Электронные библиотеки: перспективные методы и технологии, электронные коллекции

XVI Всероссийская научная конференция RCDL-2014

Дубна, Россия, 13 – 16 октября

2014

Dubna, Russia, October 13 – 16

XVI All-Russian Scientific Conference RCDL-2014

Digital Libraries: Advanced Methods and Technologies, Digital Collections

Российский фонд фундаментальных исследований
Объединенный институт ядерных исследований
Институт проблем информатики Российской академии наук
Московская секция ACM SIGMOD

Электронные библиотеки: перспективные методы и технологии, электронные коллекции

XVI Всероссийская научная конференция RCDL-2014

Дубна, 13–16 октября 2014 г.

Труды конференции

Russian Foundation for Basic Research
Joint Institute for Nuclear Research
Institute of Informatics Problems of the Russian Academy of Sciences
Moscow ACM SIGMOD Chapter

Digital Libraries: Advanced Methods and Technologies, Digital Collections

XVI All-Russian Scientific Conference RCDL-2014

Dubna, October 13–16, 2014

Proceedings of the Conference

Дубна • 2014

УДК [002:004.9] (063)
ББК [73+32.973.233]я431
Э 45

Э 45 **Электронные** библиотеки: перспективные методы и технологии, электронные коллекции: XVI Всероссийская научная конференция RCDL-2014 (Дубна, 13–16 октября 2014 г.) : труды конференции / сост. Л. А. Калмыкова, М. Р. Когаловский. — Дубна : ОИЯИ, 2014. — 455, [1] с.

ISBN 978-5-9530-0397-1

Электронные библиотеки — область исследований и разработок, направленных на развитие теории и практики обработки, распространения, хранения, анализа и поиска цифровых данных различной природы. Основная цель серии конференций RCDL (http://rcdl.ru) заключается в формировании сообщества специалистов России, ведущих исследования и разработки в области электронных библиотек и близких областях. Всероссийская научная конференция 2014 г. (RCDL-2014) является шестнадцатой по данной тематике (1999, 2003 гг. — Санкт-Петербург, 2000 — Протвино, 2001, 2009 — Петрозаводск, 2002, 2008 — Дубна, 2004 — Пущино, 2005, 2013 — Ярославль, 2006 — Суздаль, 2007, 2012 — Переславль-Залесский, 2010 — Казань, 2011 — Воронеж). Настоящий сборник включает тексты докладов, коротких сообщений и стендовых докладов, отобранных программным комитетом для RCDL-2014 (Дубна, 13–16 октября 2014 г.).

Конференция организована при поддержке Российского фонда фундаментальных исследований (грант РФФИ № 14-07-20386) и Российской академии наук.

Digital Libraries: Advanced Methods and Technologies, Digital Collections : XVI All-Russian Scientific Conference RCDL-2014 (Dubna, October 13–16, 2014) : Proceedings of the Conference / composed by L. A. Kalmykova, M. R. Kogalovsky. — Dubna : JINR, 2014. — 455, [1] p.

ISBN 978-5-9530-0397-1

Digital Libraries is a field of research and development aiming to promote the theory and practice of processing, dissemination, storage, search and analysis of various digital data. The purpose of the series of All-Russian Scientific Conferences on Digital Libraries (RCDL, http://rcdl.ru) is to stimulate consolidation of the Russian digital libraries community and encourage research in this field. The All-Russian Scientific Conference RCDL-2014 is the sixteenth conference on this subject (1999, 2003 — St. Petersburg, 2000 — Protvino, 2001, 2009 — Petrozavodsk, 2002, 2008 — Dubna, 2004 — Pushchino, 2005, 2013 — Yaroslavl, 2006 — Suzdal, 2007, 2012 — Pereslavl-Zalessky, 2010 — Kazan, 2011 — Voronezh). The RCDL-2014 Proceedings include the texts of reports, short papers and posters selected by the Programme Committee for RCDL-2014 (Dubna, October 13–16, 2014).

The conference was organized with the support of the Russian Foundation for Basic Research (RFBR Grant No. 14-07-20386) and the Russian Academy of Sciences.

УДК [002:004.9] (063)
ББК [73+32.973.233]я431

ISBN 978-5-9530-0397-1

СОДЕРЖАНИЕ / CONTENTS

ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ, МЕТОДЫ ИХ ФОРМИРОВАНИЯ / LINGUISTIC RESOURCES, METHODS OF THEIR PRODUCTION

МЕТАДАННЫЕ И ОНТОЛОГИИ / METADATA AND ONTOLOGIES

СЕМАНТИЧЕСКАЯ ОБРАБОТКА ПОЛНОТЕКСТОВЫХ РЕСУРСОВ / SEMANTIC PROCESSING OF FULL-TEXT RESOURCES

Сеть естественных иерархий терминов новостных текстов по событиям «Евромайдана»

© Д. В. Ландэ © А. А. Снарский © Е. В. Ягунова
Институт проблем регистрации информации НАН Украины,
НТУУ «Киевский политехнический институт»,
Киев, Украина
dwlande@gmail.com asnarskii@gmail.com

Санкт-Петербургский государственный университет,
Санкт-Петербург, Россия
iagounova.elena@gmail.com

Аннотация

Описывается методика построения сетей иерархий терминов на основе тематических массивов новостных сообщений. Построены и исследованы такие сети, сформированные на основе автоматической обработки полных текстов сообщений о событиях, связанных с «Евромайданом» в Киеве.

