Vitalii Velychko, Olexii Voloshyn, Krassimir Markov
(editors)

XX-th International Conference

"Knowledge-Dialogue-Solution"

Proceedings

ITHEA

2014

ITHEA International Scientific Society

XX-th International Conference

Knowledge-Dialogue-Solution

September 8-10, 2014, Kyiv (Ukraine)

P R O C E E D I N G S

ITHEA®

Kyiv - Sofia, 2014

Olexii Voloshyn, Vitalii Velychko, Krassimir Markov (eds.)

Proceedings of the XX-th International Conference "Knowledge-Dialogue-Solution"

ITHEA®

2014, Kyiv, Ukraine, Sofia, Bulgaria,

ISSN 1313-0087 (printed)

ISSN 1313-1206 (online)

ITHEA IBS ISC No.: 31

First edition

Printed in Ukraine

The XX-th International Conference "Knowledge-Dialogue-Solution" (KDS 2014) continues the series of annual international KDS events organized by Association of Developers and Users of Intelligent Systems (ADUIS).

The conference is traditionally devoted to discussion of current research and applications regarding three basic directions of intelligent systems development: knowledge processing, natural language interface, and decision making.

Edited by:

Association of Developers and Users of Intelligent Systems, Ukraine

Institute of Information Theories and Applications FOI ITHEA, Bulgaria

Publisher: ITHEA®

Sofia-1090, P.O.Box 775, Bulgaria

e-mail: office@ithea.org

www.ithea.org

ISSN 1313-0087 (printed)

ISSN 1313-1206 (online)

TABLE OF CONTENTS

Preface *3*
Table of Contents *4*
Index of Authors *7*

Аксіоматика Армстронга: повнота, критерій повноти та незалежність
Дмитро Буй, Анна Пузікова *9*

Гештальты в процессах образного мышления
Юрий Валькман *13*

Концепція аналізу та прийняття рішень при моделюванні сталого розвитку національної економіки
Олексій Волошин, Володимир Кудін, Володимир Кулик *17*

Анализ свойств нечётких обобщений методов распределения
Алексей Волошин, Василий Лавер *21*

Дистрибуция и развёртывание программной системы “ИКОН” с применением современных технологий виртуализации и облачных сервисов
Виталий Величко, Кирилл Малахов *25*

Спосіб автоматизованого виділення відношень між термінами з природномовних текстів технічної тематики
Віталій Величко, Віталій Приходнюк *27*

Медицинская грид-система для накопления и обработки электрокардиограмм
Виталий Вишневский *29*

Багатопараметричний вибір найбільш доцільних маршрутів ліній доступу з використанням моделі балансних мереж
Галина Гайворонська, Світлана Сахарова *31*

Феномен статистической устойчивости
Игорь Горбань *35*

Прогнозирование социально-экономического развития с помощью ассоциативных сетей
Валентин Григорьевский *38*

Прогнозирование на основе генетического алгоритма обучения
Леонид Гуляницкий, Анна Павленко *41*

Повышение эффективности алгоритма омк для прогнозирования третичной структуры протеинов методами распараллеливания
Леонид Гуляницкий, Виталина Рудык *43*

К вопросу об особенностях построения знаниеориентированных социальных сетей в интернете в сфере обучения
Андрей Данилов *48*

Матричная множественная регрессия
Владимир Донченко, Инна Назарага, Ольга Тарасова *50*

Разработка хранилища данных для информационно-ресурсного центра

Анна Жолнарская 54

Прогнозирование риска банкротства корпораций в условиях неопределенности с использованием нечетких нейронных сетей

Юрий Зайченко, Ови Нафас Агаи аг Гамиш 57

Анализ и автоматическое выделение мимических проявлений жестового языка в видеопотоке

Юрий Крак, Антон Тернов, Владислав Кузнецов 61

Метод линейной полосной классификации сложноразделимых дактилем украинского жестового языка

Юрий Крак, Григорий Кудин, Инга Соломянюк 63

Соответствие системы управления потребностям объекта управления

Виталий Косс 66

Автоматическое построение терминологических онтологий

Дмитрий Ландэ, Андрей Снарский 68

Деякі задачі оцінки платоспроможності підприємств у сфері кредитування

Микола Маляр, Володимир Поліщук 72

От физических отражений материального мира к основаниям информатики и математики

Анатолий Мержвинский 74

Алгебри квазіарних відношень та їх властивості

Нікітченко М.С., Шкільняк С.С., Матвіюк Д.А. 78

Моделирование рыночных механизмов в агентно-ориентированной модели функционирования экономики

