УДК 001.8:004.7

ПОДХОД К СОЗДАНИЮ ТЕРМИНОЛОГИЧЕСКИХ ОНТОЛОГИЙ

Д.В. Ландэ[1], А.А. Снарский[2]

Институт проблем регистрации информации НАН Украины, г. Киев, Украина
[1]*dwlande@gmail.com*, [2]*asnarskii@gmail.com*

Аннотация

Описывается методика построения сети естественных иерархий терминов на основе анализа массива текстов по выбранной проблематике. Данная сеть формируется в автоматическом режиме на основе обучающей коллекции текстов и может рассматриваться как основа для построения терминологических онтологий. Методика базируется на применении компактифицированных графов горизонтальной видимости для терминов – отдельных слов, биграмм и триграмм, а также на установлении связей между терминами. Предложенная авторами сеть естественных иерархий терминов охватывает связи типа «общее-частное» и может рассматриваться как основа построения сети с ассоциативными связями. Рассмотрена сеть естественных иерархий терминов, сформированная на основе полных текстов научно-популярных статей. Предложено использование алгоритма HITS для данной сети, с помощью которого обеспечивается выбор наилучших «авторов» – узлов, на которые введут ссылки, и «посредников» – узлов, от которых идут ссылки цитирования.

Ключевые слова: *языковая сеть, сеть иерархии терминов, текстовый корпус, контекстные связи, ассоциативные связи*

Введение

Для решения задачи построения терминологической онтологии предметной области требуется проведение комплексных исследований, определённым этапом которых является построение так называемых словарных номенклатур, предметных словарей, тезаурусов. Эффективный автоматический отбор отдельных терминов для таких конструкций – нерешенная окончательно задача, а проблема установления связей, автоматического построения сетей из таких терминов до сих пор остается открытой.

Как терминологическую основу для формирования онтологии предлагается использовать сеть естественной иерархии терминов, которая базируется на информационно-значимых элементах текста [1], методология выявления которых приведена в [2]. Опорные термины, как правило, выбираются с учетом такого свойства, как дискриминантная сила. Однако одного этого свойства часто недостаточно для качественного отражения содержания предметной области. Иногда слова с низкой дискриминантной силой, например, наиболее частотные слова из выбранной предметной области (например, слова «Android», «IOS», «Приложение» в корпусе текстов по тематике современных гаджетов) оказываются важнейшими для рассматриваемой задачи.

1 Постановка задачи

Как подход к решению актуальной задачи построения терминологической онтологии, в данной работе рассматриваются принципы и методика формирования сети естественных иерархий терминов (СЕИТ), базирующейся на контенте научно-популярных статей выбранной направленности [3, 4]. «Естественность» иерархий терминов в этом случае понимается как отказ при формировании сети от методов смыслового анализа текстов, ограничиваясь факти-

чески статистическим анализом. Связи в такой сети определяются естественным взаимным положением слов и словосочетаний из текстов. Такая сеть, создаваемая полностью автоматически, может рассматриваться как основа для дальнейшего автоматизированного формирования терминологической онтологии с участием экспертов.

2 Методы решения

2.1 Формирование сети естественных иерархий терминов

Методика формирования сети естественных иерархий терминов, представленная в данной работе, предусматривает реализацию последовательности шагов, которые рассмотрим подробно.

1) на первом этапе формируется исходный текстовый корпус. Как пример такого корпуса рассматриваются полные тексты научно-популярных статей, опубликованных на вебсайте «Компьютерра онлайн» (http://www.computerra.ru), посвящённых проблематике мобильных устройств, представленных на русском языке. В состав текстового корпуса было включено около 230 статей общим объёмом свыше 800 тыс. символов. Предварительная обработка такого текстового корпуса предусматривала выделение фрагментов текстов (отдельных статей, абзацев, предложений, слов), исключение нетекстовых символов, отсечение флективных окончаний.
2) на втором этапе каждому отдельному термину из текста (слову, биграмме или триграмме) ставится в соответствие оценка их «дискриминантной силы» (TFIDF[1]), которая в каноническом виде впервые была предложена Г. Солтоном. Эта оценка равна произведению частоты соответствующего термина (Term Frequency) во фрагменте текста и двоичного логарифма величины, обратной к количеству фрагментов текста, в которых этот термин встретился (Inverse Document Frequency) [5]. Для последовательностей терминов и их весовых значений по TFIDF строятся компактифицированные графы горизонтальной видимости (КГГВ) и выполняется переопределение весовых значений слов уже по этому алгоритму. Данная процедура позволяет учитывать в дальнейшем, кроме терминов с большой дискриминантной силой, также высокочастотные термины, которые имеют большое значение для общей тематики. В соответствии с [4], сеть слов с использованием алгоритма горизонтальной видимости строится также 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 [6]. 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, связи таких узлов также объединяются. В качестве весовых оценок отдельных слов в дальнейшем используются степени соответствующих им узлов в КГГВ. После этого все термины текста сортируются по убыванию рассчитанных весовых значений соответствующих узлов КГГВ. 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:

[1] TFIDF (от англ. *TF — term frequency, IDF — inverse document frequency*)

- *http://code.google.com/p/stop-words/source/browse/trunk/stop-words/stop-words/stop-words-russian.txt?spec=svn3&r=3;*
- *https://github.com/punbb/langs/blob/master/Russian/stopwords.txt;*
- *http://www.ranks.nl/stopwords/russian.html;*
- *http://trac.mysvn.ru/punbb/punbb/browser/trunk/Russian/stopwords.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.

3) из отобранных терминов строятся сети естественных иерархий терминов, в которых как узлы рассматриваются сами термины, а связи соответствуют вхождениям одних терминов в другие. На рисунке 1 проиллюстрирован принцип построения связей СЕИТ. Различные геометрические фигуры на этой иллюстрации соответствуют различным словам. Первой строке 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.

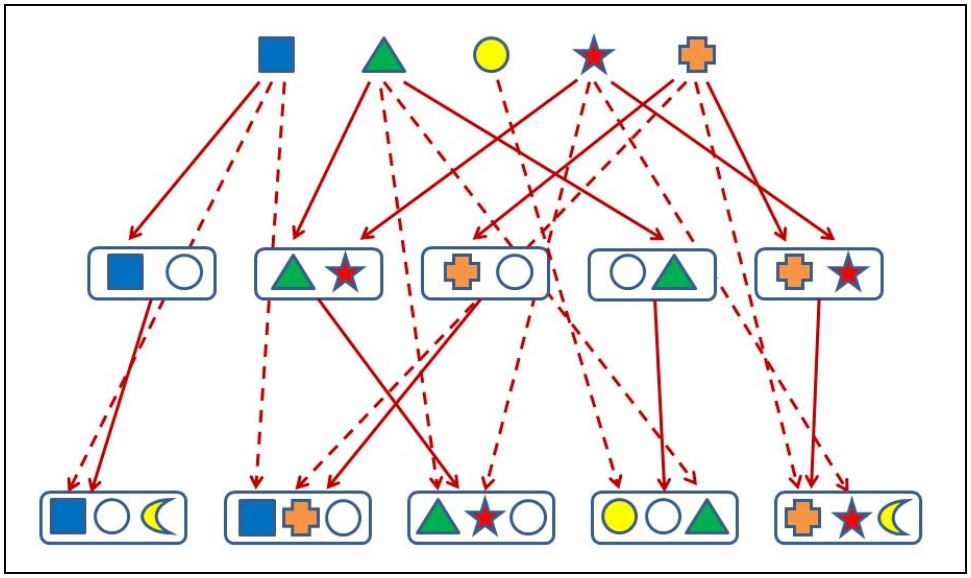


Рисунок 1 – Трехуровневая сеть естественной иерархии терминов

После формирования СЕИТ (построения матрицы инцидентности) осуществляется её отображение программными средствами анализа и визуализации графов. Для загрузки сетей естественных иерархий терминов в базы данных формируется матрица инцидентности общепринятого формата CSV[2] размерностью $(N+N+N) \times (N+N+N)$ элементов.

На рисунке 2 представлена небольшая сеть естественной иерархии терминов размером 30+30+30, которая визуализирована средствами системы Gephi (https://gephi.org/).

На рисунке 3 приведены отдельные фрагменты более крупной сети естественной иерархии терминов размером 200+200+200.