1 Постановка проблемы

Построение большой тематической онтологии – сложная и затратная проблема. Определенным этапом разработки общих онтологий является формирование словарных номенклатур, терминологических онтологий. Эффективный автоматический отбор отдельных терминов для таких конструкций на основании неразмеченных текстовых массивов – не решенная окончательно задача [5, 6]. Проблема автоматического установления связей, построения сетей из таких терминов также до сих пор остается открытой.

Другой важной задачей является формальная оценка всплеска новых тем в информационных потоках, и, соответственно, терминов, маркирующих эти темы. Сегодня лингвист, работающий с новостными текстами, не может не заинтересоваться спецификой разных современных сегментов (срезов) по данным СМИ, потоков новостных сообщений [4, 8]. В частном случае, по терминам-маркерам можно понимать соответствие отдельных новостных сюжетов тематикам целых информационных потоков, оценивая используемую в них лексику.

Ниже описаны подходы к формированию терминологической основы цепочки событий, отражаемых в сообщениях электронных СМИ, а также отдельных сюжетов тематических новостей за определенные временные периоды, а также формирование на основании некоторых принципов языковой сети из отобранных терминов. Соответствие терминологии отдельного событийного сюжета и общей тематической терминологии (или терминологии цепочки связанных событий) можно рассматривать как формальный критерий релевантности данного события и рассматриваемой тематики (цепочки событий).

Предварительные этапы формирования языковой сети, связанной с цепочкой взаимосвязанных событий, включают такие шаги:

1. Нахождение релевантных тематике сообщений – формирование корпуса тематических новостных сообщений.

2. Определение динамики тематических сообщений.

3. Определение критических точек (дат) в динамике тематических сообщений.

4. Определение объектов мониторинга (терминов).

Рассмотрим их более подробно.

2 Формирование корпуса тематических новостных сообщений

На первом этапе выбирается исходный текстовый корпус, в качестве которого рассматриваются новостные сообщения по тематике противостояний в Киеве в 2013–2014 гг., связанных с так называемым «Евромайданом». Для отбора и последующего анализа тематических сообщений была использована система контент-мониторинга InfoStream [3]. Для нахождения релевантных тематике новостных сообщений был составлен запрос:

(майдан|евромайдан)&(избиен|разгон|штурм|беркут|молотов|титушк|погиб)&lang.RUS,

по которому в период с ноября 2013 г. по март 2014 г. было найдено свыше 200 тысяч новостных сообщений на веб-сайтах РуНета (рис. 1).

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

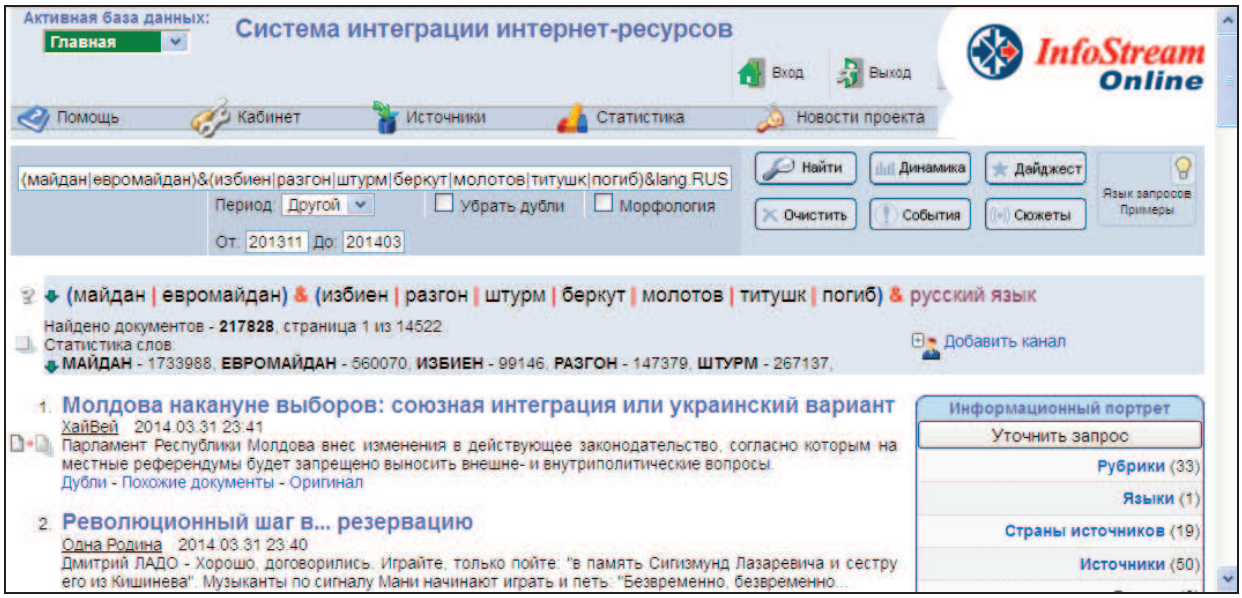

Рис. 1. Поисковый интерфейс системы InfoStream

Рис. 2. Динамика количества публикаций по запросу

3 Определение динамики тематических сообщений

Режим «Динамика событий» системы контент-мониторинга позволяет получить данные о количестве публикаций по заданному запросу за указанный промежуток времени. Эти данные отображаются в виде графика (рис. 2).

После этого данные временной динамики за каждые сутки нормируются, т.е. формируется временной ряд, содержащий относительные значения, равные отношению количества тематических сообщений к общему потоку сообщений за сутки (рис. 3). Это, в частности, позволяет избавиться от недельной периодичности в количестве тематических публикаций. Затем происходит переход к процедуре определения критических точек в данном временном ряду.

