КОМПЬЮТЕРНЫЙ АНАЛИЗ СЛОВАРНЫХ СЕТЕЙ

К. А. Рудниченко, А. И. Куликов*, И. И. Титов**

Институт вычислительной математики и математической геофизики СО РАН, 630090, Новосибирск, Россия
*Новосибирский государственный университет, 630090, Новосибирск, Россия
**Институт цитологии и генетики СО РАН, 630090, Новосибирск, Россия

УДК 622.276.1

Для исследования структурно-функциональной организации словарных сетей проведен анализ ключевых статистических характеристик словарных сетей литературных текстов и словарных сетей, построенных с использованием баз публикаций медико-биологической тематики. Разработаны и программно реализованы быстрые алгоритмы расчета одно-, двух- и трехвершинных характеристик графов.

Ключевые слова: анализ текстов, словарные сети, структура стохастических сетей, безмасштабные сети.

For investigation of the structural-functional organization of word networks the key statistical characteristics of the word networks of literary texts and the word networks constructed on the base of scientific publications of medicine and biology were analyzed. The fast algorithms calculating singe-, pair- and triple-vertex graph characteristics were developed and realized.

Key words: text mining, word networks, stochastic network structure, scale-free networks.

Введение. В настоящее время число научных публикаций по биологии и медицине, содержащихся только в базе PubMed, составило почти 20 млн. Такое огромное количество публикаций привело к необходимости автоматического анализа информации, содержащейся в базах данных. Созданы многочисленные компьютерные средства извлечения, категоризации, кластеризации и суммирования информации из текстов [1], различными способами отображающие организацию науки в сложную иерархическую структуру.

Однако в последнее десятилетие стало ясно, что сложные динамические системы обладают рядом универсальных свойств, обусловленных особенностями их структуры. Исследования большого числа природных и искусственных сетей (биологических, социальных, технологических, информационных, словарных и программных приложений) показали, что они обладают одинаковой архитектурой и как следствие одинаковыми динамическими свойствами [2]. Структура этих сетей оказалась неоднородной, т. е. узлы значительно различаются по числу входящих в них связей. При этом распределение узлов по связности спадает по степенному закону, в отличие от экспоненциального падения для случайных графов. Столь медленное спадание функции распределения по связности свидетельствует о большом числе вершин со значительным числом связей. Такие вершины являются ключевыми элементами архитектуры сети, образуя "хабы" и определяя ее помехоустойчивость, скорость и синхронность передачи информации. Таким образом, структурные особенности определяют характер разнообразных свойств сетей. Исследование структуры сетей с целью выявления общих статистических закономерностей оказывается полезным для понимания организации, динамики и происхождения систем, моделью которых являются изучаемые сети.

Работа выполнена в рамках Интеграционных проектов СО РАН № 111, 113.

1. Стохастические сети и их основные статистические характеристики. Одними из первых графы со случайными связями между вершинами исследовали П. Эрдеш и А. Реньи в работе [3]. Установлено, что при постепенном добавлении связей вершины графа объединяются в компоненты — группы вершин, в которых любая пара вершин связана путем по графу. Когда средняя связность превышает определенное критическое значение, в нем образуется гигантская компонента, содержащая фиксированную долю вершин всего графа. По сравнению с размером этой компоненты остальные компоненты в большом графе исчезающе малы, и, таким образом, свойства гигантской компоненты определяют характеристики всего графа.

При случайно-равномерном соединении вершин графа распределение $P(k)$ (k — связность, т. е. число входящих в вершины ребер) является биномиальным, а в пределе большого размера графа — пуассоновским. Такие графы соответственно называются пуассоновскими. Другие классы стохастических графов получаются с использованием иных способов случайного размещения ребер, каждый из которых задает определенную архитектуру сети и функциональные характеристики $P(k)$. Первый момент $P(k)$ представляет собой среднюю связность в графе $< k >$. Для графа с конечным числом вершин N и ребер K средняя связность вычисляется по формуле $\langle k \rangle = 2K/N$.

При анализе свойств сетей обычно используются следующие статистические характеристики. Среднее расстояние между вершинами, или диаметр сети, L вычисляется усреднением по всем парам минимального числа посещаемых на пути вершин. При этом в случайных сетях величина $L \propto \ln N$ значительно меньше соответствующего значения в регулярных сетях: $L \propto N$. Коэффициент кластеризации C задается вероятностью пары соседних вершин иметь общего соседа, т. е. относительной частотой циклов длиной 3. В пуассоновских сетях связность любой вершины не зависит от связности других вершин, поэтому кластеризация исчезающе мала: $C \propto 1/N$. По той же причине в пуассоновских сетях мал коэффициент корреляции связности вершин r.

