УДК 004.622

Д.О. Донцов

АЛГОРИТМ ГЕНЕРАЦИИ ТЕЗАУРУСНЫХ РАСШИРЕНИЙ ДЛЯ КОРПОРАТИВНОГО ИНФОРМАЦИОННОГО ПОИСКА

Донцов Д.О. **Алгоритм генерации тезаурусных расширений для корпоративного информационного поиска.**

Аннотация. Целью работы является создание алгоритма генерации тезауруса синонимов для названий продуктов. Такие тезаурусы используются в современных поисковых машинах для расширения пользовательского запроса и улучшения качества поиска. При этом подходе из поискового индекса выбираются документы, включающие в себя не только слова, содержащиеся в запросе, но и близкие по смыслу термины. В ходе работы был реализован полуавтоматический метод обучения распознавателя именованных сущностей. Для валидации извлеченных сущностей был предложен метод полуавтоматической валидации.

Ключевые слова: информационный поиск, расширение пользовательского запроса, тезаурусные расширения, извлечение синонимов, распознавание именованных сущностей, строковая кластеризация.

Dontsov D.O. **Algorithm of thesaurus extension generation for enterpise search.**

Abstract. The main goal of this paper is to create algorithm of synonyms thesaurus generation. Modern search engines use such thesauri for query expansion. Such approach allows to return not only documents containing words from query, but also ones containing their synonyms or semantically similar terms. Semi-automatic method of named entity recognizer training was developed as a part of this work. Semi-automatic method of extracted entities validation is also given.

Keywords: information retrieval, query extension, thesaurus extension, synonyms extraction, named entity recognition, string clustering.

1. Введение. Вебсайт — это основной канал коммуникации IT-компании. Он наиболее важен для производителей и продавцов компьютерного аппаратного обеспечения. Посетители таких вебсайтов часто ищут информацию о продуктах компании, о возможности их заказа, о наличии необходимых драйверов, документации, пользовательских руководств и т.д. Таким образом, высокое качество поиска – одно из ключевых требований для поисковой машины корпоративного вебсайта.

Разные названия одного и того же продукта представляют серьезное препятствие, снижающее качество поиска на коммерче-

ском вебсайте. Устройство может иметь артикул, специальный кодовый номер, официальное, неофициальное и маркетинговое название. На практике часто происходит так, что пользователь знает только одно из этих названий и пытается найти релевантную информацию о продукте по его неофициальному названию. Для того, чтобы сделать это возможным, в информационном поиске используются различные подходы к расширению запроса. Как правило в них используются специальные словари синонимов, использующихся для расширения или переформулировки пользовательского запроса [9]. Словари синонимов считаются ценным источником семантической информации для поисковых машин [7]. Эти словари получили название "тезаурусные расширения".

Ручное составление словарей синонимов - долгий и дорогостоящий процесс. К тому же, качество такого словаря сильно зависит от компетентности составителя. Другой ключевой недостаток таких словарей состоит в том, что их ручное обновление также отнимает много времени. Для решения этой проблемы была предложена *автоматическая генерация словарей синонимов*. В данной статье описывается подход к решению этой задачи.

Основное предположение состоит в том, что можно извлечь пары синонимичных названий одного и того же продукта из html-страниц корпоративного вебсайта. Также предполагалось использование паттернов синонимов, извлеченных из англоязычной Википедии в работе [18]. В данной статье описывается применение этих паттернов для генерации тезаурусного расширения для сайта hp.com.

2. Обзор методов автоматической генерации тезаурусов. Существует множество подходов к извлечению семантических отношений и автоматической генерации тезаурусов. Все эти подходы так или иначе исследуют внешние источники знаний на предмет наличия отношений. Одна группа методов использует распределение терминов в коллекции документов. Например, авторы работы [5] используют максимальное расстояние между документами, содержащими термы с целью группирования термов в кластеры. Расстояние в данном случае определяется как расстояние между двумя векторами в модели bag-of-words (документ представляется в виде набора слов без учета их положения и связей в тексте).