4 Определение критических точек в динамике сообщений

Критические точки как локальные максимумы временного ряда динамики публикаций можно определить, например, визуально по графику, представленному на рис. 3. Вместе с тем, существуют несколько научно-обоснованных методик, одна из которых базируется на вейвлет-анализе [2]. В работе [3] показано, что вейвлет «мексиканская шляпа» наиболее точно отражает динамику информационных операций, результаты применения этого вейвлета приведены на рис. 4, благодаря чему выбраны три даты (2013.11.30, 2014.01.22, 2014.02.19), соответствующие критическим точкам исследуемого процесса.

5 Выбор объектов мониторинга

После определения критических точек во временном ряду с помощью системы контент-мониторинга выполняется построение основных сюжетных цепочек из сообщений, соответствующих запросу за выбранные даты, которые определяют основные события за указанные даты (рис. 5).

Для последующего анализа отбирается три массива сообщений, соответствующие трем выбранным датам, особенности лексического состава которых являются объектами мониторинга.

	A	B	C	D	E	F
1	**Дата**	**Всего**	**Понятие**			
2	01.11.2013	84687	30		0.000354	
3	02.11.2013	31369	18		0.000574	
4	03.11.2013	27291	12		0.00044	
5	04.11.2013	64836	28		0.000432	
6	05.11.2013	88322	32		0.000362	
7	06.11.2013	89133	68		0.000763	
8	07.11.2013	86815	35		0.000403	
9	08.11.2013					
10	09.11.2013					
11	10.11.2013					
12	11.11.2013					
13	12.11.2013					
14	13.11.2013					
15	14.11.2013					
16	15.11.2013					
17	16.11.2013					
18	17.11.2013					
19	18.11.2013					
20	19.11.2013					
21	20.11.2013					
22	21.11.2013					
23	22.11.2013					
24	23.11.2013					
25	24.11.2013	31015	733		0.023634	
26	25.11.2013	89885	2721		0.030272	

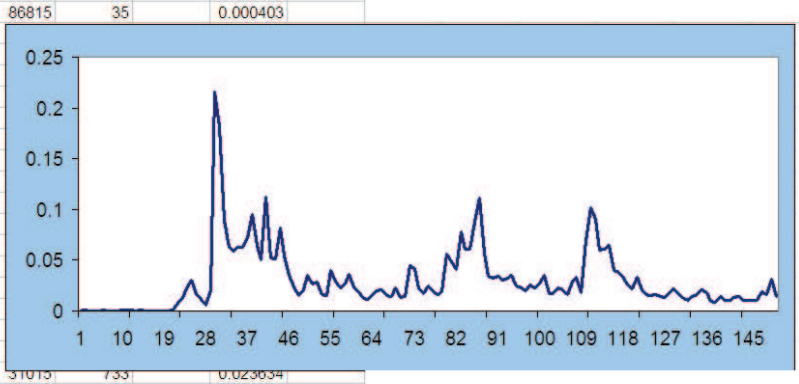

Рис. 3. Нормированная динамика тематических публикаций

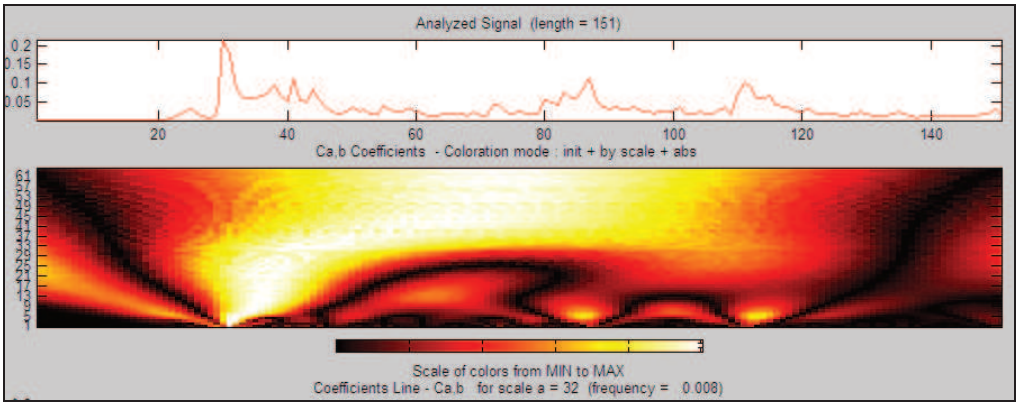

Рис. 4. Вейвлет-спектограммы исследуемого временного ряда

Рис. 5. Основные сюжетные цепочки за выбранные даты

Предварительная обработка отобранных текстовых массивов предусматривает выделение фрагментов текстов (отдельных сообщений, абзацев, предложений, слов, биграмм, триграмм), исключение нетекстовых символов, отсечение флективных окончаний – стемминг.

Далее каждому отдельному терму из текста (слову, биграмме или триграмме) ставится в соответствие оценка его «дискриминантная сила», а именно TFIDF, которая в каноническом виде равна произведению частоты соответствующего термина (Term Frequency) в фрагменте текста на двоичный логарифм от величины, обратной к количеству фрагментов текста, в которых этот терм встретился (Inverse Document Frequency) [14].

6 Сеть естественных иерархий терминов

Сеть естественной иерархии терминов (СЕИТ) базируется на разработанной ранее авторами данного доклада методологии выявления информационно-значимых элементов текста, опорных словах и словосочетаний [10, 12]. Использование таких элементов позволяет формировать сетевые информационные портреты, охватывать отдельные области знаний. Опорные слова и словосочетания как правило выбираются с учетом такого их свойства, как дискриминантная сила. Вместе с тем, одного этого свойства часто оказывается недостаточно для построения терминологических онтологий. Иногда слова с низкой дискриминантной силой, в частности, наиболее частотные слова из выбранной предметной области (например, слова «Украина», «Майдан», «Протест» в корпусе новостных сообщений о событиях, связанных с так называемым «Евромайданом» в Киеве) оказываются важнейшими для задач, которые рассматривается ниже.