В конце XX в. было обнаружено, что в сети Интернет распределение числа связей вершин k описывается степенным, а не экспоненциальным, как в пуассоновских сетях, законом [2]. При этом вероятность того, что случайная вершина связана с k другими, пропорциональна $1/k^n$. По аналогии с физикой сети со степенным распределением были названы безмасштабными, поскольку в такой сети отсутствует характерное значение связности. Объекты, распределенные по степенному закону, нередко устроены иерархически. В основном это сильнофлуктуирующие системы вблизи точки фазового перехода, обладающие свойством масштабной инвариантности, т. е. свойством сохранения внешнего сходства при изменении масштаба рассмотрения. Исследования последних лет показали, что большинство сетей в живой и неживой природе (информационные, экологические, генные, метаболические, социальные, технологические, словарные, программных приложений и др.) отличаются от пуассоновских сетей и обладают безмасштабной архитектурой.

Как и пуассоновские, безмасштабные сети имеют малый диаметр: $L \propto \ln N$. Впервые это было выявлено С. Мильграмом при определении среднего числа знакомых посредников между двумя незнакомыми людьми [4]. Популярное утверждение “все люди знакомы друг с другом через шесть рукопожатий” возникло именно отсюда. Направленный и ненаправленный графы одинаковой топологии оказываются близкими по значению диаметра [5].

Существенным свойством, отличающим исследованные безмасштабные сети от пуассоновских, является большое значение коэффициента кластеризации C. Кластеризованная сеть с малым диаметром представляет собой “малый мир”, в котором вершины соединены

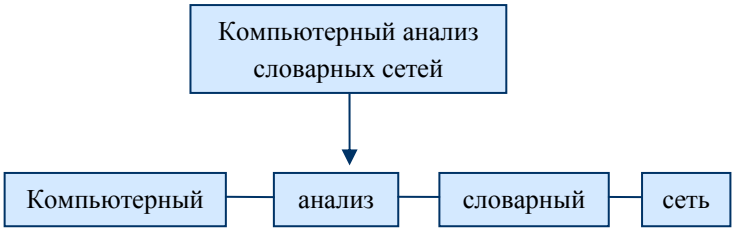

Рис. 1. Схема построения словарной сети слов-соседей в тексте

короткими путями. В подобной сети можно выделить ядро, состоящее из ключевых вершин (концентраторов), соединенных большим числом связей друг с другом и с остальной, слабоструктурированной частью. Кратчайший путь через сеть обычно проходит через один из концентраторов. Количество узлов-концентраторов у безмасштабной сети очень мало по сравнению с числом всех вершин, поэтому безмасштабные сети сравнительно легко переносят случайные потери вершин, но могут серьезно пострадать от атак, направленных на само ядро. Серьезное повреждение всей сети может наступить даже при потере небольшого числа концентраторов. Таким образом, защита ядра обеспечивает эффективное сохранение свойств сети.

2. Определение словарных сетей. Преобразование линейного текста в многомерный объект (граф) может оказаться полезным при анализе текстов, отражая сложность структуры человеческого языка, восприятия и мышления людей. На каждом уровне организации (семантическом, синтаксическом, морфологическом, фонетическом и др.) человеческий язык составлен из элементов различного типа. Даже при использовании слов в качестве элементарных языковых модулей богатство форм взаимодействия слов в человеческом языке определяет разнообразие типов возможных сетей. Функции психолингвистической и грамматической архитектуры, построения предложений выполняют соответственно сети семантические, синтаксические и одновременной встречаемости слов в тексте [6]. Сети последнего типа будем называть словарными, используя их в данной работе в качестве объекта компьютерного анализа. Для того чтобы избавиться от остатков синтаксических и семантических взаимоотношений между словами, будем пренебрегать порядком слов в тексте, ограничиваясь ненаправленными взвешенными графами. Исследовалось три варианта сетей совместной встречаемости слов. С целью анализа структуры и классификации научных направлений с использованием англоязычной базы научных публикаций PubMed построены словарные сети двух типов: сеть ключевых слов и аннотационная сеть по словам MeshWords. В отличие от обычных авторских ключевых слов с помощью MeshWords статьи аннотируются при записи в базу публикаций. В настоящий момент в PubMed содержится около 19 370 000 записей, но лишь те из них, которые имеют поля key words или meshwords, дают вклад в соответствующие сети и были использованы для анализа. Для сравнения были выбраны словарные сети иной природы — сети соседства слов в 39 классических литературных текстах XIX-XX вв. разных жанров на русском языке.

