УДК 004.912

АЛГОРИТМ ОТБОРА ТЕМАТИЧЕСКИ СХОЖИХ ДОКУМЕНТОВ С ПОСТРОЕНИЕМ КОНТЕКСТНО-СЕМАНТИЧЕСКОГО ГРАФА НА ОСНОВЕ ВЕРОЯТНОСТНО-ЭНТРОПИЙНОГО ПОДХОДА

И. А. Молошников, А. Г. Сбоев, Д. В. Гудовских

Национальный исследовательский центр «Курчатовский институт»

Поступила в редакцию 14 апреля 2015 г.

Аннотация: *в работе предложен алгоритм поиска тематически схожих документов на основе эталонной коллекции текстов с возможностью наглядной визуализации результатов поиска в виде контекстно-семантического графа вложенных тем. Алгоритм основан на интеграции множества вероятностно-энтропийных индикаторов для выделения набора ключевых слов и словосочетаний, описывающего тему для поиска. Результаты тестирования продемонстрировали среднюю точность отбора документов 99 % при полноте 84 % на основе выборки, предоставленной экспертами. Также предложен подход к построению графа на базе алгоритма извлечения ключевых словосочетаний с весами, что позволяет отобразить структуру вложенных тем в больших коллекциях документов в компактном виде.*
Ключевые слова: *семантический алгоритм Гинзбурга, поиск схожих документов, контекстно-семантический граф.*

Abstract: *the paper presents the algorithm of the search of semantically close documents. The algorithm is based on a selected reference collection of documents. It creates a context-semantic graph for visualizing themes in search results is presented. The algorithm is based on integration of set of probabilistic, entropic, and semantic markers for extractions of ranged key words and combinations of words, which describe the given topic. Test results demonstrate an average accuracy of 99 % and the recall of 84 % on expert selection of documents. Also developed special approach to constructing graph on base of algorithms extract key phrases with weights. It gives the possibility to demonstrate a structure of subtopics in large collections of documents in compact graph form.*
Key words: *Ginzburg semantic algorithm, search similar documents, context-semantic graph.*

Задача анализа больших, постоянно растущих объемов информации требует комплексного решения ряда подзадач: описание в краткой форме темы интересов пользователя; отбор документов, соответствующих теме; представление в наглядной форме результата анализа данных документов.

Существует ряд систем, частично позволяющих решить задачу поиска тематически схожих документов.

В частности, в поисковых системах, основанных на Apache Lucene, поиск подобных текстов производится с использованием заранее определенного документа и реализуется по методу «мешка слов». Несмотря на такие достоинства данного метода, как производительность и универсальность, его использование для анализа тем «эволюционирующих» с изменением в течение времени состава ключевых слов вызывает трудности.

Другой подход заключается в представлении документов в виде вектора заданной размерности и использовании различных метрик близости двух векторов. В рамках этого подхода используется набор как статистических методов: LDA, PLSA [1], так и основанных на нейронных сетях – Doc2vec [2]. Недостатками подхода являются требование к наличию большого корпуса для обучения модели, невысокая точность и сложность определения необходимого уровня близости.

Представленный нами алгоритм поиска тематически схожих документов похож на описанный в статье V. Govindaraju и K. Ramanathan [3]. Он основан на выделении набора ключевых слов и словосочетаний из представленной пользователем подборки текстов по теме и дальнейшем поиске на основе выделенных слов с ранжирование результатов. Отличие нашего метода состоит в способе выделения ключевых слов и словосочетаний (комбинация 2-х или 3-х слов в рамках одного предложении, без учета порядка), использования комбинации вероятностно-энтропийных характеристик текстов, семантического алгоритма Гинзбурга и дополнительных источников информации, таких как Национальный корпус русского языка.

Предложенный метод позволяет также строить контекстно-семантический граф, отражающий основные темы, представленные в результатах поиска, что

дает пользователю возможность быстро оценить предмет поиска.