Формирование сети естественных иерархий терминов базируется на контенте текстовых корпусов выбранной для анализа направленности. «Естественность» в этом случае понимается как отказ при формировании сети от специальных методов смыслового анализа, в том числе, разбора предложений по частям речи. Все связи в такой сети определяются естественным взаимным расположением слов и словосочетаний, которые экстрагируются из текстов статистически значимых объемов. Сеть естественных иерархий терминов, создаваемая полностью автоматически, может рассматриваться как основа для дальнейшего автоматизированного формирования терминологической онтологии с участием экспертов. Методика формирования сети естественных иерархий терминов, которая рассматривается в этой работе, предусматривает формирование компактифицированного графа горизонтальной видимости (CHVG), расчет новых весовых значений слов, биграмм и триграмм, а также непосредственное построение СЕИТ (соединение узлов связями «включения») и ее отображение [11].

Для последовательностей терминов и их весовых значений по TFIDF строятся компактифицированные графы горизонтальной видимости (CHVG) и выполняется повторное определение весовых значений слов уже по этому алгоритму [10]. Данная процедура позволяет учитывать в дальнейшем кроме терминов с большой дискриминантной силой также высокочастотные термины, которые имеют большое значение для общей тематики текстового корпуса. Сеть слов с использованием алгоритма горизонтальной видимости строится в три этапа. На первом на горизонтальной оси отмечается ряд узлов, каждый из которых соответствует словам в порядке появления в тексте, а по вертикальной оси откладываются весовые численные оценки (TFIDF). На втором этапе строится традиционный граф горизонтальной видимости [13]. Для этого между узлами существует связь, если они находятся в «прямой видимости», т.е. если их можно соединить горизонтальной линией, не пересекающей никакую другую вертикальную линию. На третьем, заключительном этапе, сеть компактифицируется. Все узлы с одним и тем же словом объединяются в один узел. Все связи таких узлов также объединяются. Важно отметить, что между любыми двумя узлами при этом остается не более одной связи – кратные связи изымаются. В качестве весовых оценок отдельных слов в дальнейшем используются степени соответствующих им узлов в CHVG. После этого все термины текста сортируются по убыванию рассчитанных весовых значений соответствующих узлов CHVG. Дальнейшему анализу не подлежат термины из так называемого стоп-словаря, являющиеся важными для связности текста, но не несущие информационной нагрузки. Это, как правило, фиксированный набор служебных слов. Используемый в рамках данной работы стоп-словарь был построен на основе различных стоп-словарей, представленных в доступном виде на веб-ресурсах:

https://code.google.com/p/stop-words/downloads/list;

http://www.ranks.nl/stopwords/;

http://www.textfixer.com/resources/common-english-words.txt.

Экспертным методом определяется необходимый размер СЕИТ (число N), после чего выбирается соответствующее количество единичных слов, биграмм и триграмм (всего N+N+N элементов) с наибольшими весовыми значениями по CHVG. Из отобранных терминов строятся сети естественных иерархий терминов, в которых как узлы рассматриваются сами термины, а связи соответствуют вхождением одних терминов в другие. На рис. 6

проиллюстрирован принцип построения связей СЕИТ. Следует отметить, что ранее этот алгоритм применялся к другим видам документов, в частности, докладам тематических конференций и реферативным базам данных [15].

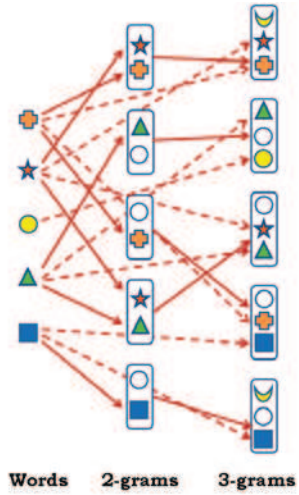

Рис. 6. Формирование связей в трехуровневой сети естественной иерархии терминов

Различные геометрические фигуры на этой иллюстрации соответствуют различным словам. Первой колонке соответствует выбранное множество единичных слов, второй – множество биграмм, а третьей – множество триграмм. Если единичное слово входит в биграмму или триграмму, или биграмма входит в триграмму, образуется связь, которая обозначается стрелкой. Множество узлов, которым соответствуют термины, и связи образуют трехуровневую сеть естественной иерархии терминов [11].

После формирования СЕИТ осуществляется ее отображение программными средствами анализа и визуализации графов. Для загрузки сетей естественных иерархий терминов в базы данным формируется матрица инцидентности общепринятого формата csv.

В таблице 1 приведены списки 20 наиболее весомых терминов (слов, биграмм и триграмм) из новостных сообщений, соответствующих сюжетной цепочке.

На рис. 7 представлена небольшая сеть естественной иерархии терминов размером 20+20+20, которая визуализирована средствами системы Gephi (https://gephi.org/).

На рис. 8 приведен фрагмент более крупной сети естественной иерархии терминов размером 200+200+200.

Для построенных сетей естественных иерархий терминов различных размеров по выбранному тексту было определено распределение исходящих степеней узлов, которое оказалось близким к степенному ($p(k) = Ck^{\alpha}$), т.е. эти сети являются безмасштабными. Оказалось, что коэффициент α для сетей различных размеров (от 20+20+20 до 500+500+500) составляет от 2,1 до 2,3 (рис. 9), что вполне соответствует сетям языка (Language Networks) [1].

