BUILDING OF NETWORKS OF NATURAL HIERARCHIES OF TERMS BASED ON ANALYSIS OF TEXTS CORPORA

D.V. Lande, Institute of Data Recording Problems, NAS Ukraine

Summary. *The technique of building of networks of hierarchies of terms based on the analysis of chosen text corpora is offered. The technique is based on the methodology of horizontal visibility graphs. Constructed and investigated language network, formed on the basis of electronic preprints arXiv on topics of information retrieval.*

Keywords: *language network, hierarchies of terms, electronic preprint, visibility graph, visualization.*

СОЗДАНИЕ СЕТЕЙ ЕСТЕСТВЕННЫХ ИЕРАРХИЙ ТЕРМИНОВ НА ОСНОВЕ АНАЛИЗА ТЕКСТОВЫХ КОРПУСОВ

Ландэ Д.В., Институт проблем регистрации информации НАН Украины

Аннотация. *Предлагается методика построения сетей иерархий терминов на основе анализа текстовых корпусов по выбранной тематике. Методика базируется на применении методологии компактифицированных графов горизонтальной видимости. Построена и исследована сеть языка, сформированная на основе подборки аннотаций электронных препринтов arXiv по тематике информационного поиска.*

Ключевые слова: *сеть языка, иерархия терминов, электронные препринты, граф видимости, визуализация.*

В настоящее время актуальными являются задачи построения онтологий по определенным областям знаний. Очевидно, построение большой отраслевой онтологии – сложная проблема, которая требует больших ресурсных затрат. В любом случае, определенным этапом построения общих онтологий является построение соответствующих тезаурусов, терминологических онтологий.

Предлагается методика построения сети естественной иерархии терминов, которую можно рассматривать как "квазионтологию", основу для формирования соответствующей терминологической онтологии. Сеть естественной иерархии терминов базируется на информационно-значимых элементах текста, опорных словах и словосочетаниях, методология выявления которых приведена в [1, 2]. Использование таких элементов позволяет формировать поисковые образы, охватывать целые области знаний в качестве основ для дальнейшего построения общих онтологий. Опорные слова и словосочетания для построения естественных иерархий терминов выбираются с учетом такого их свойства, как дескриминантная сила. Вместе с тем, одного этого свойства оказывается недостаточным при построении тезаурусов и онтологий. Иногда слова с низкой дескриминантной силой, в частности, наиболее частотные слова выбранной предметной области (например, слова "Information", "Retrieval", "Search" в корпусе по информационному поиску) оказываются важнейшими для задачи, которая рассматривается.

Формирование сети естественных иерархий терминов (СЕИТ) базируется на контенте текстовых корпусов соответствующей направленности. "Естественность" иерархий терминов в этом случае понимается как отказ при формировании сети от специальных методов семантического анализа. Все связи в такой сети определяются естественным применением слов и словосочетаний, которые экстрагируются из текстовых корпусов статистически значимых объемов. Сеть естественных иерархий терминов, создаваемая полностью автоматически, может рассматриваться как основа для дальнейшего автоматизированного формирования терминологической онтологии.

Алгоритм формирования сети естественных иерархий терминов, которая рассматривается в этой работе, предусматривает реализацию последовательности шагов, охватывающей предварительную обработку исходного текстового корпуса, определение и сортировку терминов, выбор необходимого количества наиболее весомых (наибольших узлов компактифицированного графа горизонтальной видимости [3]), построение СЕИТ и ее отображение. Рассмотрим эти шаги подробней.

1. На первом этапе выбирается исходный текстовый корпус. Как пример такого корпуса ниже рассматривается массив аннотаций электронных препринтов arXiv (www.arxiv.org) за 2007-2010 годы по тематике информационного поиска (рубрика cs.IR) объемом 550 записей.

Предварительная обработка такого текстового корпуса предусматривает выделение текстовых частей записей, исключение нетекстовых символов, стемминг.

2. На втором этапе каждому отдельному слову из текстового корпуса ставится в соответствие оценка его "дескриминантной силы", а именно TFIDF, которая в каноническом виде равна произведению частоты этого слова (Term Frequency) в фрагменте текста на двоичный логарифм от величины, обратной к количеству фрагментов текста, в которых это слово встретилось (Inverse Document Frequency)[4].

3-4. Выполняется то же, что и на предыдущем шаге, только для словосочетаний из двух слов (биграмм) и из трех слов (триграмм).