Общая схема системы поиска тематически схожих документов

Общая схема системы представлена на рис. 1. Система позволяет:

1) формировать набор ключевых слов и словосочетаний, описывающих предметную область, на основе заданной небольшой коллекции тематических текстов – эталонной коллекции;

2) находить тематически схожие документы;

3) отображать вложенные темы в виде контекстно-семантического графа.

Система состоит из хранилища данных, аналитического модуля, модуля поиска тематически схожих документов и модуля визуализации результатов поиска. Для анализа темы пользователь предоставляет коллекцию документов, отражающих предмет поиска, которая называется эталонной коллекцией. Система позволяет выделить основные ключевые слова темы. Основываясь на них и эталонной коллекции производится отбор и ранжирование документов из хранилища. Используя аналитический модуль на результатах поиска, формируются тематические кластеры слов и словосочетаний, которые отображаются в виде контекстно-семантического графа.

Методы выделения ключевых слов и словосочетаний

Аналитический модуль производит ранжирование слов и словосочетаний, отражающее их принадлежность к теме. Под словосочетанием, биграммой или триграммой подразумевается комбинация из двух или трех слов встречающихся внутри одного предложения без учета последовательности в предложении. Входными данными для модуля служит предобработанный текст. В предобработку входит приведение слов к нормальной форме, разбиение текста на предложения (производится методами морфологического модуля АОТ), словарная фильтрация наиболее употребительных слов (предлогов, союзов и т. п.). Модуль использует вероятностно-энтропийные и семантические индикаторы для расчета ранга термина, базирующиеся на:

1) дивергенции Кульбака — Лейблера, используемой для сравнения распределений терминов;

2) информационной энтропии отражающей равномерность распределения термина по документам коллекции;

3) весах, основанных на распределении Бернулли;

4) семантическом алгоритме Гинзбурга для определения близости двух слов.

На основе нормализованных значений указанных индикаторов вычисляется единый ранг для каждого термина. Слово «термин» обозначает «слово», в случае, если индикатор рассчитывается для ключевых слов или означает «словосочетание», применимо к ключевым словосочетаниям.

Отбирается 100 слов с наивысшим рангом и формируются словосочетания (биграммы и триграммы) с ними без учета последовательности слов в предложении. Далее вычисляются ранги для словосочетаний и выбираются биграммы и триграммы с наивысшим рейтингом. Используя выделенные слова и словосочетания для каждого документа эталонной коллекции, рассчитывается суммарный вес документа, равный сумме рангов ключевых слов и словосочетаний, входящих в него. Минимальный суммарный ранг документа эталонной коллекции выбирается как базовая линия для фильтрации не релевантных текстов.