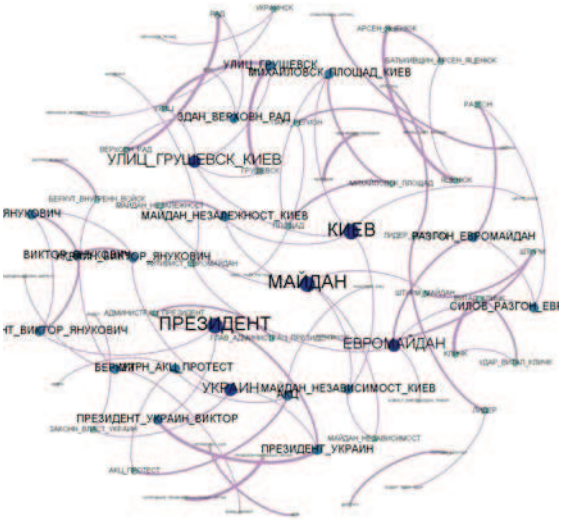

Рис. 7. Вид СЕИТ размером 20+20+20

Очевидно, что в соответствии с предложенным алгоритмом, максимальное количество входных связей для узлов данной сети составляет 5: для узлов из одного слова – 0 входящих связей, для узлов из 2 слов – максимально 2 связи, для узлов из 3 слов – максимально 5 связей – три связи от отдельных слов и две от пар слов. Топ-20 узлов с максимальной входной степенью для СЕИТ размером 200+200+200 приведен в таблице 2.

Наиболее интересными с семантической точки зрения в рассматриваемой СЕИТ оказались узлы с максимальным количеством входных связей, среди которых можно выделить такие словосочетания: «УЧАСТНИКИ АКЦИИ ПРОТЕСТА»; «УЛИЦА ГРУШЕВСКОГО КИЕВ»; «СИЛОВОЙ РАЗГОН ЕВРОМАЙДАНА»; «МИРНАЯ АКЦИЯ ПРОТЕСТА»; «БОЙЦЫ СПЕЦПОДРАЗДЕЛЕНИЯ БЕРКУТ».

По отдельным сюжетам также были рассчитаны значения CHVG для слов, биграмм и триграмм, построены сети естественных иерархий терминов. В качестве примера, отражающего направленность сюжетной цепочки, на рис. 10 приведена визуализация СЕИТ для трех выбранных массивов. Взаимосвязь терминов из новостей, входящих в состав выбранных сюжетов, приведена на рис. 11.

7 Релевантность отдельных сюжетов сюжетным цепочкам

На рис. 11 можно видеть, что каждому массиву (узлы, идентифициированные датами) соответствуют термины. При этом в центральной части сети располагаются термины, общие для нескольких дат (О-зона), а «гребешки» на периферии соответствуют специальным терминам, отражающим специфику конкретных сюжетов (С-зоны).

Таблица 1. ТОП-20 по значениям CHVG терминов

№	Слова	Биграммы	Триграммы
1	УКРАИНА	ВИКТОР ЯНУКОВИЧ	ПРЕЗИДЕНТ ВИКТОР ЯНУКОВИЧ
2	КИЕВ	ЦЕНТР КИЕВА	СОТРУДНИКИ ПРАВООХРАНИТЕЛЬНЫХ ОРГАНОВ
3	ВЛАСТЬ	ВЕРХОВНАЯ РАДА	ВВЕДЕНИЕ ЧРЕЗВЫЧАЙНОГО ПОЛОЖЕНИЯ
4	СТРАНА	УЛИЦА ГРУШЕВСКОГО	БАТЬКИВЩИНА АРСЕНИЙ ЯЦЕНЮК
5	ЯНУКОВИЧ	ПРЕЗИДЕНТ УКРАИНЫ	ОЛИМПИЙСКИЕ ИГРЫ СОЧИ
6	МАЙДАН	МАЙДАН НЕЗАВИСИМОСТИ	ГЛАВА АДМИНИСТРАЦИИ ПРЕЗИДЕНТА
7	ЛЮДИ	ПАРТИЯ РЕГИОНОВ	ФРАКЦИЯ ПАРТИИ РЕГИОНОВ
8	МИЛИЦИЯ	ПРЕСС-СЛУЖБА	ШТАБ НАЦИОНАЛЬНОГО СОПРОТИВЛЕНИЯ
9	БЕРКУТ	АРСЕНИЙ ЯЦЕНЮК	ДЕЙСТВИЕ БЛАГОДАТИ ПРЕСВЯТОЙ
10	ОППОЗИЦИЯ	МИХАЙЛОВСКАЯ ПЛОЩАДЬ	МАЙДАН НЕЗАЛЕЖНОСТИ КИЕВ
11	ПРЕЗИДЕНТ	ЛИДЕРЫ ОППОЗИЦИИ	СТРАНИЦЫ СОЦИАЛЬНЫХ СЕТЕЙ
12	ЯЦЕНЮК	РАЗГОН ЕВРОМАЙДАНА	УДАР ВИТАЛИЙ КЛИЧКО
13	УКРАИНСКИЙ	ОБЪЯВЛЕНИЕ ПЕРЕМИРИЯ	ГЕРМАНИЯ ФРАНЦИЯ ВЕЛИКОБРИТАНИЯ
14	ЕВРОМАЙДАН	ВИТАЛИЙ КЛИЧКО	УЛИЦА ГРУШЕВСКОГО КИЕВ
15	ШТУРМ	МАЙДАН НЕЗАЛЕЖНОСТИ	ОФИС ПАРТИИ РЕГИОНОВ
16	АКЦИЯ	АКЦИЯ ПРОТЕСТА	МИХАЙЛОВСКАЯ ПЛОЩАДЬ КИЕВ
14	ЗДАНИЕ	ПРАВЫЙ СЕКТОР	СИЛОВОЙ РАЗГОН ЕВРОМАЙДАНА
15	АКТИВИСТ	ОГНЕСТРЕЛЬНОЕ ОРУЖИЕ	БЕРКУТ ВНУТРЕННИЕ ВОЙСКА
16	МВД	ПРАВООХРАНИТЕЛЬНЫЕ ОРГАНЫ	ПРЕМЬЕР НИКОЛАЙ АЗАРОВ
17	ПЛОЩАДЬ	ШТУРМ ЗАЧИСТКА	МИРНАЯ АКЦИЯ ПРОТЕСТА
18	УЛИЦА	ШТУРМ МАЙДАНА	ЗДАНИЕ ВЕРХОВНОЙ РАДЫ
19	ГРУШЕВСКОГО	ВНУТРЕННИЕ ВОЙСКА	ЗАКОННАЯ ВЛАСТЬ УКРАИНЫ
20	ЛИДЕР	ПРИМЕНЕНИЕ СИЛЫ	ЛИДЕР ПАРТИИ УДАР