Другая группа методов использует гипотезу о распределении и строит кластеры связанных термов основываясь на близости кон-

текстов, в которых они встречаются. Примером применения такого подхода может служить работа [23]. Особенностью этой работы является различие контекстов по грамматике, например, авторы отделяют контексты, в которых терм является субъектом, от тех, в которых тот же терм появляется в качестве объекта.

Третий подход использует знания о структуре ссылок в гипертекстовой разметке. Два наиболее ярких проекта, служащих источником информации о семантических отношениях - это WordNet и Википедия. Пример использования WordNet описан в работе [6]. Статья [14] описывает использование Википедии для поиска связанных сущностей. Основная разница между ними состоит в том, какие типы ссылок анализируются, например, ссылки по категориям [20] или ссылки внутри статьи [15].

Более современные подходы основываются на декомпозиции матриц, например, латентно-семантический анализ, описанный в статье [21], может применяться в связке вышеописанными методами.

Также следует отметить другое направление по генерации тезаурусов, использующее явную синонимию и другую семантическую информацию, содержащуюся в тексте. Первой в этой области была работа [8]. Этот подход успешно применялся во многих проектах, включая извлечение синонимов биомедицинской тематики [13], построение классификатора отношений “часть-целое” [19] и т.д. Этот метод используется в данной работе в сочетании с заранее определенным набором лексических паттернов для извлечения синонимических названий продуктов.

Целью данной работы является создание алгоритма генерации тезаурусных расширений на основе корпуса текстов и исходного словаря типовых именованных сущностей.

Основные эксперименты проводились на материалах сайта компании Hewlett-Packard (*hp.com*).

3. Подготовка данных. Общая схема проекта представлена на рисунке 1. Страницы веб-сайта были скачаны и из них был построен корпус. Часть корпуса использовалась как данные для обучения, в оставшейся части распознаватель выделял именованные сущности. Для разметки корпуса использовался исходный словарь, который был получен из pdf-каталогов продукции HP. Кластеры строк этого словаря использовались на этапе валидации извлеченных сущностей. На выходе системы - синонимы, прошедшие вали-

дацию. Результаты, полученные на этапе валидации также использовались для пополнения словаря.

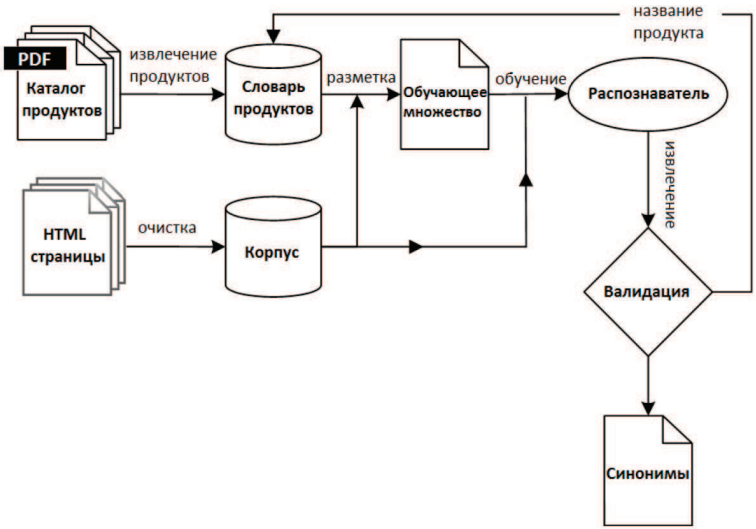

Рис. 1. Общая схема проекта.

Создание корпуса. Для создания корпуса использовались страницы веб-сайта hp.com. Всего было скачано и сохранено более 1500 страниц. Затем эти страницы необходимо было очистить от баннеров, рекламных, навигационных блоков и других подобных "зашумляющих"блоков. Существует несколько основных подходов к задаче очистки веб-страниц от информационных шумов. Большинство из них используют особенности объектной модели документа (DOM) или выявляют наиболее часто встречающиеся html-сегменты.