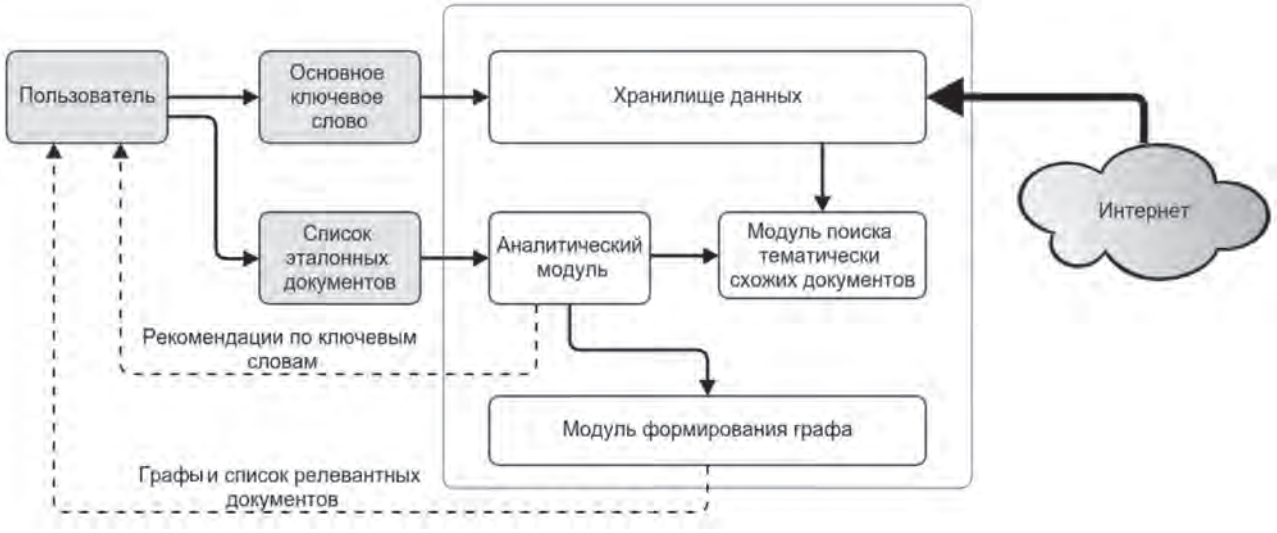

Рис. 1. Общая схема системы

Дивергенция Кульбака – Лейблера

Индикатор, основанный на дивергенции Кульбака – Лейблера, рассчитывается для слов и словосочетаний, согласно формуле 1. Он характеризует различие между реальным распределением термина *w* и теоретическим в соответствии с длиной документа (чем больше документ, тем больше в нем различных терминов, а значит, больше вероятность случайного попадания термина *w* в документ *d*).

$$D(w) = \sum_{d \in D} p_{doc}(w,d) \cdot \ln\left(\frac{p_{doc}(w,d)}{p_n(d)}\right), \quad (1)$$

где p_n – вероятность встретить термин *w* во всей коллекции документов относительно длины документа *d*.

$$p_n(d) = \frac{N(d)}{\sum_{x \in D} N(x)}, \quad (2)$$

где *N(d)* – общее количество терминов в документе *d*; сумма *N(x)* – общее количество терминов во всей коллекции; $p_{doc}(w, d)$ – вероятность встречаемости термина *w* в документе *d*.

$$p_{doc}(w,d) = \frac{tf(w,d)}{F(w)}, \quad (3)$$

где *tf(w, d)* – встречаемость термина *w* в документе *d*; *F(w)* – встречаемость термина *w* в коллекции.

Малое значение величины *D(w)* показывает, что данное слово характерно для представленной выборки. Это может быть общеупотребительное слово, фильтруемое за счет других индикаторов, или тематическое ключевое слово.

Информационная энтропия

Информационная энтропия показывает равномерность распределения термина *w* в документах коллекции и рассчитывается по формуле:

$$H(w) \quad \sum_{d \in D} p_{doc}(w,d) \ln\left(\frac{1}{p_{doc}(w,d)}\right) \quad (4)$$

Если данный показатель большой, то термин равномерно представлен в коллекции документов, если он равен 0, то это означает, что термин *w* встречается только в одном документе. В тематической коллекции ключевые слова равномерно распределены по набору документов.

Индикатор выделения общеупотребительных слов

Индикатор показывает отличие распределения слова *w* в эталонной коллекции и в Национальном корпусе русского языка. Рассчитывается по формуле *r = pe/pk*, где *pe* – относительная частота встречаемости термина в эталонной коллекции, *pk* – относительная частота встречаемости термина в Национальном корпусе русского языка.

Данный индикатор позволяет выделить большую часть общих слов, если они хорошо представлены в корпусе. Для общеупотребительных он будет иметь значение около 1, а для специализированных слов он много большее 1.

Веса на основе распределения Бернулли

Данный тип индикаторов основывается на сравнении реального распределения терминов в коллекции с теоретическим распределением Бернулли.