5. Для последовательностей терминов и их весовых значений по TFIDF строятся компактифицированные графы горизонтальной видимости (CHVG) [1, 2] и выполняется повторное определение весовых значений слов по этому алгоритму. Эта процедура позволяет учитывать в дальнейшем кроме терминов с большой дескриминантной силой также высокочастотные термины, которые имеют большое значение для общей тематики текстового корпуса. После этого все термины сортируются по убыванию рассчитанных весовых значений соответствующих узлов CHVG.

Дальнейшему анализу не подлежат термины из так называемого стоп-словаря. Это, как правило, фиксированный набор служебных слов, не играющих существенной роли в содержании текстов.

6. Экспертным методом определяется необходимый объем СЕИТ (число N), после чего избирается соответствующее количество единичных слов, биграмм и триграмм (всего $N+N+N$ элементов) с наибольшими весовыми значениями по CHVG.

7. Из отобранных на предыдущем шаге элементов строятся сети естественных иерархий терминов, в которых как узлы рассматриваются сами термины, а связи соответствуют вхождением одних терминов в другие. На рис. 1 проиллюстрирован принцип построения связей СЕИТ. Отдельные геометрические фигуры на этой иллюстрации соответствуют единичным словам.

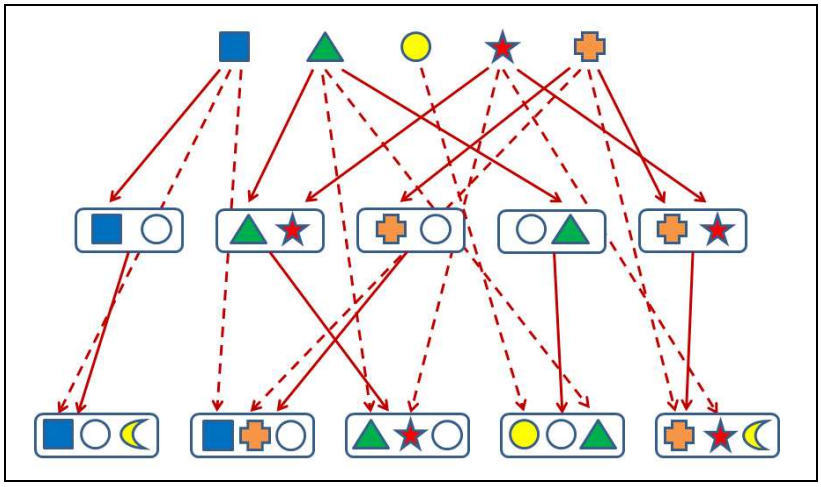

Рис. 1 - Формирование связей в трехуровневой сети естественной иерархии терминов

Первой строке соответствует выбранное множество единичных слов, второму – множество биграмм, а третьему – 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.

8. На последнем этапе формирования СЕИТ осуществляется ее отображение программными средствами анализа и визуализации сложных сетей. Для загрузки сетей естественных иерархий терминов в базы данным формируется матрица инцидентности общепринятого формата csv.

Для построенных сетей естественных иерархий терминов различных размеров по выбранному текстовому корпусу 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 200+200+200) составляет от 2,1 до 2,3.

На рис. 2 представлена небольшая сеть естественной иерархии терминов размером 20+20+20, которая визуализирована в виде спирали по предложенному автором методу.