Кратко опишем преобразование текста в ненаправленный взвешенный граф, которое использовалось в данной работе. В словарной сети каждая вершина соответствует слову. Сначала для большего единообразия литературных текстов из них были удалены числительные, даты и имена собственные. Затем для каждого из 39 произведений строилась своя сеть путем соединения вершин связями, если соответствующие вершинам слова соседствовали в одном предложении (рис. 1). Часть текстов принадлежала одним и тем же авторам.

Оставшиеся две сети строились аналогично с несколько иными критериями соответствия вершин и образования ребер. Нередко ключевые (key word) или аннотационные (meshword) слова в научных статьях представляют собой словосочетания, которые имеют самостоятельное значение и являются элементарными смысловыми модулями. Поэтому вершине в этих сетях сопоставлялось одиночное слово или словосочетание (рис. 2) в зависимости от того, что записано в соответствующем поле базы. Пара вершин соединялась ребром в случае присутствия соответствующих слов в одной и той же публикации.

Каждому ключевому слову в отдельном поле включая словосочетания ставилась в соответствие своя вершина графа. Все вершины, соответствующие ключевым словам в одной и той же публикации, соединялись ребрами.

Ниже приведен фрагмент карточки с полями ключевых слов для одной из публикаций в базе PubMed:

```
<KeywordList Owner="PIP»
     < Keyword MajorTopicYN="N»Americas</Keyword>
     < Keyword MajorTopicYN="N»Canada</Keyword>
     < Keyword MajorTopicYN="Y»Cancer</Keyword>
     < Keyword MajorTopicYN="Y»Causes Of Death</Keyword>
     < Keyword MajorTopicYN="N»Demographic Factors</Keyword>
     < Keyword MajorTopicYN="N»Developed Countries</Keyword>
     < Keyword MajorTopicYN="N»Diseases</Keyword>
     < Keyword MajorTopicYN="N»Mortality</Keyword>
     < Keyword MajorTopicYN="N»Neoplasms</Keyword>
     < Keyword MajorTopicYN="N»North America</Keyword>
     < Keyword MajorTopicYN="N»Northern America</Keyword>
     < Keyword MajorTopicYN="N»Population</Keyword>
     < Keyword MajorTopicYN="N»Population Dynamics</Keyword>
</KeywordList >
```

В полученных сетях частоты наблюдения слов и их одновременной встречаемости определяют веса вершин и ребер взвешенных графов, результаты анализа которых представлены в п. 3.

3. Статистика словарных сетей. Часть полезной информации из словарной сети можно извлечь без учета взаимосвязей (ребер). Классическим примером является распределение слов по рангу встречаемости в текстах на английском языке, которое подчиняется известному закону Ципфа. По ранговому распределению реальные тексты неотличимы от случайных, что обусловлено комбинаторной природой текста, который можно представить в виде непрерывной линейной символьной последовательности из алфавита, расширенного на один символ пробела между словами [7]. Степенным является также распределение $P(k)$, как было показано для ряда европейских [6] и китайского [8] языков. Для сравнения были построены распределения по частоте употребления слов для литературных текстов на русском языке и текста компьютерной программы blast, которая используется для выравнивания генетических последовательностей. Усредненные показатели степени оказались близкими к одному и тому же значению, равному -0,41. Такая универсальность для столь различных по своей природе текстов свидетельствует о малой информативности закономерности.

Одним из инструментов классической лингвистики является частотный анализ текста, который, в частности, использовался для идентификации авторства по частоте чередования