[2] Текстовый формат **CSV** (от англ. *Comma-Separated Values* — значения, разделённые запятыми) — формат для представления табличных данных. Значения отдельных столбцов разделяются символами: запятой, точкой с запятой или двоеточием. Под CSV, как правило, понимают набор значений, разделенных какими угодно разделителями, в какой угодно кодировке с какими угодно окончаниями строк.

2.2 Ранжирование узлов СЕИТ

Ранжирование узлов в СЕИТ возможно также по свойствам, обуславливаемым сетевой структурой, связями. Например, для определения авторитетности узла как слова – источника порождения словосочетаний или как составного термина, состоящего из отдельных важных слов, можно анализировать СЕИТ, выбирая при этом наиболее важных «авторов» или «посредников». Для решения этой задачи предлагается использовать известный алгоритм ранжирования веб-страниц, основанных на связях - HITS (hyperlink induced topic search), предложенный Дж. Клейнбергом [7].

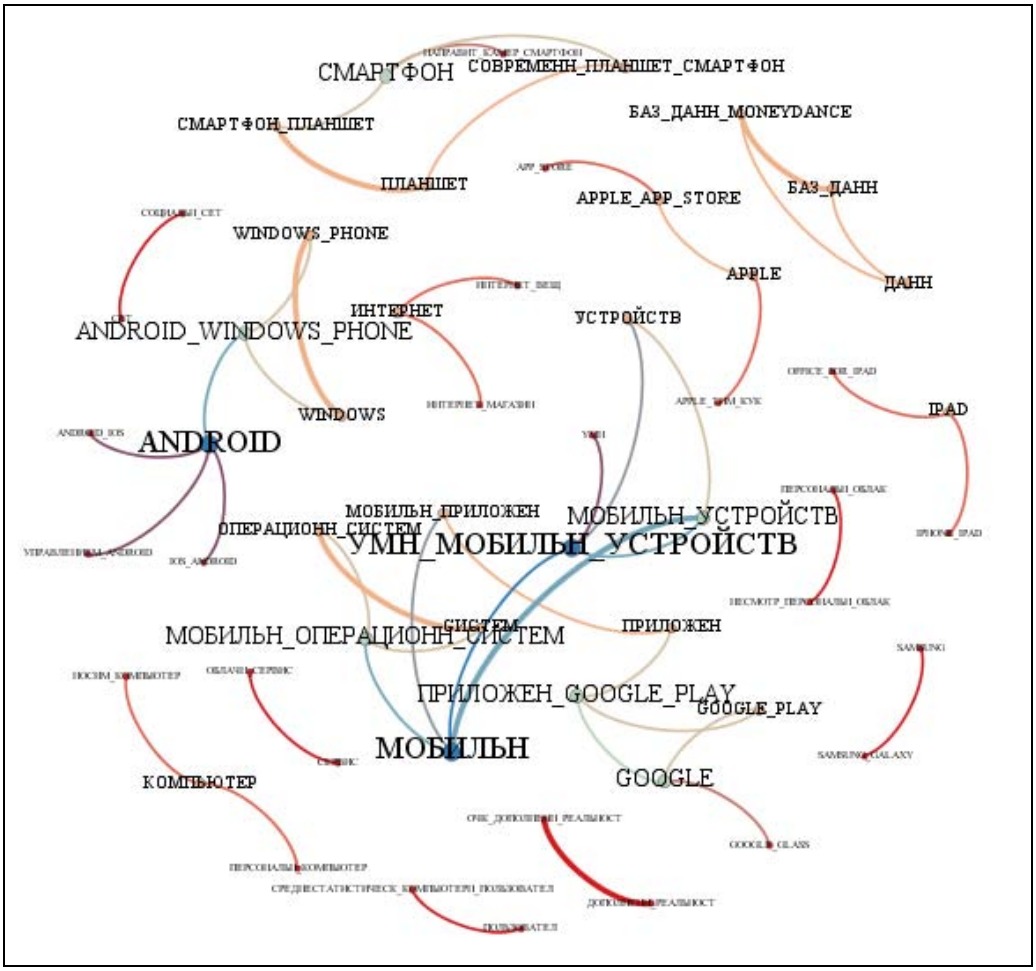


Рисунок 2 – Визуализация связного фрагмента СЕИТ размером 30+30+30

Алгоритм HITS обеспечивает выбор из информационного массива лучших «авторов» (узлов, на которые введут ссылки) и «посредников» (узлов, от которых идут ссылки цитирования). Понятно, что в нашем случае термин является хорошим посредником, если от него идут связи на важные словосочетания, и наоборот, термин (словосочетание) является хорошим автором, если на него ведут связи от важных авторов. В соответствии с алгоритмом HITS в нашем случае для каждого узла сети v_j рекурсивно вычисляется его значимость как автора $a(v_j)$ и посредника $h(v_j)$ по формулам:

$$a(v_j) = \sum_i h(v_i);\ h(v_j) = \sum_i a(v_i).$$

В данных формулах суммирование производится по всем узлам, которые ссылаются (или на которые ссылаются – во второй формуле) на данный узел.

Наиболее интересными с семантической точки зрения в рассматриваемой СЕИТ оказались узлы с наибольшим значением авторства и посредничества. В таблице 1 приведены термины, соответствующие таким узлам.

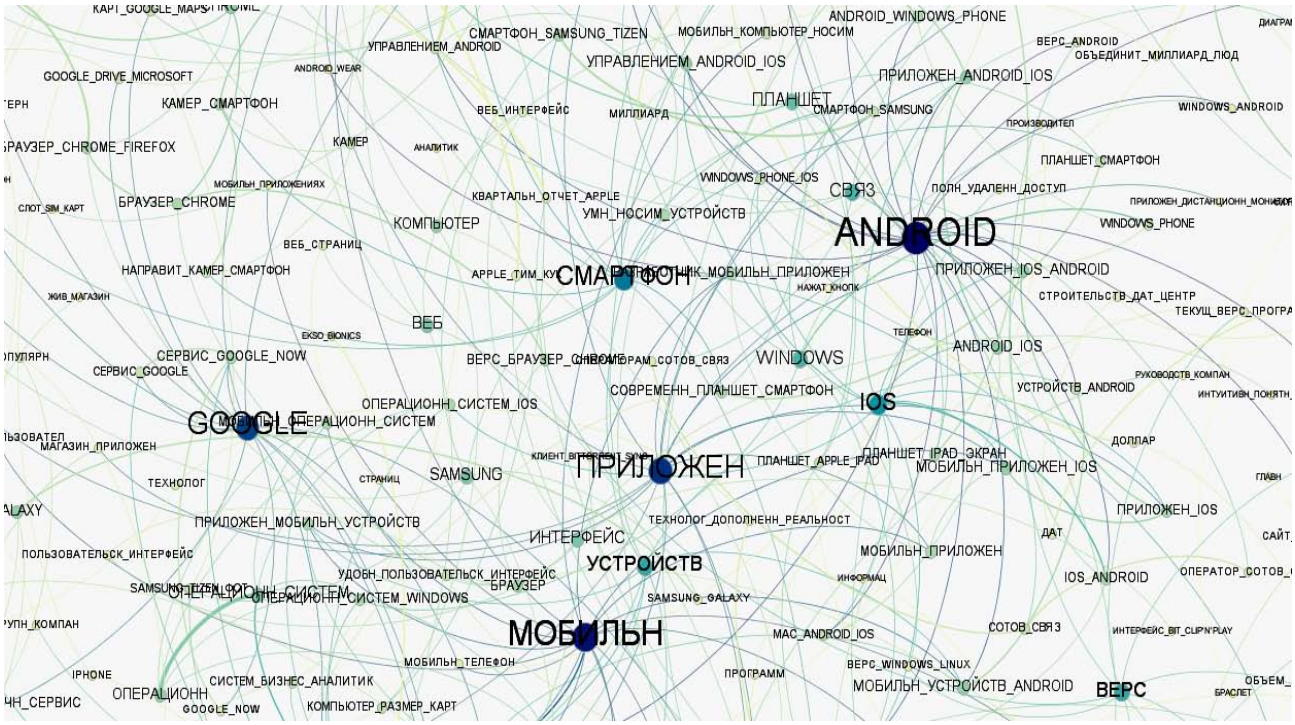


Рисунок 3 – Фрагмент СЕИТ размером 200+200+200

Таблица 1. Термины, соответствующие узлам с наибольшим авторством и посредничеством