Мы используем веса W_1 и W_2 в качестве индикаторов, основываясь на результатах из статьи G. Amati и C. J. Van Rijsbergen [4]. Они рассчитываются по формулам:

$$W_1(w) \quad \sum_{x \in D} Wrisk_1(w,x), \quad (5)$$

$$W_2(w) \quad \sum_{x \in D} Wrisk_2(w,x), \quad (6)$$

$$Wrisk_1(w,x) \quad \frac{-\log_2 Prob_{norm}(w,d)}{tf(w,d)+1}, \quad (7)$$

$$Wrisk_2(w,x) \quad \frac{F(w)(-\log_2 Prob_{norm}(w,d))}{df(w)(tf(w,d)+1)}, \quad (8)$$

Здесь *df(w)* – количество документов в коллекции, содержащих термин *w*; $Prob_{norm}$ рассчитывается по следующим формулам:

$$Prob_{norm}(w,d) = \frac{Prob(w,d)}{\sum_{x \in D} Prob(w,x)}, \quad (9)$$

$$Prob(w,d) \quad 2^{-\log_2 Prob_1(w,d)}, \quad (10)$$

$$Prob_1(w,d) \quad B(N,F,X) \quad \binom{F(w)}{tf(w,d)} p^{tf(w,d)} q^{F(w)-tf(w,d)}. \quad (11)$$

Основываясь на величинах $Prob_{norm}$, $Wrisk_2$, мы предлагаем два новых индикатора $D_{fe}1$ и $D_{fe}2$:

$$D_{fe}1(w) \quad \sum_{x \in D} p_{doc}(w,x) log_2\left(\frac{p_{doc}(w,x)}{Prob_{norm}(w,x)}\right) \quad (12)$$

$$D_{fe}2(w) \quad \sum_{x \in D} p_{doc}(w,x) log_2\left(\frac{p_{doc}(w,x)}{Wrisk2_{norm}(w,x)}\right), \quad (13)$$

$$Wrisk2_{norm}(w,d) \quad \frac{Wrisk_2(w,d)}{\sum_{x \in D} Wrisk_2(w,x)}. \quad (14)$$

Для ключевых слов индикаторы W_1, W_2, $D_{fe}1$, $D_{fe}2$ должны иметь большое значение.

Семантический алгоритм Гинзбурга

Семантический алгоритм Гинзбурга [5] предназначен для выделения слов, схожих по контексту, в

котором они употребляются, при этом вводится индекс значимости:

$$ind\left(a \mid c\right) = \frac{N_{ac} N_t}{N_{tc} N_a}. \quad (15)$$

Он используется для определения семантической близости двух слов по их окружению (в рамках одного предложения), где N_{ac} – встречаемость слова *a* со словом *c*, N_t – общее число слов в коллекции документов, N_{tc} – общее число слов в окружении слова *c*, N_a – встречаемость слово *a* в коллекции.

Индекс значимости рассчитывается для всех слов, встречающихся в одном предложении со словом *A* или *C*, для которых рассчитывается близость по контексту. Если *ind(a|c)* больше 1 – это означает, что данный показатель значим при расчете. Индексы значимости представлены на рис. 2 как ребра *ind 1*, *ind 2*, *ind 3* и т.д.

Индикатор связанности по Гинзбургу, определяющий силу семантической связи двух слов, рассчитывается на основе индексов значимости по следующей формуле:

$$ginz(A,C) \quad 1 - \frac{sum(A) + sum\_razn + sun(C)}{sum\_all}, \quad (16)$$

где *sum(A)* – сумма индексов значимости со словом *A*, больших чем 1 и не принадлежащих к окружению слова *C* (на рис. 2 это сумма индексов *ind 1*, *ind 2*, *ind 3*, *ind 4*); *sum(C)* – сумма индексов значимости со словом *C*, больших чем 1 и не принадлежащих к окружению слова *A* (на рис. 2 это сумма индексов *ind 11*, *ind 12*, *ind 13*); sum_razn – сумма абсолютных значений разностей индексов значимости в общей части (для рис. 2 эта сумма равна |ind7–ind8|+|ind6–ind9|+|ind5–ind10|); sum_all – сумма всех индексов значимости больших, чем 1.