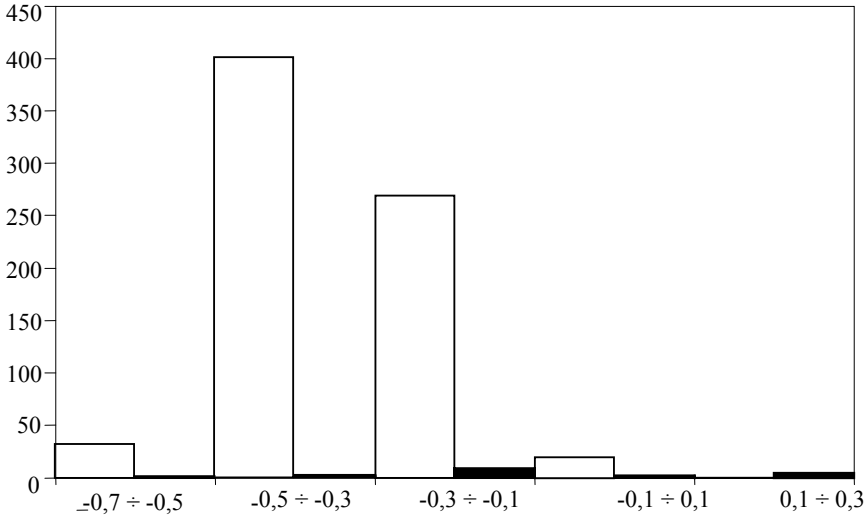

Рис. 2. Гистограмма коэффициентов корреляции относительных частот использования слов для текстов разных (светлые области слева) и одних и тех же (темные области справа) авторов для вторых 300 (т. е. со 101-го по 400-е) наиболее часто встречающихся слов (по оси абсцисс отложены диапазоны значений коэффициента корреляции)

букв [9]. Для каждого текста строился свой частотный словарь на основе расчета относительных частот наблюдения слов. Затем оценивалась близость текстов по коэффициенту корреляции между наборами относительных частот использования одних и тех же слов. Из $39 \cdot 19 = 741$ пар текстов 722 принадлежали разным авторам. Еще 19 пар образованы 21 произведением 8 совпадающих авторов, от А. С. Пушкина до А. Н. Толстого. Для выбранных текстов русской литературы обнаружено, что частота наблюдения слов оказывает влияние на характер их использования (и близость текстов) (рис. 2). Тексты, принадлежащие перу одного и того же и разных авторов, оказались неразличимыми, если использовались словари, включающие 100 наиболее употребимых слов. Эти слова, преимущественно низкого семантического содержания, например предлоги, соответствуют вершинам-концентраторам словарной сети. Авторство становится заметным при рассмотрении вторых 300 (т. е. со 101-го по 400-е) наиболее часто наблюдаемых слов (см. рис. 2). Видимо, авторский стиль проявляется ярче не в использовании слов, а при конструировании предложений. Следует отметить, что тот же порог в 100 слов преодолевает ребенок в своем взрослении при переходе к грамматическому построению речи [10].

Среди пар исследованных текстов положительные значения коэффициента корреляции наблюдались только для произведений одного и того же автора.

Временные пики или провалы употребления слов в научных текстах могут сигнализировать о “горячих” или “холодных” направлениях в науке. В ходе анализа динамики и структуры словарной сети эволюция науки была прослежена по 50 наиболее часто встречающимся словам из постоянно цитируемых публикаций журнала PNAS [11]. Временная неравномерность использования слов была выявлена при исследовании словарных сетей уже для всей базы научных публикаций по биологии и медицине. Например, число статей с использованием популярного термина “nanotechnology” увеличивается экспоненциально (рис. 3), опережая рост общего числа публикаций. В то же время практически не употребляются наиболее часто используемые в прошлом аннотационные слова meshwords (рис. 4).

Природу и динамику научных направлений можно детализировать, включив взаимодействие слов в словарных сетях (рис. 5). Замечено, что слово “population” оказалось единственным общим среди наиболее часто используемых слов в обеих сетях научной классификации