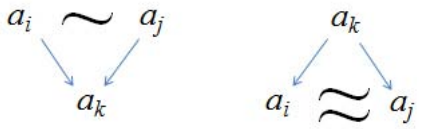
№	Термины с наибольшим значением посредничества	Термины с наибольшим значением авторства
1	МОБИЛЬНЫЙ	ПРИЛОЖЕНИЕ МОБИЛЬНОГО УСТРОЙСТВА
2	ANDROID	МОБИЛЬНОЕ ПРИЛОЖЕНИЕ IOS
3	IOS	МОБИЛЬНАЯ ОПЕРАЦИОННАЯ СИСТЕМА
4	СИСТЕМА	ОПЕРАЦИОННАЯ СИСТЕМА ANDROID
5	УСТРОЙСТВО	ПОЛЬЗОВАТЕЛЬ МОБИЛЬНОГО УСТРОЙСТВА
6	ОПЕРАЦИОННАЯ	МОБИЛЬНОЕ ПРИЛОЖЕНИЕ
7	ОПЕРАЦИОННАЯ СИСТЕМА	УМНОЕ МОБИЛЬНОЕ УСТРОЙСТВО
8	МОБИЛЬНОЕ УСТРОЙСТВО	ОПЕРАЦИОННАЯ СИСТЕМА IOS
9	WINDOWS	ПРИЛОЖЕНИЕ ANDROID
10	ПОЛЬЗОВАТЕЛЬ	ОПЕРАЦИОННАЯ СИСТЕМА WINDOWS
11	УМНЫЙ	МОБИЛЬНОЕ УСТРОЙСТВО
12	ПРИЛОЖЕНИЕ ANDROID	МОБИЛЬНЫЙ ПОЛЬЗОВАТЕЛЬ
13	МОБИЛЬНОЕ ПРИЛОЖЕНИЕ	ПРИЛОЖЕНИЕ GOOGLE_PLAY
14	GOOGLE	ANDROID WINDOWS PHONE
15	ВЕРСИЯ	МОБИЛЬНЫЙ ТЕЛЕФОН

Представления об информационной значимости наборов терминов для построения СЕИТ, степени их важности для отражения смысла научного текста были подтверждены в

ходе экспериментов с информантами. Так, для всех текстов были проведены эксперименты со стандартной инструкцией «Прочитайте текст. Подумайте над его содержанием. Выпишите 10-15 слов, наиболее важных для его содержания» [8].

2.3 Выявление ассоциативных связей

Рассматриваемые в предложенной модели СЕИТ связи являются направленными и могут рассматриваться как отношения «общее-частное» при построении общей онтологии. Вместе с тем, построенная сеть СЕИТ может рассматриваться как основа для формирования других связей между её узлами. Если обозначить матрицу инцидентности СЕИТ буквой A, то матрицы AA^T и A^TA будут отражать связи вхождения таких типов: если два термина-узла данной сети a_i и a_j порождают третий термин a_k, то будем считать, что такие термины связаны ассоциативной связью, назовем ее ассоциативной связью первого рода (рисунок 5а); если два термина-узла данной сети a_i и a_j порождаются третьим термином a_k, который также входит в данную сеть, то будем считать, что такие термины связаны ассоциативной связью второго рода (рисунок 5б).



а) б)

Рисунок 5 – Ассоциативные связи, построенные по СЕИТ, со связями, обозначенными стрелками: а) первого рода «~»; б) второго рода «≃»

На рисунках 6 и 7 приведены фрагменты сети СЕИТ с ассоциативными связями. В частности, на рисунке 6 жирными кривыми обозначены ассоциативные связи между словами «мобильное» и «приложение», что объясняется наличием общего узла-термина «мобильное приложение».