Таблица 2. Топ-20 узлов с максимальной входной степенью

№	Выходная степень	Узел
1	5	УЧАСТНИКИ АКЦИИ ПРОТЕСТА
2	5	УЛИЦА ГРУШЕВСКОГО КИЕВ
3	5	(ПРЕЗИДЕНТ) УКРАИНЫ ВИКТОР ЯНУКОВИЧ
4	5	СИЛОВОЙ РАЗГОН ЕВРОМАЙДАНА
5	5	МИРНАЯ АКЦИЯ ПРОТЕСТА
6	5	ГЛАВА АДМИНИСТРАЦИИ ПРЕЗИДЕНТА
7	5	ФРАКЦИЯ ПАРТИИ РЕГИОНОВ
8	5	БОЙЦЫ СПЕЦПОДРАЗДЕЛЕНИЯ БЕРКУТ
9	5	БАТЬКИВЩИНА АРСЕНИЙ ЯЦЕНЮК
10	4	АДМИНИСТРАЦИЯ ПРЕЗИДЕНТА УКРАИНЫ
11	4	ЗДАНИЕ ВЕРХОВНОЙ РАДЫ
12	4	ЗДАНИЯ ЦЕНТРА КИЕВА
13	4	ВЕРХОВНАЯ РАДА УКРАИНЫ
14	4	УДАР ВИТАЛИЙ КЛИЧКО
15	4	СОТРУДНИКИ СПЕЦПОДРАЗДЕЛЕНИЯ БЕРКУТ
16	4	СОТРУДНИКИ ПРАВООХРАНИТЕЛЬНЫХ ОРГАНОВ
17	4	СИЛОВОЙ РАЗГОН МИТИНГУЮЩИХ
18	4	ПОЛИТИЧЕСКИЙ КРИЗИС УКРАИНА
19	4	ПРИМЕНЕНИЕ СИЛЫ СТОРОНАМИ
20	4	ПРЕСС-СЛУЖБА МВД

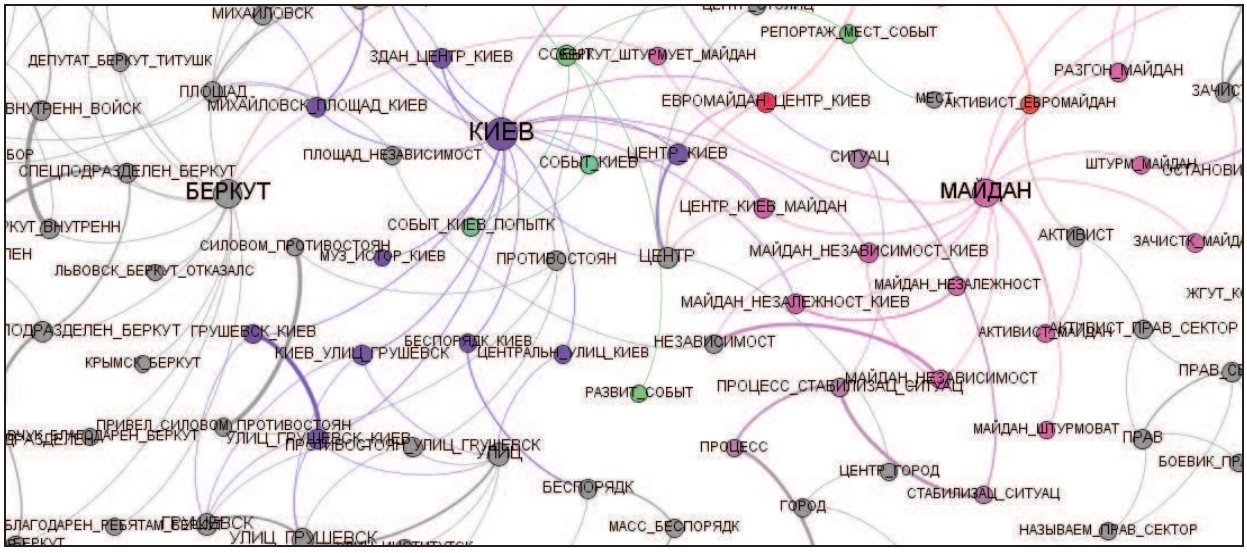

Рис. 8. Фрагмент СЕИТ (визуализация средствами Gephi)