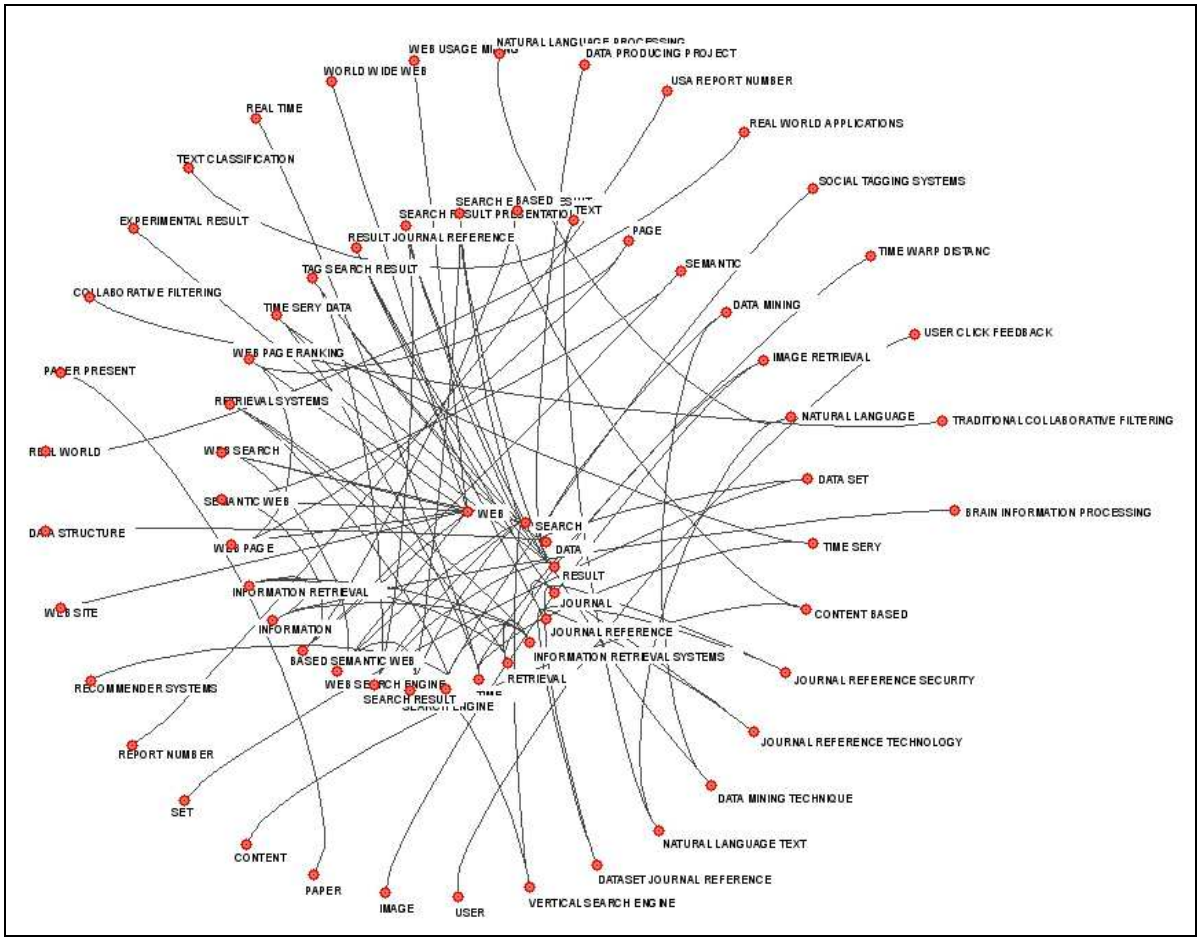

Рис. 2 - Вид СЕИТ размером 20+20+20

На рис. 3 представлен общий вид сети естественной иерархии терминов размером 200+200+200, которая визуализирована средствами системы Gephi (https://gephi.org/).

На рис. 4 приведены отдельные фрагменты сети естественной иерархии терминов, которые соответствуют выбранным базовым терминам.