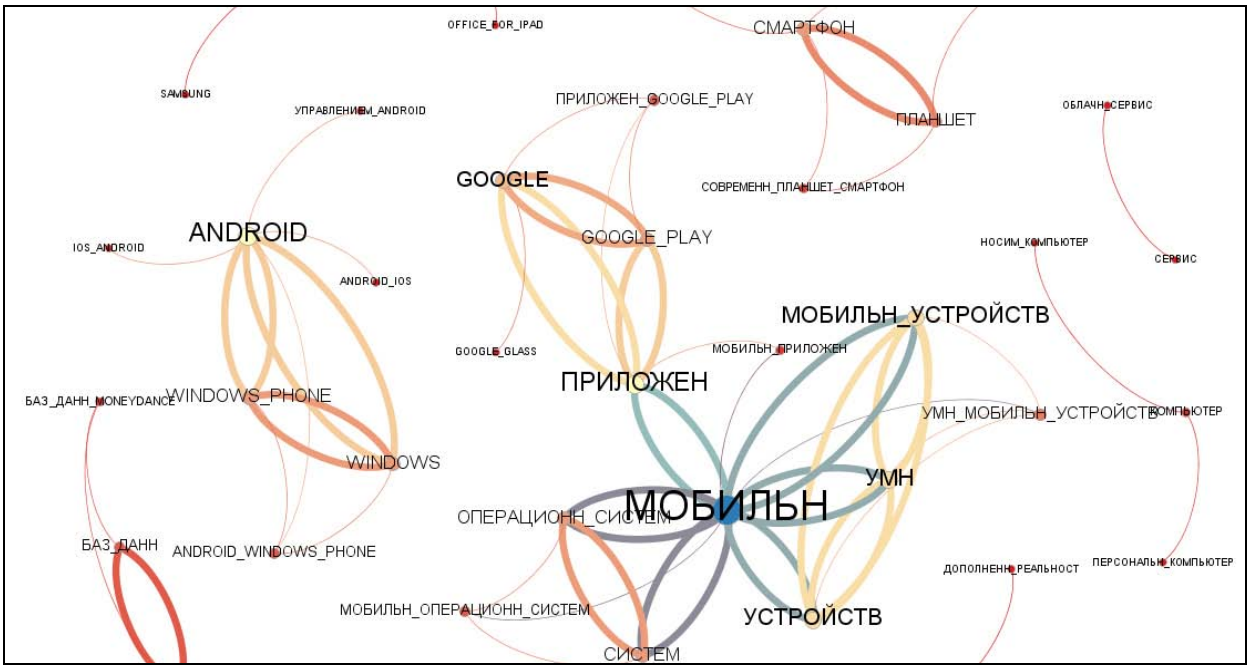


Рисунок 6 - Фрагмент СЕИТ размером 30+30+30 с ассоциативными связями 1-го рода

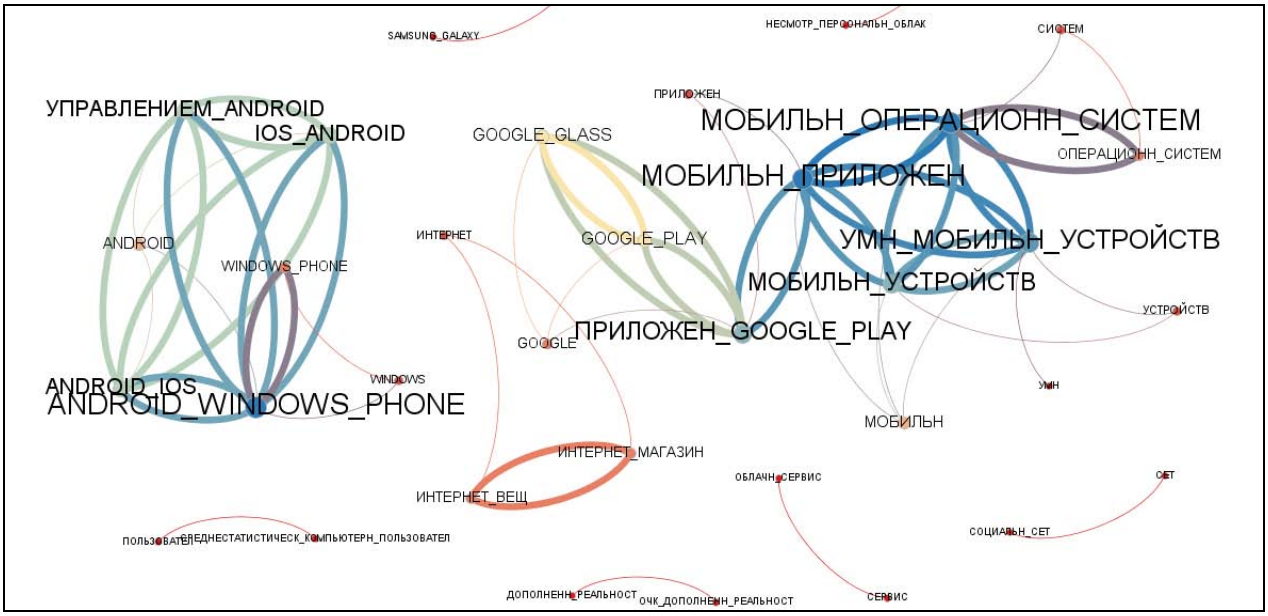


Рисунок 7 - Фрагмент СЕИТ размером 30+30+30 с ассоциативными связями 2-го рода

Заключение

Таким образом, в данной статье:

- предложен алгоритм построения СЕИТ на основе анализа текстовых корпусов;
- на основании этого алгоритма по текстам научных статей по проблематике мобильных устройств построена сеть естественной иерархии терминов;
- предложен и обоснован алгоритм построения ассоциативных связей 1-го и 2-го рода между терминами в СЕИТ;
- предложено использование алгоритма HITS для выбора важнейших элементов СЕИТ;
- выбраны программные средства визуализации СЕИТ.

Сеть языка, автоматически построенную с помощью предложенного алгоритма с использованием относительно небольшого тематического текстового корпуса, можно использовать в качестве основы для построения онтологии предметной области (в рассмотренном примере – по проблематике мобильных устройств). Кроме того, данную СЕИТ можно использовать на практике в качестве готового к применению средства навигации в информационных массивах, а также для организации контекстных подсказок пользователям информационно-поисковых систем.

Список источников

[1] ***Yagunova E., D. Lande D.*** 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, October 15-18, 2012. – P. 150-159.

[2] ***Lande D.V., Snarskii A.A., Yagunova E.V., Pronoza E.V.*** 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.

[3] ***Lande D.V., Snarskii A.A.*** Compactified Horizontal Visibility Graph for the Language Network // E-preprint ArXiv 1302.4619. - http://poiskbook.kiev.ua/art/arxiv1302.4619/chvg.pdf

[4] ***Lande D.V.*** Building of Networks of Natural Hierarchies of Terms Based on Analysis of Texts Corpora // E-preprint ArXiv 1405.6068 - http://dwl.kiev.ua/art/arxiv1405.6068/1405.6068.pdf
[5] ***Salton G., McGill M.J.*** Introduction to Modern Information Retrieval. – New York: McGraw-Hill, 1983. – 448 p.