Значения индикатора связанности по Гинзбургу лежат в интервале от 0 до 1; 0 – слова не связаны, 1 – максимальная связанность слов.

Алгоритм формирования ключевых словосочетаний

Алгоритм формирования ключевых словосочетаний состоит из нескольких шагов:

1. Выбрать N наиболее частотных слов в эталонной коллекции в кандидаты на ключевые слова (экспериментально было подобрано оптимальное значение N = 1000).

2. Рассчитать для кандидатов все описанные ранее индикаторы, за исключением индикатора связанности по Гинзбургу.

3. Нормализовать полученные значения индикаторов. Нормализованные значения формируют одно значение для каждого кандидата, называемое ранг, отражающее принадлежность данного слова к теме.

4. Слова с наивысшим значением ранга являются ключевыми словами темы.

5. На основе выделенных ключевых слов формируются биграммы и триграммы. Эти словосочетания создаются из слов одного предложения без учета последовательности и их положения в предложении.

6. Для ранжирования биграмм и триграмм рассчитываются ранее описанные индикаторы, включая индикатор связанности по Гинзбургу, и аналогично ключевым словам формируется ранг ключевого словосочетания.

7. Биграммы и триграммы с наивысшим значением ранга являются ключевыми словосочетаниями темы.

Результатом работы данного алгоритма является взвешенный относительно темы, заданной эталонной коллекцией, набор ключевых слов и словосочетаний.

Поиск тематически схожих документов

Если представлена эталонная коллекция и выделены ключевые слова и словосочетания темы, то можно переходить к этапу поиска тематически схо-

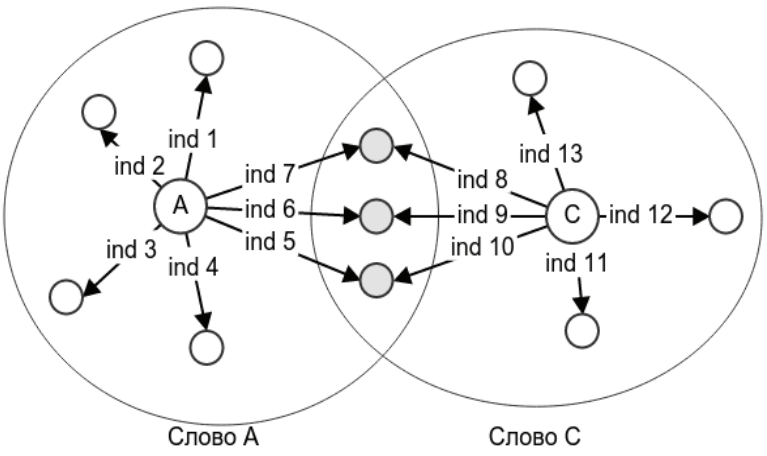

Рис. 2. Алгоритм Гинзбурга

жих документов в хранилище. Схематично этот процесс представлен на рис. 3. Он основан на ранее описанных алгоритмах и трех источниках информации. Первый – это заданная пользователем эталонная коллекция документов, подсказывающая, что необходимо искать, второй – это Национальный корпус русского языка, использующийся для вычисления частоты общеупотребительных слов, и третий – набор документов, полученный в результате поиска по основным ключевым словам. Используя эти входные данные, модуль формирует ключевые слова и словосочетания для ранжирования документов, на релевантность заданной эталоном теме. При этом рассчитывается минимальная граница релевантности теме документов эталонной коллекции.