Например, *Site Style Tree (SST)* - дерево стилей сайта, использованное в работе [22], - это представление разных html-блоков всех страниц сайта в виде единого дерева со взвешенными узлами. Такой подход демонстрирует хорошие результаты в случаях, когда все страницы вебсайта построены на основе одного или нескольких шаблонов. К текущей задаче такой подход неприменим, т.к. количество шаблонов составляет несколько десятков. Другой под-

ход к задаче избавления от шумов получил название "плотностный"(densitometric) [11]. В основе метода лежит предположение о том, что основной сегмент страницы содержит наибольшее количество связного, не разбитого на отдельные блоки текста. Исходя из этой идеи для каждого блока измеряется "плотность"текста в нем, и по этому показателю производится фильтрация блоков. Также была использована эвристическая модель фильтрации "зашумляющих"блоков на основе черного и белого списков идентификаторов и дескрипторов html-блоков. Нужно отметить, что такой подход не является универсальным и применим только в тех случаях, когда вебсайт имеет хорошо организованную и систематизированную html-верстку, ориентированную на применение каскадных таблиц стилей (CSS). Для данной задачи наиболее высокие результаты показал плотностный подход, но одним из важных недостатков его применения является частое удаление знаков препинания в конце текста, который был классифицирован как основное содержание страницы. Отсутствие необходимых знаков препинания часто ухудшает результаты разбивки текста на предложения - одного из следующих этапов предварительной обработки.

Одним из источников для создания исходного словаря были результаты работы [9], содержащие около 122 000 продуктов HP. Тем не менее, из-за низкого уровня покрытия и точности этот словарь не мог быть единственным источником для разметки корпуса и был расширен с помощью официальных каталогов продукции HP. Общий объём словаря составил около 144 000 строк.

Разметка корпуса. В данной работе использовалась BIO-разметка, которая считается наиболее эффективной для обучения распознавателей [17]. BIO-разметка – это общепринятая схема для разметки именованных сущностей, которая ставит метку в соответствии с каждым токеном: началом сущности (B_X), продолжением сущности (I_X) и всеми остальными токенами (O). Пример такой разметки приведен в таблице 1.

Таблица 1. Пример BIO-разметки

The	HP	Media	Center	is	available	at	Best	Buy	.
O	B-HP	I-HP	I-HP	O	O	O	O	O	O

4. Распознавание и валидация извлеченных названий. Все названия продуктов HP отличаются от привычных именованных сущностей, которые можно встретить в повседневных текстах. Например, *"HP Pavilion dv6-1220sd Entertainment Notebook PC"*. Такая особенность в именовании требует обучения специального распознавателя. Для этой цели был выбран алгоритм машинного обучения CRF (Conditional Random Fields или условные случайные поля) [12]. CRF признан одним из лучших алгоритмов для задач разметки последовательностей и в частности для распознавания именованных сущностей. В данной работе использовалась программная реализация алгоритма, входящая в состав пакета для обработки естественного языка LingPipe [1].

На момент написания статьи большинство лучших распознавателей и классификаторов именованных сущностей были созданы на базе алгоритма CRF или на его расширении дополнительными наборами признаков. В данной работе стандартный для задач распознавания именованных сущностей набор признаков был дополнен признаками форм слова, которые учитывают особенности именования продуктов HP.

Рис. 2. Схема этапа распознавания названий продуктов.

После обучения из всего корпуса извлекались сущности, к которым применялись синонимические паттерны (см. рис. 2).

На каждой итерации распознаватель генерирует два выходных файла: список извлеченных синонимов и список сущностей, которые не были размечены как синонимы. Все сущности из второго списка могут быть использованы для расширения словаря, что, в свою очередь, может увеличить покрытие разметки на следующей

итерации.

Список извлеченных синонимов подвергается валидации в полуавтоматическом режиме. Сначала проверяется, содержит ли словарь какие-либо из извлеченных сущностей. В случае если такие сущности обнаруживаются, они удаляются из списка валидации.