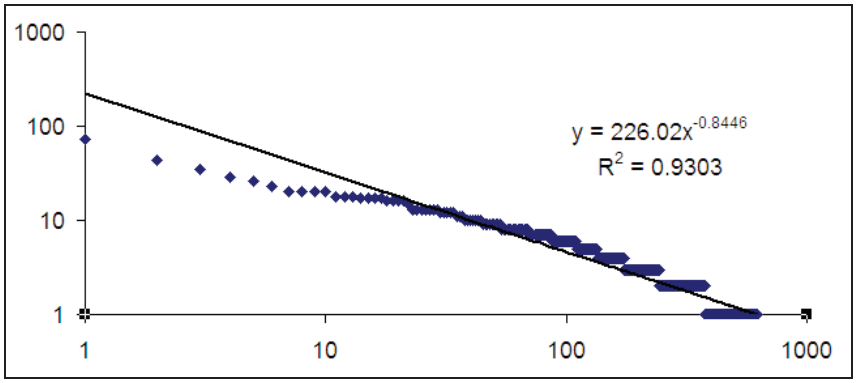

Рис. 9. Ранговое распределение степеней узлов в логарифмической шкале (по оси абсцисс – порядковый номер узла, по оси ординат – степень узла)

а б в

Рис. 10. СЕИТ размером 20+20+20 по массивам (а – 2013.11.30, б – 2014.01.22, в – 2014.02.19)

О-зона не обязательно включает термины из всех сюжетов, достаточно, чтобы термины соответствовали лишь их определенной части (порогу). Чем в сообщениях сюжета больше терминов, попадающих О-зону, тем он лучше вписывается в тематику всей сюжетной цепочки, тем он точнее попадает в ее тренд. В данном случае (рис. 11) именно сюжет 22 января наиболее точно соответствует тематическому направлению всей сюжетной цепочки.

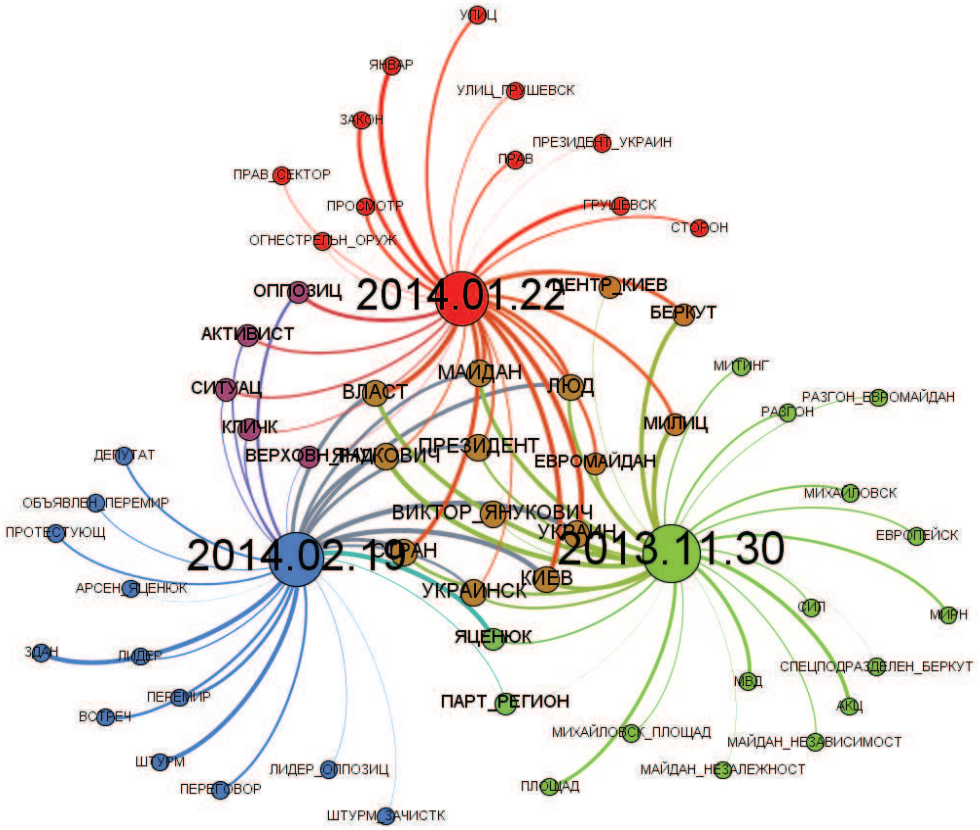

Рис. 11. Сеть связи терминов выбранных событий

Таким образом, можно предложить такой лингвистический критерий релевантности сюжета общей сюжетной цепочке: чем большая часть лексики из него попадает в О-зону, тем он более релевантен. Формально значение этого критерия $k_{i,N}$ для сюжета i сюжетной цепочки s может быть записано:

$$k_{i,N} = \frac{\left|T_{i,N} \cap T_{s,N}\right|}{3N},$$

где N – параметр СЕИТ (количество слов, биграмм и триграмм), $T_{i,N}$ – множество значимых терминов сюжета i, $T_{s,N}$ – множество значимых терминов всей сюжетной цепочки.

Представления об информационной значимости наборов терминов для построения СЕИТ, степени их важности для отражения смысла сюжетной цепочки были подтверждены в ходе экспериментов с информантами. Так, для всех сюжетов были проведены эксперименты с вариантами стандартной инструкции «Вспомните сюжет. Подумайте над его содержанием. Выпишите 10-15 слов, наиболее важных для его содержания» (более 20 информантов для каждого сюжета) [7, 8[1]]. При этом количество предложенных экспертами отдельных слов, биграмм и триграмм, характеризующих общий сюжет, совпало с данными табл. 1 на 65, 50 и 45 %, соответственно. Вместе с тем, качество предложенной терминологической сети исследовалось лишь на уровне экспертных оценок, сравнение СЕИТ с другими подобными сетями остается открытым вопросом.