[6] ***Luque B., Lacasa L., Ballesteros F., Luque J.*** Horizontal visibility graphs: Exact results for random time series // Phys. Review E, 2009. – P. 046103-1 – 046103-11.
[7] ***Kleinberg J.*** Authoritative sources in a hyperlinked environment // In Processing of ACM-SIAM Symposium on Discrete Algorithms, 1998, 46(5):604-632.
[8] ***Ягунова, Е.В.*** Эксперимент и вычисления в анализе ключевых слов художественного текста / Е.В. Ягунова // Сборник научных трудов кафедры иностранных языков и философии ПНЦ УрО РАН. Вып. 1: Философия языка. Лингвистика. Лингводидактика. – Пермь, 2010. — С. 85-91.

APPROACH TO THE CREATION OF TERMINOLOGICAL ONTOLOGIES

D.V. Lande[1], A.A. Snarskii[2]

Institute of data recording problems NAN Ukraine, Kiev, Ukraine
[1]*dwlande@gmail.com,* [2]*asnarskii@gmail.com*

The technique for creating networks of natural hierarchies of terms based on the analysis of chosen sets of texts on selected issues is offered. The network is formed automatically on the basis of the teaching collection of texts and can be considered as the basis for the design of terminological ontologies. The technique is based on the methodology of horizontal visibility graphs for individual words, bigrams and trigrams, as well as establishing links between the terms. The network of natural hierarchies of terms covers connection "general-private" type and can be considered as a basis of creation of networks with associative links. Designed and investigated language network, formed on the basis of full texts of popular scientific papers is reviewed. Use of HITS algorithm for this network is proposed. The named algorithm makes the choice of the best "authors" – nodes that have the most citations, and "intermediaries" – nodes that establish the biggest number of citation links is offered.
Key words: *Language network, Networks of hierarchies of terms, vitality, Text corpora, Contextual links, Associative links, Terminological ontologies.*