Как показали результаты экспериментов, после каждой итерации около 52% извлеченных названий продуктов уже содержатся в словаре. Затем все оставшиеся названия сравниваются с кластерами словаря с использованием модифицированного расстояния Жаккарда. Под кластерами в данном случае понимаются группы строк, объединенные по значению определенной метрики близости. Модификация расстояния Жаккарда заключается в предварительной модификации сравниваемых строк: из них с помощью регулярных выражений удаляются типовые номера моделей, что затем позволяет объединить названия в продуктовые линейки.

Если расстояние между названием и одним из кластеров равно нулю, то оно добавляется в кластер и удаляется из списка валидации. В подавляющем большинстве случаев такое название является новой моделью в существующей линейке продуктов. Оставшиеся в списке валидации названия (после каждой итерации) валидируются вручную.

Около 94% названий, не прошедших автоматическую валидацию, являются ошибками распознавателя, то есть подстроки с неправильно определенной левой или правой границей. Тем не менее, среди них встречаются строки, принципиально отличающиеся от тех, которые уже содержатся в словаре (например, строка “HP 1 Year Post Warranty Next business day c-Class SAN Switch Hardware Support” была извлечена из корпуса, хотя в словаре похожие строки не содержатся).

Добавление таких строк в словарь сильно улучшают качество распознавания на следующих итерациях. Обычно около 30% названий из списка валидации некорректны. Количество таких “дельта” строк не увеличивается с итерациями. Их вклад в расширение словаря приведен в таблице 2.

Также был разработан эвристический алгоритм извлечения аббревиатур продуктов HP. Он заключается в нахождении совместно встречающихся названий продукта и набора заглавных букв, заключенных в скобки. Он также имеет простую процедуру ав-

Таблица 2. Расширение словаря по мере итераций алгоритма

Итерация	Добавлено строк	Общий объем
исходный размер	107,313	107,313
извлечено из pdf-каталогов	11,241	118,554
дельта 1	1629	120,183
дельта 2	1582	121,765
дельта 3	1598	123,363
...	...	...
дельта 15	1607	142,715

томатический валидации, основанную на сравнении извлеченных заглавных и первых букв в названии продукта. Таким способом было извлечено около 150 аббревиатур для продуктов и сервисов HP.

Кластеризация строк словаря. Как уже упоминалось выше, в работе была предпринята попытка сократить вычислительную сложность на этапе валидации строк. Исходная задача состоит в том, чтобы сравнить извлеченную строку с каждой строкой словаря. На каждой итерации извлекается около 5000 строк, объем словаря – около 150000 строк, таким образом, для валидации потребуется 7.5×10^8 сравнений. Т.к. словарь содержит множество похожих строк (например, ноутбуки одной серии, отличающиеся номером модели), было решено объединить их в кластеры и проводить сравнение строк-кандидатов с данными кластерами.

Для выбора оптимального алгоритма кластеризации было проведено отдельное исследование, детальное описание которого выходит за рамки данной статьи. Наилучшие результаты в кластеризации именованных сущностей показал гибридный алгоритм, использующий k-means для первичной, “грубой” кластеризации и последующей перегруппировкой полученных кластеров с помощью метода kNN (k ближайших соседей).

5. Оценка.

Распознавание именованных сущностей. Для оценки качества распознавания продуктов использовались стандартные метрики информационного поиска: полнота (R), точность (P) и F-мера

(F). Точность показывает, какой процент из извлеченных сущностей распознан правильно. Полнота – процент распознанных сущностей из всего множества. F-мера – среднее гармоническое взвешенное полноты и точности. Оценка получена в ходе перекрестной проверки (итерационное разделение всего множества на обучающее и тестовое), результаты представлены в таблице 3.

