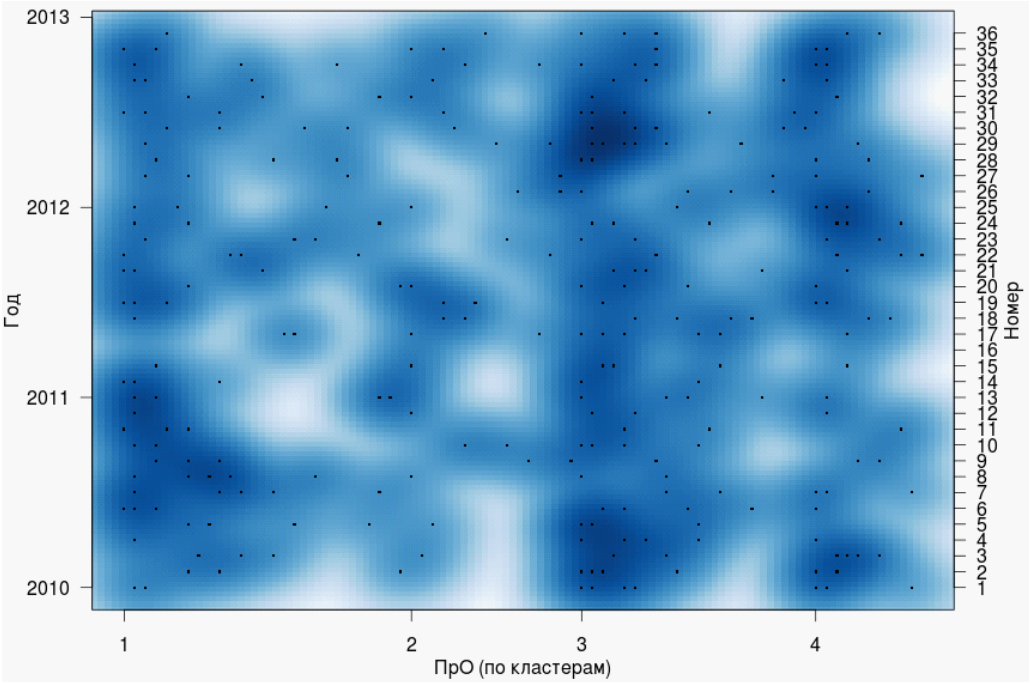

Рисунок 4. Статьи по ПрО и номерам

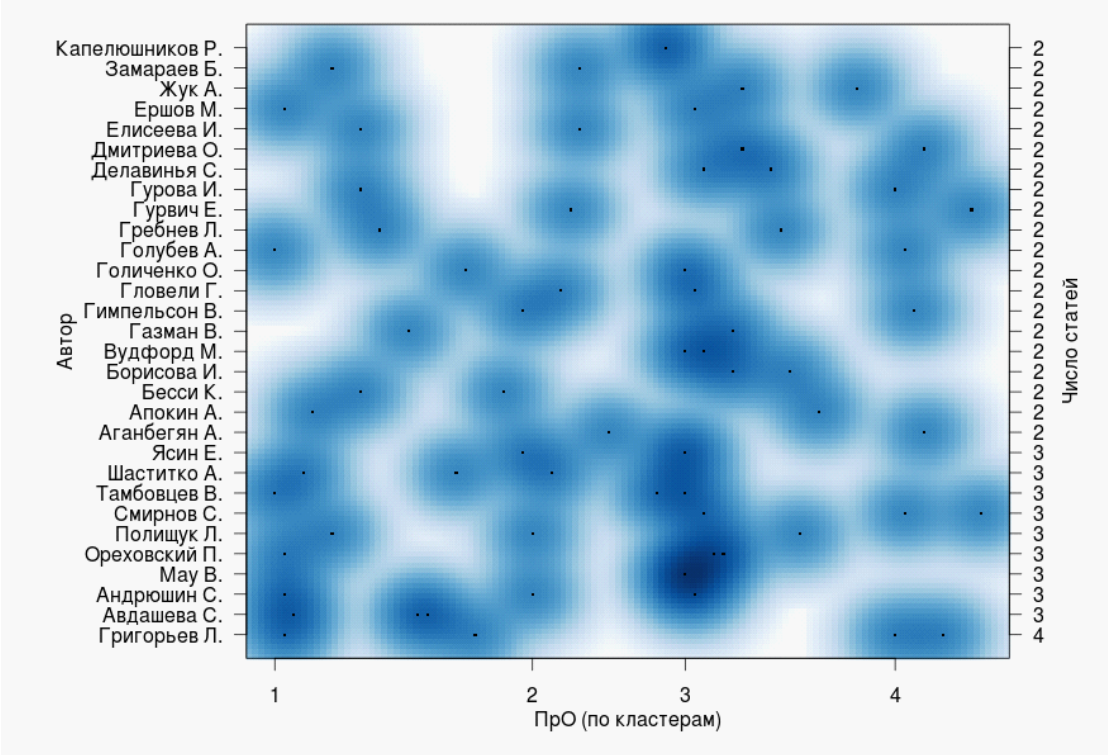

Рисунок 5. Статьи по ПрО и авторам

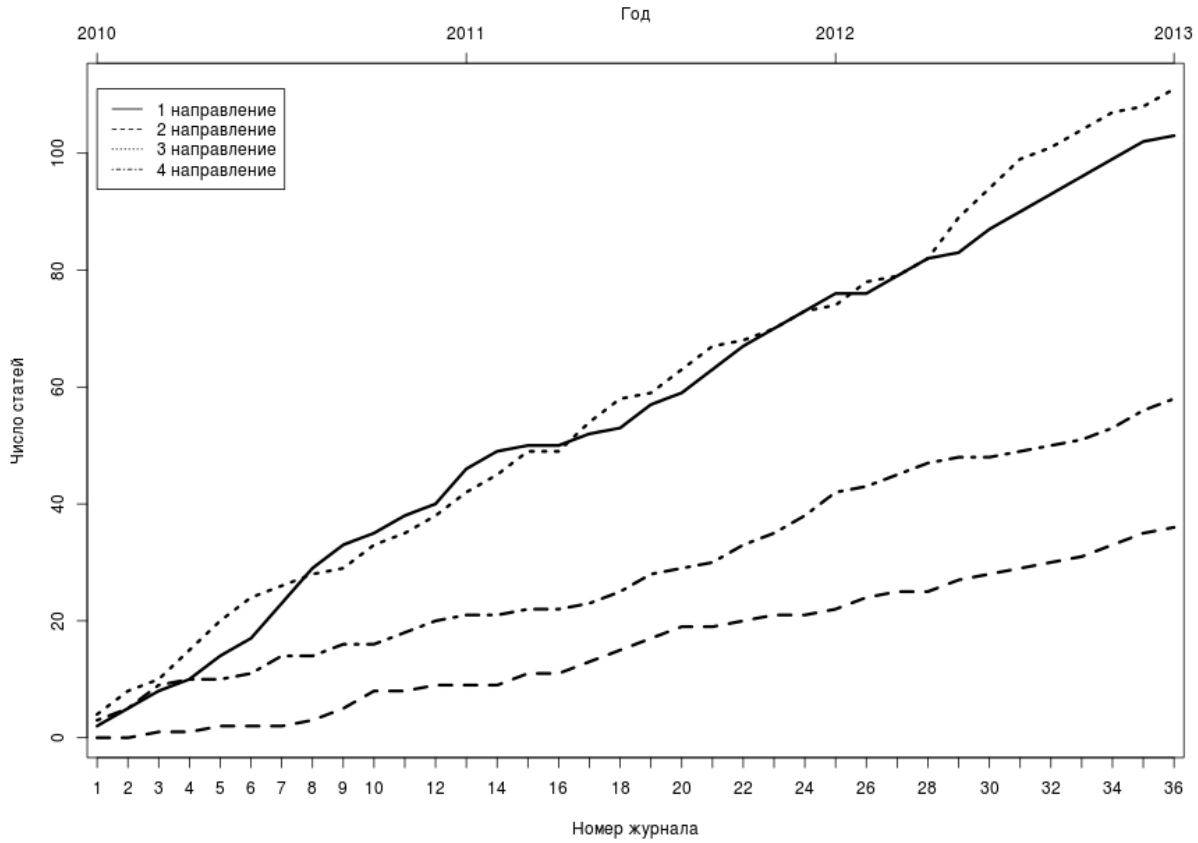

Рисунок 6. Динамика изменения научных направлений

вание, планирование и принятие решений. Перспективы предложенной методики могут быть значительно расширены за счёт использования большего объёма исходных данных (журналы различных направлений, сборники материалов конференций, патентные базы данных и т.д.).

5.07.2013

Исследование выполнялось в рамках государственного задания ОГУ № 8.2714.2011 и при финансовой поддержке Российского гуманитарного научного фонда (проект № 12034в)

Список литературы:

1. Citation Statistics : Rep. / IMU, ICIAM, IMS ; Executor: Robert Adler, John Ewing, Peter Taylor : 2008. — P. 28.
2. Lawrence Peter. Lost in publication: how measurement harms science // Ethics in Science and Environmental Politics. — 2008. — no. 8.
3. Boyack Kevin W, Klavans Richard. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? — 2010. — Vol. 61, no. 12. — P. 2389–2404.
4. Small H., Upham S. P. Co-citation structure of an emerging research area on the verge of application // Scientometrics. — 2009. — Vol. 79, no. 2. — P. 15–38.
5. Белоусов К. И. Теория и методология полиструктурного синтеза текста. — М. : Флинта, 2009. — С. 216.
6. Белоусов К. И., Зелянская Н. Л., Баранов Д. А. Концептуально гипертекстовая модель управления контентом в ИС «Семограф» // Вестник ОГУ. — 2012. — Т. 11, No 147. — С. 56–61.
7. Математическая формализация метода графосемантического моделирования : Техника и технология: новые перспективы развития / Материалы VIII Международной научно-практической конференции ; исполн.: Д. А. Баранов. — М. : 2013. — С. 70–78.
8. Семограф [Электронный ресурс]: Система графосемантического моделирования. — режим доступа: http://new.semograf.com.
9. Система графосемантического моделирования / Д. А. Баранов, К. И. Белоусов, И. В. Влацкая, Н. Л. Зелянская. — М. : Свидетельство о государственной регистрации в Федеральной службе по интеллектуальной собственности, патентам и товарным знакам. Зарегистрировано в Реестре программ для ЭВМ No 20111617192 от 15.09.2011.
10. Kaymak U., Setnes M. Extended fuzzy clustering algorithms // Erasmus research institute of management. — 2000. — P. 24.

Сведения об авторах:

Афанасьев Владимир Николаевич, заведующий кафедрой статистики и эконометрики ренбургского государственного университета, доктор экономических наук, профессор, e-mail: afanassiev@rambler.ru
Баранов Дмитрий Александрович, аспирант кафедры математического обеспечения информационных систем Оренбургского государственного университета, e-mail: baranov@semograph.com
Влацкая Ирина Валерьевна, заведующая кафедрой математического обеспечения информационных систем Оренбургского государственного университета, кандидат технических наук, доцент, e-mail: irina.vlatskaya@yandex.ru
Ичкинеева Дилара Ахметовна, старший преподаватель кафедры английской филологии и методики преподавания английского языка Оренбургского государственного университета, кандидат филологических наук, e-mail: dilaraichkineeva@gmail.com
460018, г. Оренбург, пр-т Победы 13, ауд. 20521, тел. (3532) 372534, e-mail: mois@mail.osu.ru

UDC 51-77
Afanasyev V. N., Baranov D. A., Vlatskaya I. V., Ichkineeva D. A.
Orenburg state university, e-mail: mois@mail.osu.ru

APPLICATION OF GRAFOSEMANTIC MODELING FOR THE ANALYSIS OF SUBJECT DOMAINS OF SCIENTIFIC PRODUCTION AGENTS (ON THE EXAMPLE OF THE JOURNAL "VOPROSY ECONOMIKI")

The paper deals with the possibilities of method of grafosemantic modeling in relation to a problem of the analysis and assessment of subject domains of scientific production agents. As an example the articles from journal «Voprosy Economiki» for 2010-2012 (36 issues) are investigated. The main steps of grafosemantic model construction are described, the problems arising at modeling of subject domains of scientific production agents are considered. Some types of the analysis applicable to grafosemantic model are shown.

Key words: grafosemantic modeling, modeling of subject domains, scientometrics, cluster analysis.

Bibliography:

1. Citation Statistics: Rep. / IMU, ICIAM, IMS ; Executor: Robert Adler, John Ewing, Peter Taylor : 2008. – P. 28.
2. Lawrence Peter. Lost in publication: how measurement harms science // Ethics in Science and Environmental Politics. – 2008. – no. 8.
3. Boyack Kevin W, Klavans Richard. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? – 2010. – Vol. 61, no. 12. – P. 2389–2404.
4. Small H., Upham S. P. Co-citation structure of an emerging research area on the verge of application // Scientometrics. – 2009. – Vol. 79, no. 2. – P. 15–38.
5. Belousov K.I. Theory and methodology of polystructural text synthesis. – M.: Flint, 2009. – P. 216.
6. Belousov K.I., Zelyanskaya N. L., Baranov D. A. Conceptual model of hypertext content management in «Semograf» IS / / Vestnik OSU. – 2012. – T. 11, No 147. – P. 56-61.
7. The mathematical formalization of graphosemantic modeling method: Techniques and Technology: new prospects / Proceedings of the VIII International Scientific Conference; executed.: Baranov D. A. – M.: 2013. – P. 70-78.
8. Semograf. – URL: http://new.semograf.com (date accessed: 04.08.2013).
9. Graphosemantic modeling system / D. A. Baranov, K. I. Belousov, I. V. Vlatskaya, N. L. Zelyanskaya. – M.: State Registration Certificate of the Federal Service for Intellectual Property, Patents and Trademarks. Registered in the Registry of the computer programs No 20111617192 from 15.09.2011.
10. Kaymak U., Setnes M. Extended fuzzy clustering algorithms / / Erasmus research institute of management. – 2000. – P. 24.