По результатам анализа ключевых слов и словосочетаний вместе с результатами поиска по основным ключевым словам модуль формирует минус-слова темы. Минус-слова темы может встречаться с основными ключевыми словами только в других предметных областях. Например, для темы «Автомобили Форд» система выдаст следующие минус-слова: Марк, Том, Харрисон. Марк Форд – поэт, Том Форд – дизайнер и кинорежиссер, Харрисон Форд – знаменитый киноактер.

Полученные из хранилища документы взвешиваются и фильтруются на основе ключевых слов, словосочетаний и минус-слов темы. Пользователю представляется отранжированный список документов, тематически схожих с документами эталонной коллекции.

Методы построения контекстно-семантического графа

Результаты тематического поиска могут содержать большое число документов. Для наглядной визуализации вложенных тем используется контекстно-семантический граф. Узлами данного графа являются ключевые слова, а ребрами ключевые биграммы, полученные в ходе анализа результатов поиска, представленными ранее методами. Размер узлов, расстояние между ними и толщина связывающих линий характеризуют связанность подтем в коллекции документов.

Для построения графа коллекция документов обрабатывается аналитическим модулем с получением списка ключевых слов (узлов) и биграмм с рангами. Из ключевых биграмм строится матрица смежности, описывающая связи будущего графа. К этой матрице применяется метод кластеризации на основе близости узлов через соседей (Affinity Propagation с использованием библиотеки scikit-learn [6]). Результатом работы алгоритма является набор вложенных тем, представленных кластерами из ключевых слов. Для визуализации графов использовался алгоритм укладки «Force Atlas 2», реализованный в программном продукте Gephi [7]. На рис. 4 представлена часть такого графа, построенного по выборке из 9000 документов по теме «Ракета Булава».

Для расчета связанности тематических кластеров и дальнейшей визуализации веса связей для узлов из разных кластеров объединяются и отражаются в виде ребра между центрами кластеров. Пример части такого графа для выборки по теме «Арктика», состоящей из 14 000 документов, представлен на рис. 5.

Тестирование системы

Совместно с экспертами было проведено тестирование по 4 темам. Экспертами были подготовлены сложные запросы к хранилищу, результат запроса эксперта принимается в качестве 100 % релевантного теме. Экспертами также были подобраны эталонные выборки по каждой теме. Размер полученного по запросу экспертов набора документов составил около 5000 документов.

Для каждой из 4 тем в результате анализа эталонной коллекции описанными выше алгоритмами автоматизированно сформирован запрос к хранилищу.