Таблица 3. Результаты перекрестной проверки

Точность	Полнота	F-мера
0,901	0,880	0,890

Т.к. эта оценка основывается на тексте размеченном сущностями из словаря, результаты зависят от качества разметки. Некоторые неразмеченные названия были извлечены правильно, чаще всего это разные модели продукта из одной серии. Например, словарь может содержать не строку *"hp pavilion u5053cn desktop pc"*, но в нем есть похожее название из этой же серии: *"hp pavilion a6652gr desktop pc"*. Даже если в этом случае строка извлечена верно, она считается ложно положительной (false positive), т.к. отсутствует в разметке. Такие случаи составляют около 75% от всех ложно положительных строк. Таким образом, с учетом неполноты разметки оценка точности может быть экстраполирована до 0,973, а F-мера, соответственно, до 0,924.

Извлечение синонимов. В данной работе предполагалось извлечение разных лексических представлений одной сущности с помощью паттернов синонимов, предложенных в работе [8]. Набор таких паттернов был извлечен из Википедии и отфильтрован в работе [18] и содержит около 250 паттернов с определенными коэффициентами доверия. Эксперименты проводились с паттернами из списка топ-25 и топ-60. Оставшиеся паттерны были заведомо неподходящими для данной задачи. Самые высокие коэффициенты доверия у паттернов "also known as"и "also referred to as". Основная идея состоит в том, чтобы разметить в тексте такие паттерны, затем распознать в нем все названия продуктов и отобрать предложения, в которых встретились пары продуктов, соединенные паттерном. Предполагается, что такие сущности являются синонимичными. Список также предполагалось дополнить

паттернами, которые являются специфичными для вычислительной техники и для оборудования в целом. Для этого был применен оригинальный подход М. Хирст [8] с использованием поисковых машин. Идея состоит в том, что в поисковый запрос вводятся два названия, являющиеся синонимами, а затем в найденных документах производится поиск явных лексических конструкций, соединяющих два данных синонима. Единственным найденным таким образом паттерном является "(or". Эксперименты показали, что только 5 паттернов могут дать точный результат.

Во всем корпусе было найдено 13,300 названий продуктов и 560 паттернов, но они очень редко встречаются в одном предложении. Только около 40 таких пар было найдено. Большинство извлеченных пар – это ссылки на прошлые названия продуктов, например, “HP was the first company to ship PCI-X-enabled servers with the launch of the *HP ProLiant DL760* server (formerly called the *ProLiant 8500*) in December 2000.”

Также набор паттернов позволил извлечь аббревиатуры названий продуктов и технологий. Всего было извлечено 82 акронима.

6. Вывод. В ходе данной работы была разработана архитектура для итерационного извлечения синонимов и акронимов для продуктов, технологий и сервисов ИТ-компании. Редкая совместная встречаемость паттерна синонимов и названий продуктов обуславливает применимость алгоритма только к большим входным корпусам текста.

В качестве дополнительного результата был получен эффективный алгоритм пополнения исходного словаря сущностей.

Одним из основных результатов работы является bootstrap-архитектура для расширения исходного словаря продуктов коммерческой компании с использованием собранных с ее сайта данных.

Список литературы

[1] Alias-i, “LingPipe 4.0.1” (Online; accessed 10 October 2011), http://alias-i.com/lingpipe.

[2] Beauliev, M., “Experiments of interfaces to support query expansion”, *Journal of Documentation*, **53** (1997), 8–19.

[3] Brajnik, G. and Mizzaro, S. and Tasso, C., “Evaluating user interfaces to information retrieval systems: A case study on user support”,

Proceedings of the 19th annual conference on Research and Development in Information Retrieval, ACM/SIGIR, Zurich, Switzerland, 1996, 128–136.

[4] Charikar, Moses S., "Similarity estimation techniques from rounding algorithms", *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, ACM, New York, NY, USA, 2002, 380–388, ISBN: 1-58113-495-9.

[5] Crouch, Carolyn J. and Yang, Bokyung, "Experiments in automatic statistical thesaurus construction", *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, 1992, 77–88, ISBN: 0-89791-523-2.

[6] Fox, Edward A. and Nutter, J. Terry and Ahlswede, Thomas and Evens, Martha and Markowitz, Judith, "Building a large thesaurus for information retrieval", *Proceedings of the second conference on Applied natural language processing*, Association for Computational Linguistics, Stroudsburg, PA, USA, 1988, 101–108.