Диана Омельянчик 79

Анализ существующих классификаций компетенций

Юлия Панасовская, Максим Вороной 81

Разработка методов семантического поиска web-сервисов на основе онтологического аннотирования

Юлия Рогушина 83

Кластерный анализ: проблема адекватности

Ирина Рясная 85

Автоматизированная система агрегации, поиска, визуализации данных

Михаил Соколов, Сергей Николаев 87

О следующем этапе компьютерной семантики

Владимир Сторож 90

О применении системологии в трансдисциплинарных исследованиях

Екатерина Соловьева 92

Классификация биоэлектрических сигналов на основе нейроподобных моделей

Наталья Н. Филатова, Дмитрий М. Ханеев 94

АВТОМАТИЧЕСКОЕ ПОСТРОЕНИЕ ТЕРМИНОЛОГИЧЕСКИХ ОНТОЛОГИЙ

Дмитрий Ландэ, Андрей Снарский

Аннотация: *Представлен подход к автоматическому созданию терминологических онтологий на основе анализа массивов текстов по выбранной проблематике. Подход базируется на применении компактифицированных графов горизонтальной видимости для терминов, а также автоматическом установлении связей между ними*

Ключевые слова: *граф горизонтальной видимости, сеть иерархии терминов, терминологическая онтология, текстовый корпус*

Введение

Для решения актуальных задач построения онтологий (детальной формализации выбранных областей знаний) требуется проведение комплексных исследований, определенным этапом которых является построение словарных номенклатур, тезаурусов. Эффективный автоматический отбор отдельных терминов для таких конструкций – не решенная окончательно задача, а проблема автоматического построения сетей из таких терминов до сих пор остается открытой. Как терминологическую основу для формирования соответствующей терминологической онтологии предлагается использовать сеть естественной иерархии терминов (СЕИТ), которая базируется на информационно-значимых элементах текста, опорных словах и словосочетаниях [Ландэ, 2014].

Постановка задачи

Опорные слова и словосочетания в теории информационного поиска выбираются с учетом такого их свойства, как дискриминантная сила. Вместе с тем, одного этого свойства часто оказывается недостаточно для отражения содержания предметной области. Иногда слова с низкой дискриминантной силой, в частности, наиболее частотные слова из выбранной предметной области (например, слова «Web», «Search», «Text» в корпусе текстов по тематике информационного поиска) оказываются важнейшими для рассматриваемой задачи. В данной работе для автоматического выявления терминологической сетевой основы при построении онтологий предметной области предлагается использовать сети естественных иерархий терминов, базирующейся на контенте аннотаций научных статей выбранной направленности. Связи в такой сети определяются естественным взаимным положением слов и словосочетаний, которые экстрагируются из текстов. Такая сеть, создаваемая полностью автоматически, может рассматриваться как основа для дальнейшего автоматизированного формирования терминологической онтологии с участием экспертов.

Методика формирования СЕИТ

Методика формирования сети естественных иерархий терминов предусматривает реализацию последовательности шагов, охватывающей предварительную обработку исходного текста, определение и сортировку терминов, выбор из них необходимого количества наиболее весомых, непосредственное построение СЕИТ и ее отображение. Рассмотрим эти шаги более подробно. В начале формируется исходный текстовый корпус. Как пример такого корпуса рассматриваются аннотации электронных препринтов Arxiv (http://arxiv.org) по тематике информационного поиска (рубрика csir, свыше 500 документов за 5 лет). Предварительная обработка такого текстового корпуса предусматривала выделение фрагментов текстов (отдельных аннотаций, абзацев, предложений, слов), исключение нетекстовых символов, отсечение флективных окончаний (стемминг). На втором этапе каждому отдельному термину из