Выводы

Таким образом, в результате проведенных исследований:

- Описан алгоритм построения СЕИТ на основе анализа текстов новостных сообщений.
- При построении СЕИТ для новостных сообщений был учтен ряд особенностей, связанных с предварительным анализом потока новостей, предложен подход к выбору репрезентативных сюжетов.
- На основании предложенного алгоритма построена СЕИТ (как показала практика, минимальный объем для построения репрезентативной сети составляет около 20 КБ).
- Сеть естественных иерархий терминов оказалась скейл-фри по исходящим связям.
- Выбраны программные средства, позволяющие решать задачу визуализации СЕИТ. При этом задача выбора лучшего алгоритма визуализации не ставилась.
- Предложен критерий релевантности сюжета общей сюжетной цепочке.

[1] Ср. основную инструкцию сходного эксперимента: «Как можно детальнее вспомните события сравнительно недавнего периода, примерно от 4 декабря 2011 г. до 4 марта 2012 г. (от выборов в Государственную Думу до подведения итогов по выборам президента). Напишите 10-15 слов или словосочетаний, относящихся к этим событиям».

Сеть языка, построенную с помощью предложенной методики, можно использовать в качестве базы для построения общей онтологии по выбранной тематике, готового к применению средства навигации в базах данных, а также для организации контекстных подсказок пользователям информационно-поисковых систем.

Литература

[1] Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Е.И. Большакова, Э.С. Клышинский , Д.В. Ландэ , А.А. Носков, О.В. Пескова, Е.В. Ягунова. – М.: МИЭМ, 2011. – 272 с.

[2] А.А. Давыдов. Системная социология. – М.: Издательство ЛКИ, 2008. – 192 с.

[3] А.Г. Додонов, Д.В. Ландэ. Моделирование и анализ тематических информационных потоков // Информационное противодействие угрозам терроризма, 2013. – № 20. – С. 52–59.

[4] И.В. Крылова, Л.М. Пивоварова, А.В. Савина, Е.В. Ягунова. Исследование новостных сегментов российской «снежной революции»: вычислительный эксперимент и интуиция лингвистов // Понимание в коммуникации: Человек в информационном пространстве: сб. научных трудов: в 3 т. – Яр.-М.: Изд-во ЯГПУ, 2012. – Т. 1. – С. 377–382.

[5] Н.В. Лукашевич, Б.В. Добров, Д.С. Чуйко. Отбор словосочетаний для словаря системы автоматической обработки текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2008». – М., 2008. – С. 339–344.

[6] Ю.Н. Филиппович, А.В. Прохоров. Семантика информационных технологий: Опыты словарно-тезаурусного описания. – М.: МГУП, 2002. – 368 с.

[7] Е.В. Ягунова. Эксперимент и вычисления в анализе ключевых слов художественного текста // Сборник научных трудов кафедры иностранных языков и философии ПНЦ УрО РАН. Философия языка. Лингвистика. Лингводидактика. – Пермь, 2010. – Вып. 1. – С. 85–91.

[8] Е.В. Ягунова, И.В. Крылова, О.Е. Макарова, Л.М. Пивоварова. "Снежная революция в России": значимые номинации, события, оценки (оценка событий информантами и данные СМИ) // "Мы не немы!": творчество протестующей улицы. – М., 2014.

[9] Е.В. Ягунова, А.В. Антонов. Методика работы с коллекциями текстовой информации через анализ информационных портретов // Труды 12-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010.

[10] D.V. Lande, A.A. Snarskii. Compactified HVG for the Language Network // International Conference on Intelligent Information Systems: The Conference is dedicated to the 50th anniversary of the Institute of Mathematics and Computer Science, 20–23 aug. 2013, Chisinau, Moldova: Proceedings IIS / Institute of Mathematics and Computer Science, 2013. – P. 108–113.

[11] D.V. Lande. Building of Networks of Natural Hierarchies of Terms Based on Analysis of Texts Corpora // E-preprint arXiv 1405.6068.

[12] D.V. Lande, A.A. Snarskii, E.V. Yagunova, E.V. Pronoza. The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text // 12th Mexican International Conference on Artificial Intelligence, 2013. – P. 209–215.

[13] B. Luque, L. Lacasa, F. Ballesteros, J. Luque. Horizontal visibility graphs: Exact results for random time series // Phys. Review E, 2009. – P. 046103-1–046103-11.

[14] G. Salton, M.J. McGill. Introduction to Modern Information Retrieval. – New York : McGraw-Hill, 1983. – 448 p.

[15] E. Yagunova, D. Lande. Dynamic Frequency Features as the Basis for the Structural Description of Diverse Linguistic Objects // CEUR Workshop Proceedings. Proceedings of the 14th All-Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections" – Pereslavl-Zalessky, Russia, 2012. – P. 150–159.

Network of Natural Hierarchies of Terms of News Messages on the "Euromaydan" Events

Dmitri V. Lande, Andrew A. Snarskii,
Elena V. Jagunova

The technique of building of networks of hierarchies of terms based on the analysis of scientific texts is offered. The technique is based on the methodology of horizontal visibility graphs for the terms – of individual words, bigrams and trigrams, as well as of an inclusion relationships between the terms. The network formed on the basis of news texts on the "Euromaydan" events has been designed and investigated.