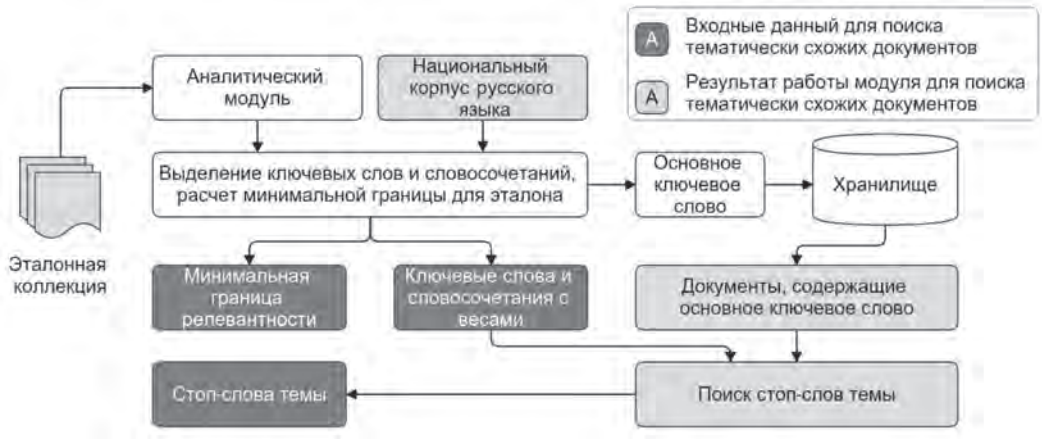

Рис. 3. Схема поиска тематически схожих документов

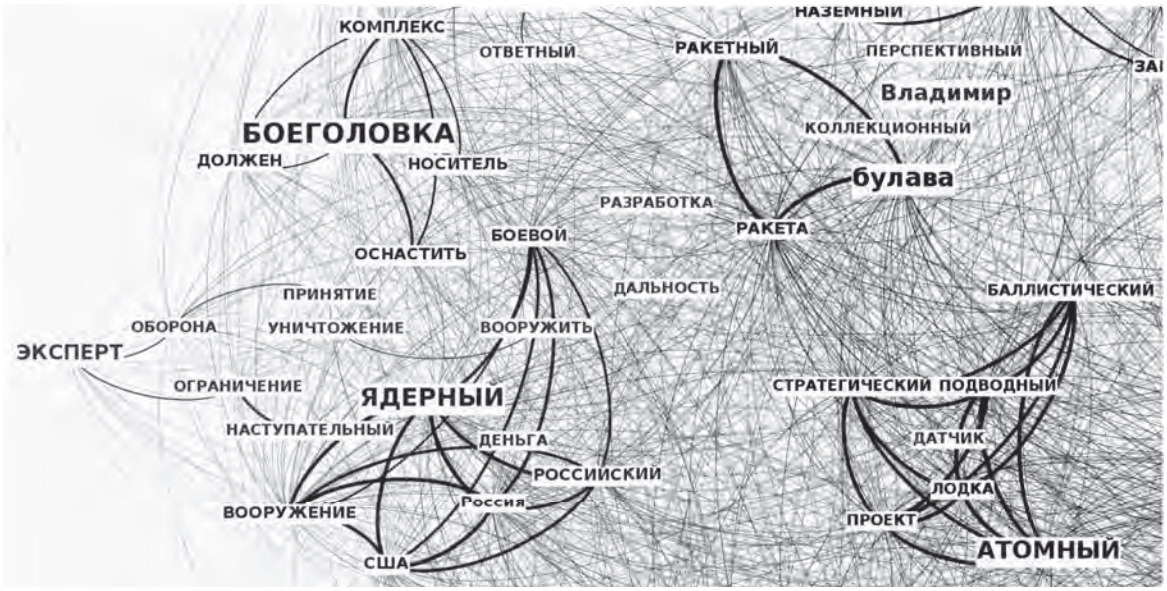

Рис. 4. Часть графа для выборки по теме «Ракета Булава»