текста (слову, биграмме или триграмме) ставится в соответствие оценка их 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 термин встретился (Inverse Document Frequency) [Salton, 1983]. Следует отметить, что TFIDF – не единственный метод для вычисления весовых значений терминов. Могут использоваться также различные дисперсионные и весовые оценки, широко применяемые в практике информационного поиска. Затем для последовательностей терминов и их весовых значений по TFIDF строятся компактифицированные графы горизонтальной видимости (CHVG) [Luque, 2009], [Ландэ, 2014] и выполняется переопределение весовых значений слов уже по этому алгоритму, что позволяет учитывать в дальнейшем кроме терминов с большой дискриминантной силой также высокочастотные термины, имеющие большое значение для общей тематики. В качестве весовых оценок отдельных слов в дальнейшем используются степени соответствующих им узлов. После этого все термины текста сортируются по убыванию рассчитанных весовых значений CHVG. Дальнейшему анализу не подлежат термины из так называемого стоп-словаря, являющиеся важными для связности текста, но не несущие информационной нагрузки. Используемый авторами стоп-словарь был построен на основе различных стоп-словарей, представленных в доступном виде на различных веб-ресурсах. Экспертным методом определяется необходимый размер СЕИТ, после чего выбирается соответствующее количество униграмм (единичных слов), биграмм и триграмм с наибольшими весовыми значениями по CHVG. Из отобранных терминов строятся сети естественных иерархий терминов, в которых как узлы рассматриваются сами термины, а связи соответствуют непосредственным вхождениям одних терминов в другие. На последнем этапе формирования СЕИТ осуществляется ее визуализация графов. На рис. 1 приведена визуализация фрагментов рассматриваемой СЕИТ в виде радиальных диаграмм, получивших сегодня широкое распространение в области изучения иностранным языкам.

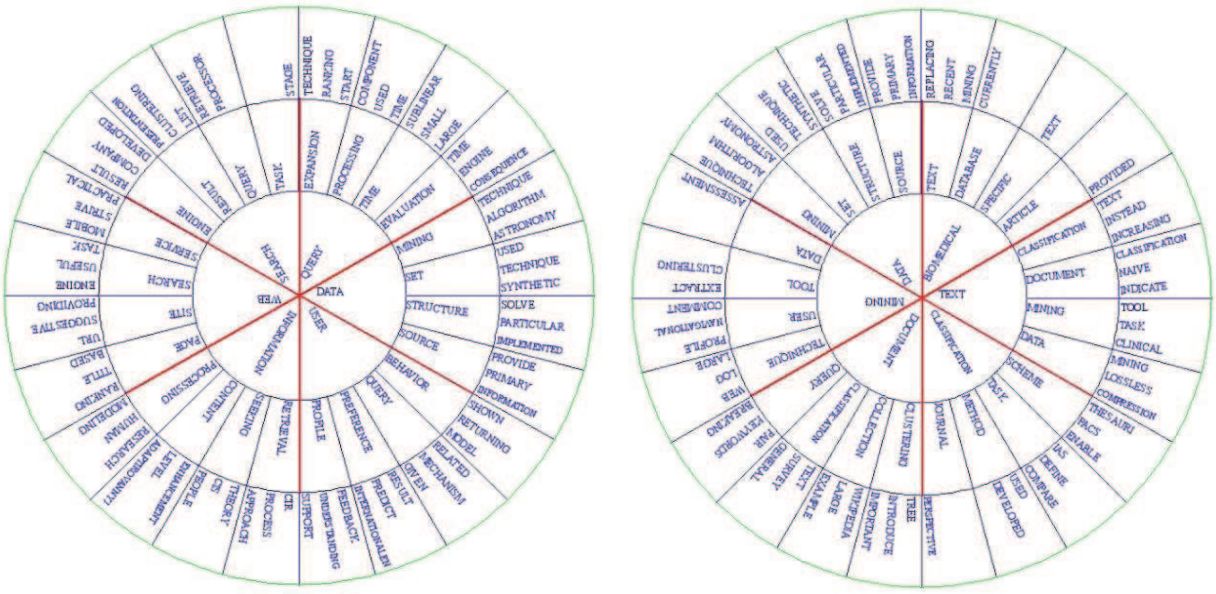

Рис. 1. Визуализация фрагментов СЕИТ в виде радиальных диаграмм

Ранжирование узлов СЕИТ

Ранжирование узлов в СЕИТ возможно по свойствам, обуславливаемым сетевой структурой, ссылками. 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 «хабов». Для решения этой задачи предлагается использовать известный алгоритм ранжирования веб-страниц, основанных на связях, HITS (hyperlink induced topic search), предложенный Дж. Клейнбергом [Kleinberg, 1998], который может применяться, например, наряду с алгоритмами PageRank (в этом случае – оценка единственная, интегрированная) или Salsa (этот метод идеально подходит для биграфов, а рассматриваемый нами граф – трехуровневый).