[7] Gonzalo, J. and Verdejo, F. and Chugur, I. and Cigarran J., "Indexing with WordNet synsets can improve Text Retrieval", *Proceedings of the COLING/ACL '98*, ACL, 1988, 38–44.

[8] Hearst, Marti, "Automatic Acquisition of Hyponyms from Large Text Corpora", *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France, 1992.

[9] Kiseleva, Julia and Simanovsky Andrey, "Exploring synonyms within large commercial site search engine queries", 2011.

[10] Kohlschütter, Christian and Fankhauser, Peter and Nejdl, Wolfgang, "Boilerplate detection using shallow text features", *Proceedings of the third ACM international conference on Web search and data mining*, ACM, New York, NY, USA, 2010, 441–450, ISBN: 978-1-60558-889-6.

[11] Kohlschütter, Christian and Nejdl, Wolfgang, "A densitometric approach to web page segmentation", *Proceeding of the 17th ACM conference on Information and knowledge management*, ACM, New York, NY, USA, 2008, 1173–1182, ISBN: 978-1-59593-991-3.

[12] Lafferty, John D. and McCallum, Andrew and Pereira, Fernando C. N., "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", *Proceedings of the Eighteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, 282–289, ISBN: 1-55860-778-1, http://dl.acm.org/citation.cfm?id=645530.655813.

[13] McCrae, John and Collier, Nigel, "Synonym set extraction from the biomedical literature by lexical pattern discovery", *BMC Bioinformatics*, **9** (2008), 159, http://www.biomedcentral.com/1471-2105/9/159.

[14] Milne, David and Medelyan, Olena and Witten, Ian H., "Mining Domain-Specific Thesauri from Wikipedia: A Case Study", *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, IEEE Computer Society, Washington, DC, USA, 2006, 442–448, ISBN: 0-7695-2747-7.

[15] Nakayama, Kotaro and Hara, Takahiro and Nishio, Shojiro, "Wikipedia mining for an association web thesaurus construction", *Proceedings of the 8th international conference on Web information systems engineering*, Springer-Verlag, Berlin, Heidelberg, 2007, 322–334, ISBN: 3-540-76992-7, http://dl.acm.org/citation.cfm?id=1781374.1781410.

[16] Panchenko, Alexander, "Could we automatically reproduce semantic relatoin of an information retrieval thesaurus?", 2011.

[17] Ratinov, Lev and Roth, Dan, "Design challenges and misconceptions in named entity recognition", *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, 147–155, ISBN: 978-1-932432-29-9, http://dl.acm.org/citation.cfm?id=1596374.1596399.

[18] Simanovsky, Andrey and Ulanov, Alexander, "Mining Text Patterns for Synonyms Extraction", *Proceedings of DEXA Workshops, E-LKR'11*, IEEE Computer Society, 2011, 473–477, ISBN: 987-0-7695-4486-1.

[19] Snow, Rion and Jurafsky, Daniel and Ng, Andrew Y., "Learning Syntactic Patterns for Automatic Hypernym Discovery", *NIPS*, 2004, http://www.stanford.edu/ jurafsky/paper887.pdf.

[20] Strube, Michael and Ponzetto, Simone Paolo, "WikiRelate! computing semantic relatedness using wikipedia", *proceedings of the 21st national conference on Artificial intelligence*, **2**, AAAI Press, 2006, 1419–1424, ISBN: 978-1-57735-281-5, http://dl.acm.org/citation.cfm?id=1597348.1597414.

[21] Wandmacher, Tonio, "How semantic is Latent Semantic Analysis?", *Proceedings of TALN/RECITAL*, 2005, 6–10.

[22] Yi, Lan and Liu, Bing and Li, Xiaoli, "Eliminating noisy information in Web pages for data mining", *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, New York, NY, USA, 2003, 296–305, ISBN: 1-58113-737-0.

[23] Takenobu Tokunaga and Iwayama Makoto and Tanaka Hozumi, "Automatic Thesaurus Construction Based on Grammatical Relations", 1995.