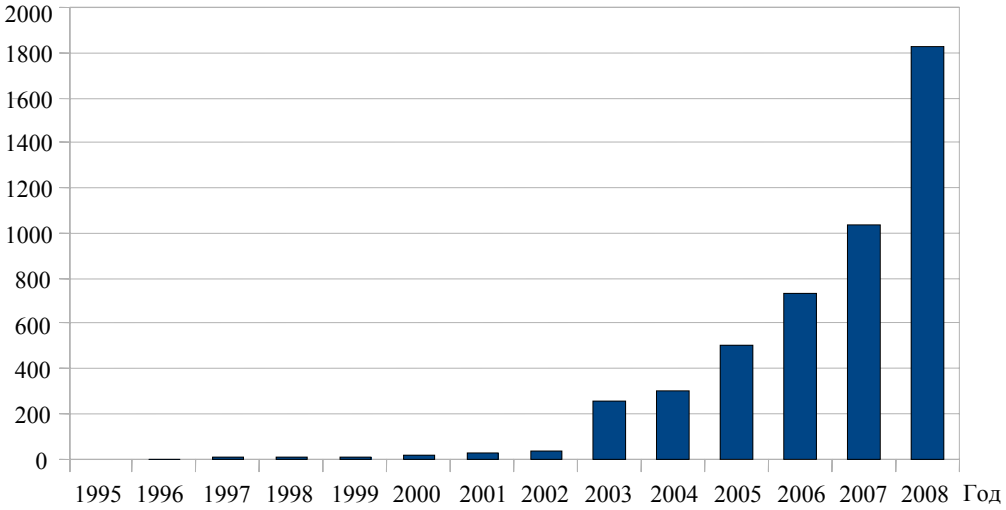

Рис. 3. Динамика числа статей биомедицинской тематики, использующих в поле Abstract термин “nanotechnology”

(ключевых слов и mesh word). Из зависимости частоты использования от времени следует, что всплеск интереса к данной теме наблюдался во второй половине 90-х гг. XX в. (см. рис. 4). По совместной встречаемости ключевых слов можно сузить “остывшую” область — это исследование демографии развивающихся стран (см. рис. 5).

Для выявления общих закономерностей рассмотренных в данной работе сетей были рассчитаны их статистические характеристики (см. таблицу).

Статистические характеристики словарных сетей

Словарная сеть	N	K	$<k>$	C_1	C_2	r
Keywords	11 362	1 902 780	334,00	0,29	0,78	−0,34
Meshwords	6407	625 096	195,00	0,26	0,74	−0,34
Литературные тексты (средние)	6176	22 528	3,55	0,02	0,05	−0,17

Примечание. N — число вершин; K — число ребер; $<k>$ — средняя связность; C_1, C_2 — локальный и глобальный коэффициенты кластеризации; r — коэффициент корреляции степеней вершин

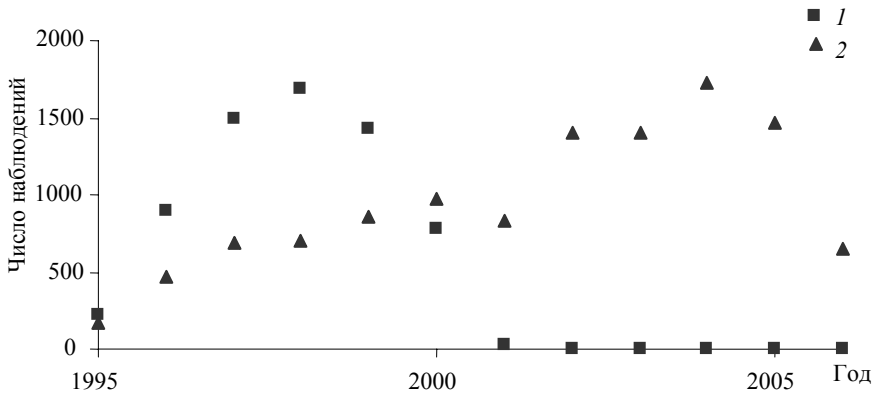

Рис. 4. Динамика использования наиболее часто встречающихся аннотационных слов meshword в 1995-2006 гг. в базе публикаций PubMed: *1* — population; *2* — biomedical and behavioral research

Из таблицы следует, что среди рассмотренных случаев сети литературных текстов наиболее разрежены, поскольку обладают меньшей средней связностью вследствие линейного механизма генерации. Вероятно, по той же причине литературные сети менее кластеризованы. Все исследованные словарные сети являются дисассортивными (когда вершины с большим числом ребер избегают быть связанными между собой). По-видимому, дисассортивность литературных сетей есть отражение имманентных принципов построения предложений. Дисассортивность сетей ключевых слов может быть обусловлена специализацией науки: области исследований ограничены использованием наиболее распространенных терминов. При этом коэффициенты корреляции научных сетей совпадают, что может свидетельствовать об объективности крупномасштабной кластеризации и категоризации науки.