Алгоритм HITS обеспечивает выбор из информационного массива лучших «авторов» (узлов, на которые введут ссылки) и «посредников» (узлов, от которых идут ссылки включения). В рассматриваемом 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 для каждого узла сети v_j рекурсивно вычисляется его значимость как автора $a(v_j)$ и посредника $h(v_j)$ по формулам:

$$a(v_j) = \sum_i h(v_i);\ h(v_j) = \sum_i a(v_i). \qquad (1)$$

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.

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 ее 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 (рис. 2 а); 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 (рис. 2 б).

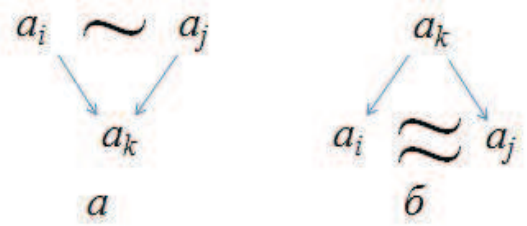

Рис. 2. Ассоциативные связи: а) первого рода «~»; б) второго рода «≈»

На рис. 3 приведен фрагмент сети СЕИТ, дополненной ассоциативными связями.

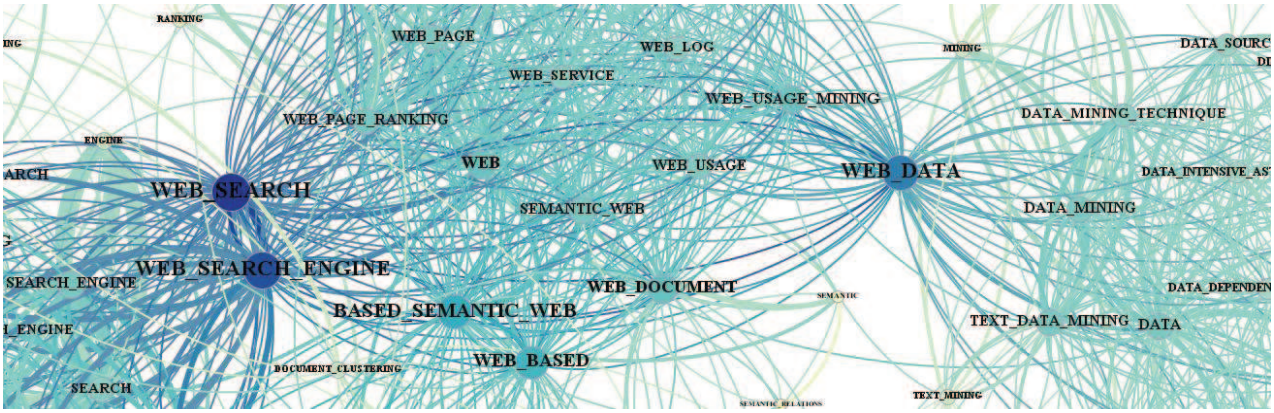

Рис. 3. Фрагмент СЕИТ с ассоциативными связями (визуализация с помощью программы Gephi)

Выводы

Таким образом, предложено:

- алгоритм построения сетей естественных иерархий терминов на основе анализа текстов;
- метод визуализации фрагментов СЕИТ в виде радиальных диаграмм;
- алгоритм построения ассоциативных связей между терминами в СЕИТ;
- использование алгоритма HITS для выбора наиболее важных элементов в СЕИТ.

Сеть языка, построенную с помощью предложенной методики, можно использовать в качестве базы для построения онтологии предметной области (в рассмотренном примере – по проблематике живучести), использовать на практике в качестве готового к применению средства навигации в информационных массивах, а также для организации контекстных подсказок пользователям информационно-поисковых систем.

Литература

[Ландэ, 2014] Ландэ Д.В., Снарский А.А. Подход к созданию терминологических онтологий // Онтология проектирования, 2014. – № 2(12). – С. 83-91.

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

[Luque, 2009] 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.

[Kleinberg, 1998] Kleinberg J. Authoritative sources in a hyperlinked environment // In Processing of ACM-SIAM Symposium on Discrete Algorithms, 1998, 46(5):604-632.

Информация об авторах

Дмитрий Владимирович Ландэ *– д.т.н., зав. отделом специализированных средств моделирования ИПРИ НАН Украины, ул. Шпака, 2, 03113 Киев, Украина, тел.(044) 4542163; e-mail: dwlande@gmail.com*

Андрей Александрович Снарский *– д.ф.-м.н., профессор кафедры общей и теоретической физики Национального технического университета Украины «КПИ», просп. Победы, 37 ,03056 Київ-56, Украина; e-mail: asnarskii@gmail.com*