Донцов Дмитрий Олегович — аспирант СПИИРАН, лаборатория информационно-вычислительных систем. Область научных интересов: обработка естественного языка, информационный поиск, машинное обучение.

d.dontsov@gmail.com; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ. Научный руководитель — Воробьев В.И.

Dontsov O. Dmitry — Ph.D. student at SPIIRAS, Laboratory of Computer and Informational Systems. Research interests: natural language processing, information retrieval, machine learning. d.dontsov@gmail.com; SPIIRAS, 14-th line V.O., 39, St. Petersburg, 199178, Russia. Research advisor — V.I. Vorobyov.

Рекомендовано лабораторией информационно-вычислительных систем СПИИРАН, зав. лабораторией Воробьев В.И., д.т.н., проф.
Статья поступила в редакцию 15.02.2013.

РЕФЕРАТ

Донцов Д.О. **Алгоритм генерации тезаурусных расширений для корпоративного информационного поиска.** Классическая модель информационного поиска текстовой информации основан на индексации документов и составлении поискового индекса, в котором каждый документ имеет представление в виде набора встречающихся в нем терминов. Улучшения качества информационного поиска как правило производится с помощью манипуляций с индексом. Например, учет терминов, расположенных рядом друг с другом, приведение слов к нормальной морфологической форме и т.д. Таким образом, поисковая выдача содержит только те документы, которые явным образом содержат слова из запроса. При этом информационная потребность пользователя может быть выражена множеством схожих по смыслу терминов (например: дом, жилье, квартира, жилплощадь, недвижимость), которые оказываются "незадействованными"при выборке документов из индекса. Чтобы улучшить полноту поиска и его качество существует подход по расширению пользовательского запроса синонимами введенных терминов. Для такого подхода необходим специальный словарь синонимов – т.н. "тезаурусное расширение".

Создание таких словарей вручную практически невозможно из-за большого количества терминов и необходимости регулярного обновления. Возникает необходимость в алгоритмах автоматической генерации тезаурусных расширений.

В статье предложен алгоритм автоматической генерации словаря синонимов для предметной области с существующим корпусом текстов. Работоспособность алгоритма доказана на примере корпуса текстов с сайта *hp.com*. Также предложен алгоритм валидации извлеченных терминов при больших объемах словаря, когда вычислительная сложность валидации сильно возрастает.

Приведено описание обучения распознавателя именованных сущностей, необходимого для извлечения не типовых терминов.

Статья содержит обзор методов очистки веб-страниц от "информационного шума".

Описан процесс разметки корпуса для последующего обучения распознавателя.

Приведены количественные результаты итеративного пополнения тезауруса и сделан вывод относительно сходимости алгоритма.

SUMMARY

Dontsov D.O. **Thesaurus extension generation for enterprise search.**

Classic text information retrieval model is based on document indexing and generation of search index where each document is presented as a set of containing terms. All improvements of classical information retrieval are basically manipulations with this index. E.g. taking into account terms collocation, morphological normalization, etc. Therefore, search result provide only documents that explicitly contains terms from query. At the same time user information need can be expressed as a set of semantically similar terms (e.g. house, apartment, accomodation, lodging, etc.) that has been "unused" in selection documents from index. In order to improve recall and, therefore, search quality method of query expansion is applied. Such approach has to operate with special synonyms dictionary, so-called "thesaurus extension".

Manual creation of such dictionaries is practically impossible due to big amount of terms and necessity of regular updates. Therefore we need an algorithm of automatic thesaurus extension generation.

Algorithm of automatic generation of thesaurus extension for specific area with existing text corpus is proposed in this article

Functionality of this algorithm is proven using text corpus from *hp.com*. Also author proposed algorithm of extracted terms validation against big dictionary, when computational complexity is dramatically high.

Description of non-typical named entity recognizer is also provided.

Article contains review of information "uncluttering"of web-pages.

Article describes of corpus markup for following named entity recognizer training.

Article also provides quantitative results of iterative dictionary extension and conclusion about algorithm convergence.