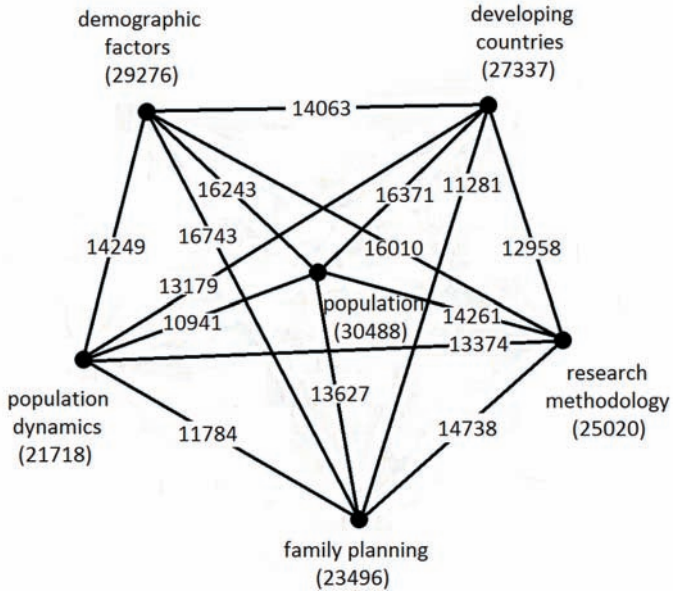

Рис. 5. Фрагмент сети ключевых слов, содержащий наиболее часто встречающиеся слова: цифры у вершин — общее число публикаций в соответствующей рубрике, т. е. содержащих данное ключевое слово; цифры у ребер — число публикаций, одновременно принадлежащих соответствующей паре рубрик

Заключение. Современная наука развивается быстрыми темпами, открывая новые перспективы и предавая забвению “немодные” темы. С высокой скоростью идет специализация, что часто приводит к фрагментации и дублированию работы. Для планирования может оказаться полезным как изучение общих закономерностей организации научных знаний, так и объективная классификация, категоризация науки, выявление перспективных направлений. Автоматизация извлечения соответствующей информации, “картирование” и прогноз научного прогресса имеют очень большое значение. До настоящего времени классификацию научных публикаций осуществляли сами авторы или аннотаторы баз данных. Более объективной является классификация, основанная на результатах компьютерной обработки научных публикаций, в которой вместо сетей ключевых слов используются словарные сети терминов, полученные удалением из текста общеупотребительной лексики. Практическое значение имеют также словари терминов, являющиеся промежуточными результатами работы.

Список литературы

1. Mack R., Hehenberger M. Text-based knowledge discovery: search and mining of life-sciences documents // Drug Discov. Today. 2002. V. 7, N 11. P. 89–98.

2. Newman M. E. J. The structure and functions of complex networks // SIAM Rev. 2003. V. 45, N 2. P. 167–256.

3. Erdös P., Rènyi A. On random graphs I // Publ. Math. Debrecen. 1959. V. 6. P. 290–297.

4. Milgram S. The small world problem // Psych. Today. 1967. V. 1, N 1. P. 61–67.

5. Титов И. И. Архитектура, динамика и эволюция безмасштабных сетей // Системная компьютерная биология. Новосибирск: Изд-во СО РАН, 2008. С. 329–333.

6. Solè R. V., Corominas-Murtra B., Valverde S., Steels L. Language networks: their structure, function and evolution // Complexity. 2010. V. 15, N 6. P. 20–26.

7. Li W. Random texts exhibit Zipf's-law-like word frequency distribution // IEEE Trans. Inform. Theory. 1992. V. 38, N 6. P. 1842–1845.

8. Zhoua S., Hua G., Zhanga Z., Guanc J. An empirical study of Chinese language networks // Phys. A: Statist. Mech. Appl. 2008. V. 387, N 12. P. 3039–3047.

9. Кукушкина О. В., Поликарпов А. А., Хмелев Д. В. Определение авторства текста с использованием буквенной и грамматической информации // Пробл. передачи информации. 2001. Т. 37, вып. 2. С. 96–108.

10. Ke J., Yao Y. Analyzing language development from a network approach // J. Quantitative Linguist. 2008. V. 15, N 1. P. 70–99.

11. Mane K. K., Börner K. Mapping topics and topic bursts in PNAS // PNAS. 2004. V. 101, suppl. 1. P. 5287–5290.

Рудниченко Константин Алексеевич — асп. Института вычислительной математики и математической геофизики СО РАН; тел.: 8-923-177-73-90, e-mail: rudnichenko.k@gmail.com;
Куликов Александр Иванович — зав. кафедрой Высшего колледжа информатики Новосибирского государственного университета, тел.: 333-24-50, e-mail: kulikov@nmsf.sscc.ru;
Титов Игорь Иванович — канд. физ.-мат. наук, ст. науч. сотр. Института цитологии и генетики СО РАН, тел.: 363-49-24, e-mail: titov@bionet.nsc.ru

Дата поступления — 17.12.09