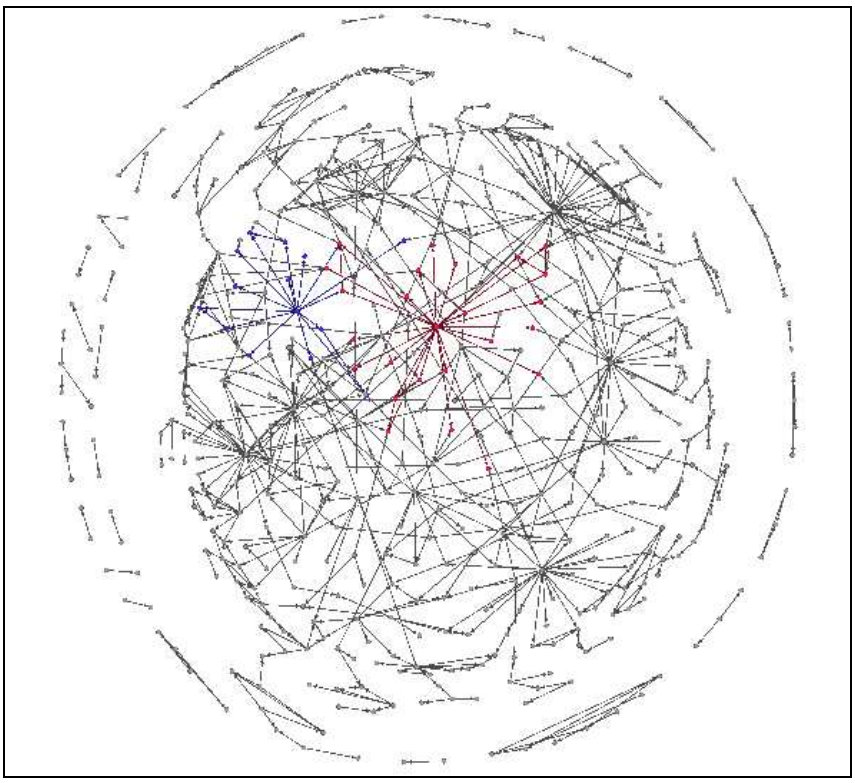

Рис. 3 - Визуализация СЕИТ размером 200+200+200 средствами Gephi

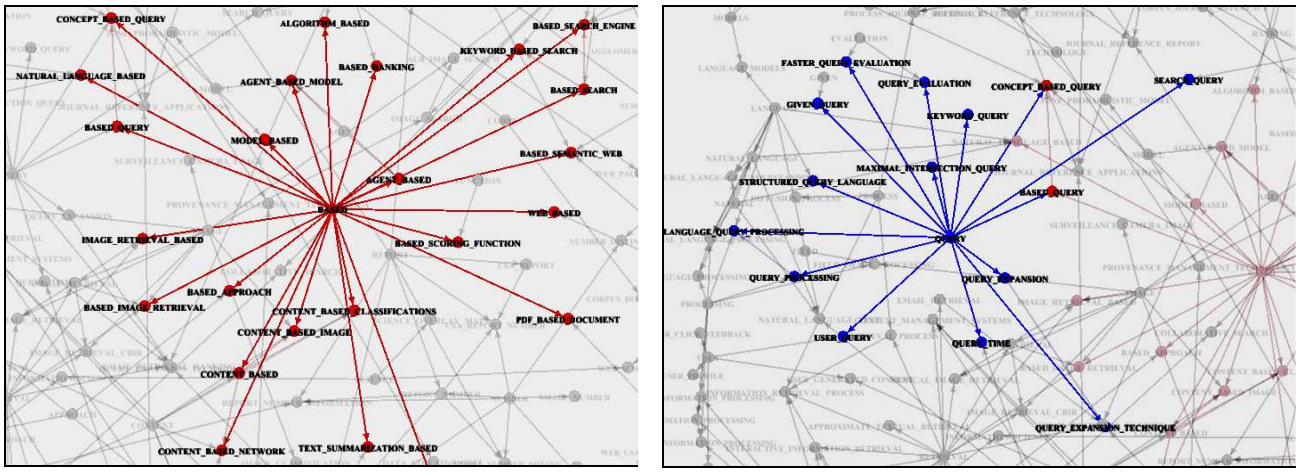

Рис. 4 – Фрагменты СЕИТ

Таким образом, в результатами проведенных исследований:

- Предложен алгоритм построения сетей естественных иерархий терминов на основе анализа текстовых корпусов.
- На основании этого алгоритма по текстовому корпусу построена сеть естественной иерархии терминов.
- Исследованы свойства сети естественных иерархий терминов, которая оказалась скейл-фри по исходящим связям.
- Выбраны средства визуализации сети естественных иерархий терминов.
- Сеть языка, построенную с помощью предложенной методики, можно использовать в качестве базы для построение общей онтологии (в рассмотренном примере – по тематике информационного поиска), использовать на практике в качестве готового к применению средства навигации в базах данных

соответствующей тематики, а также для организации контекстных подсказок пользователям информационно-поисковых систем.

Литература

1. *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.
2. *Lande D.V., Snarskii A.A.* Compactified Horizontal Visibility Graph for the Language Network // Preprint Arxiv (1302.4619)
3. *Luque B., Lacasa L., Ballesteros F., Luque J.* Horizontal visibility graphs: Exact results for random time series // Physical Review E, 2009. – P. 046103-1 – 046103-11.
4. *Salton G., McGill M.J.* Introduction to Modern Information Retrieval. – New York : McGraw-Hill, 1983. – 448 p.