References

[1] ***Yagunova E., D. Lande D.*** 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, October 15-18, 2012. – P. 150-159.
[2] ***Lande D.V., Snarskii A.A., Yagunova E.V., Pronoza E.V.*** 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.
[3] ***Lande D.V., Snarskii A.A.*** Compactified Horizontal Visibility Graph for the Language Network // E-preprint ArXiv 1302.4619. - http://poiskbook.kiev.ua/art/arxiv1302.4619/chvg.pdf
[4] ***Lande D.V.*** Building of Networks of Natural Hierarchies of Terms Based on Analysis of Texts Corpora // E-preprint ArXiv 1405.6068 - http://dwl.kiev.ua/art/arxiv1405.6068/1405.6068.pdf
[5] ***Salton G., McGill M.J.*** Introduction to Modern Information Retrieval. – New York: McGraw-Hill, 1983. – 448 p.
[6] ***Luque B., Lacasa L., Ballesteros F., Luque J.*** Horizontal visibility graphs: Exact results for random time series // Phys. Review E, 2009. – P. 046103-1 – 046103-11.
[7] ***Kleinberg J.*** Authoritative sources in a hyperlinked environment // In Processing of ACM-SIAM Symposium on Discrete Algorithms, 1998, 46(5):604-632.
[8] ***Yagunova*** E.V. Ehksperiment i vychisleniya v analize klyuchevykh slov khudozhestvennogo teksta [Experiment and calculations in the keywords analysis of an artistic text] // Sbornik nauchnykh trudov kafedry inostrannykh yazykov i filosofii PNTS UrO RAN. Filosofiya yazyka. Lingvistika. Lingvodidaktika. [Collection of scientific papers of the department of foreign languages of PSC UrD RAS. Philosophy of language. Linvistics. lingvodidactics]– Perm, 2010. – Issue 1. – pp. 85-91. (In Russian).

Сведения об авторах

Ландэ Дмитрий Владимирович, 1959 г. рождения. Окончил в 1981 г. механико-математический факультет Киевского государственного университета им. Т.Г. Шевченко, д.т.н. (2006). Заведующий отделом специализированных средств моделирования Института проблем регистрации информации НАН Украины, профессор Национального технического университета «Киевский политехнический институт». Член Российской ассоциации искусственного интеллекта, академик Российской академии естествознания (РАЕ). В списке научных трудов более 200 работ в области информационного поиска, динамики информационных потоков.

Dmitry Vladimirovich Lande (b. 1959) graduated from the Shevchenko Kiev State University, faculty of mechanics and mathematics in 1981, Dr. of Science (2006). He is department head of the Institute for data recording problems NAS Ukraine, professor at National Technical University of Ukraine “Kiev Polytechnic Institute”. He is a member of Russian Association for Artificial Intelligence (RAAI), full member of the Russian Academy of Natural History (RANH). He is co-author of over 200 scientific articles, books and abstracts in the field of information retrieval and information flows dynamics.

Снарский Андрей Александрович, 1949 г. рождения. Окончил в 1972 г. физический факультет Черновицкого государственного университета, д.ф-м.н. (1991). Профессор кафедры общей и теоретической физики Национального технического университета Украины «Киевский политехнический институт». Академик Международной академии термоэлектричества. Член-корреспондент Международной академии холода. В списке научных трудов более 200 работ в области термоэлектричества, теории протекания, методов детерминированного хаоса.

Andrei Alexandrovich Snarskii (b. 1949) graduated from the Chernovtsi State University, faculty of physics in 1972, Dr. of Science (1991). He is professor at National Technical University of Ukraine “Kiev Polytechnic Institute”, Department of General and Theoretical Physics. He is an International Academy of thermoelectricity (IAT) full member, corresponding member of the International Academy of Refrigeration (IAR). He is co-author of over 200 scientific articles, books and abstracts in the field of thermoelectricity, percolation theory, methods of deterministic chaos.