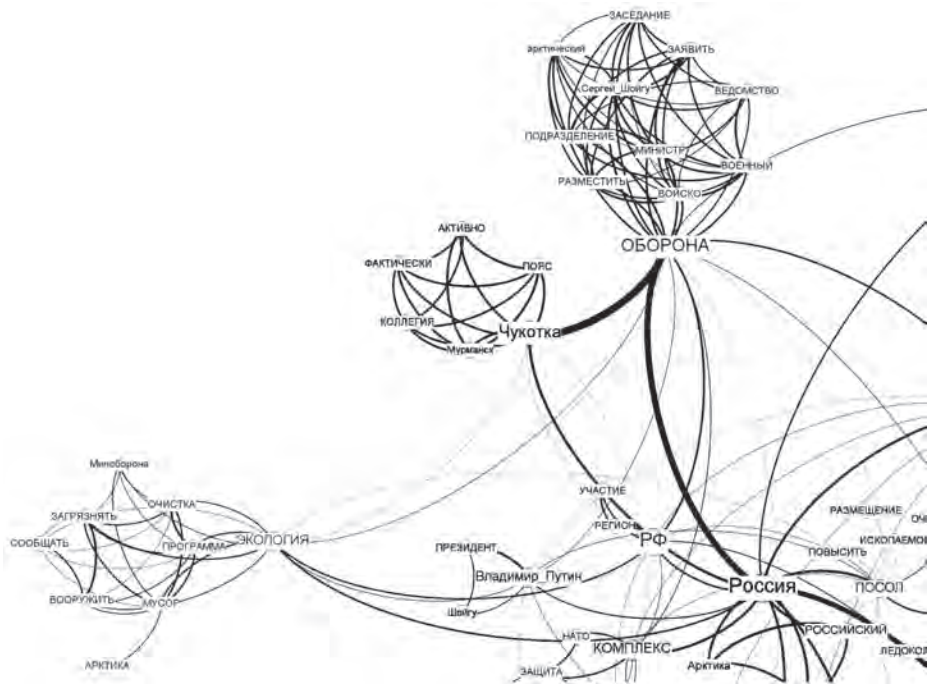

Рис. 5. Часть графа для выборке по теме «Арктика»

Сравнение результатов по 4 темам от экспертов и полученных в ходе работы системы показало точность 99 % и полноту 84 % разработанного алгоритма поиска тематически схожих документов на основе заданной эталонной коллекции.

Таким образом, на основе описанных выше алгоритмов создана система поиска тематически схожих документов по теме, задаваемой эталонной коллекцией, показавшая высокие значения точности и полноты по результатам тестирования совместно с экспертами. Также предложен алгоритм наглядной визуализации вложенных тем в виде контекстно-семантического графа для больших коллекций документов.

Представленные методы и алгоритмы закладывают основу для разработки системы отражающей эволюцию темы с течением времени и формирования новых подтем в рамках исходной темы.

ЛИТЕРАТУРА

1. *Blei David M., Ng. Andrew Y., Jordan Michael I.* Latent dirichlet allocation // Journal of machine Learning research. – 2003. – № 3. – P. 993–1022.

2. *Le Quoc, Mikolov T.* Distributed Representations of Sentences and Documents // Proceedings of The 31st International Conference on Machine Learning, 2014. – P. 1188–1196.

3. *Govindaraju V., Ramanathan K.* Similar Document Search and Recommendation // HP Laboratories, 2011.

4. *Amati Gianni and Van Rijsbergen.* Cornelis Joost Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness // ACM Transactions on Information Systems (TOIS) 20. – 2002. – № 4. – P. 357–389.

5. *Воронина И. Е.* Алгоритмы определения семантической близости ключевых слов по их окружению в тексте / И. Е. Воронина, А. А. Кретов, И. В. Попова // Вестник Воронеж гос. ун-та. Сер. : Системный анализ и информационные технологии. – 2010. – № 1. – С. 148–153.

6. *Pedregosa F., Varoquaux et.al.* Scikit-learn : Machine Learning in Python // Journal of Machine Learning Research. – 2011. – № 12. – P. 2825–2830.

7. *Bastian M., Heymann S., Jacomy M.* Gephi : an open source software for exploring and manipulating networks // International AAAI Conference on Weblogs and Social Media, 2009.

Национальный исследовательский центр «Курчатовский институт»

Молошников И. А., инженер-исследователь, аспирант

E-mail: ivan-rus@yandex.ru

Тел.: 8-926-378-09-57

Сбоев А. Г., кандидат физико-математических наук, ведущий научный сотрудник

E-mail: sag111@mail.ru

Тел.: 8-926-253-72-17

Гудовских Д. В., инженер-исследователь, аспирант

E-mail: dmitrygagus@gmail.com

Тел.: 8-963-994-26-50

National Research Center «Kurchatov Institute»

Moloshnikov I. A., Research Engineer, Post-graduate Student

E-mail: ivan-rus@yandex.ru

Tel.: 8-926-378-09-57

Sboev A. G., Candidate of Physical and Mathematical Sciences, Leading Researcher

E-mail: sag111@mail.ru

Tel.: 8-926-253-72-17

Gudovskikh D. V., Research Engineer, Post-graduate Student

E-mail: dmitrygagus@gmail.com

Tel.: 8-963-994-26-50