Автоматическое построение тезаурусных расширений для корпоративного информационного поиска

Д.О. Донцов

Санкт-Петербургский институт информатики и автоматизации РАН
d.dontsov@gmail.com

Аннотация

В данной статье описывается подход к автоматическому построению словаря синонимов для продукции компании Hewlett-Packard. Такие словари используются для расширения пользовательского запроса в поисковых механизмах корпоративного веб-сайта. Статья включает описание алгоритма генерации словаря синонимов.

1. Введение

Информационный поиск — одна из наиболее активно развивающихся областей информатики и одна из обширнейших областей программирования для веба и корпоративного сектора. Несмотря на обилие и разнообразие текстового содержимого веб-страниц, классический информационный поиск текстовой информации осуществляется по ключевым словам, то есть вся «интеллектуализация» поиска сводится к разбиению текста веб-страниц на отдельные слова (токены) и различным манипуляциям с ними.

Перед создателями поисковой системы часто встает проблема интерпретации пользовательского запроса. Поскольку поиск осуществляется по ключевым словам, от конкретной формулировки запроса зависит попадание в поисковую выдачу целого набора релевантных документов. Например, при таком подходе в ответ на запрос «подержанная машина» поисковый механизм не вернет страницы, содержащие словосочетание «подержанный автомобиль», что сильно ухудшает качество поиска. Для того чтобы избежать такой ситуации в современных поисковых системах используется технология «расширения запросов» («Query expansion»). Суть этого подхода заключается в дополнении пользовательского запроса новыми словами, которые могли бы улучшить релевантность выданных результатов. Расширение запроса позволяет существенно улучшить качество поисковой выдачи, но также имеет свои недостатки. Одним из недостатков является необходимость использования специального словаря синонимов, который должен содержать активно применяемую в запросах лексику. Обычно подобные словари принято называть «тезаурусными расширениями» [1].

Опыт исследователей и разработчиков систем информационного поиска показывает, что составить вручную словарь синонимов за разумное время возможно только для очень узкой предметной области [2]. Поэтому в области информационного поиска возникает новая задача – автоматическое построение тезаурусных расширений.

Целью данной работы было создание системы автоматического построения тезаурусного расширения для продуктов компании Hewlett-Packard. Это расширение предполагается использовать в системе поиска по порталу hp.com.

2. Общее описание алгоритма

Основная идея подхода состоит в том, чтобы извлечь синонимы из заранее подготовленного корпуса. Корпус состоит из html-страниц сайта hp.com, очищенных от информационного шума и нерелевантных блоков информации (шапка, навигационные, рекламные блоки и т.д.). Текущая версия корпуса насчитывает более 15,000 html-страниц.

Для извлечения синонимов из текста производится разметка двух типов. На первом этапе в корпусе размечаются все именованные сущности, которые в дальнейшем будут рассматриваться как кандидаты на роль синонимов.

На втором проходе размечаются так называемые паттерны синонимов. Это лингвистические конструкции, наличие которых в тексте говорит о том, что по обе стороны от них с высокой долей расположены синонимы. В английском языке одним из самых распространенных является паттерн ~~“a~~lso known as”.

После того, как была произведена разметка (рис. 1), из текста выбираются пары сущностей, разделенные паттерном. Предполагается, что такие пары являются синонимами.

Труды XIV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2011), Санкт-Петербург, Россия, 2011.

The <entity>HP DL585 G1 server</entity>, <pattern>also known as</pattern> the <entity>HP ProLiant DL858 G1</entity>, is a dynamic IT solution famous in the IT world for its performance and dependability.

Рис. 1. Пример полученной разметки

3. Подготовка данных

В задачах, связанных с обработкой массивов неструктурированных данных, полученных из веба, очистка и предварительная подготовка этих данных к последующей обработке является одной из самых трудоемких операций.

В данном случае основную сложность представляет проблема сильной зашумленности. Она состоит в том, что необходимые для нашей задачи данные (текст новости, записи в блоге, на форуме и т.д.) зашумлены малозначащими блоками информации: рекламой, меню, баннерами, списками похожих новостей, навигационными блоками и т.д. Чаще всего подобные блоки нерелевантной информации являются шаблонами для большинства страниц одного веб-сайта и задачу

Одним из дальнейших шагов алгоритма является обучение распознавателя именованных сущностей. В качестве данных для обучения используется размеченная часть корпуса html-страниц. Таким образом, обучение распознавателя на зашумленных данных может значительно снизить качество его работы.

Следовательно, на стадии подготовки данных их нужно очистить от этих шумов.

В исследованиях, посвященных информационному поиску, очистку данных от шума принято считать отдельной задачей, требующей особого подхода. Объем информационного шума, который можно отнести к шаблонам сайтов, оценивается как 40-50% от общего объема html-данных в вебе [3].

Существуют разные алгоритмы выявления html-шаблонов. Большинство из них основаны на анализе объектной модели документа (DOM), которая представляет собой дерево элементов документа. Такой подход нашел отражение в работах [4] [5] [6] [7].

Также существуют подходы, использующие статистические меры html-разметки (плотность ссылок, плотность текста, плотность тэгов, TF/IDF, отношение количества токенов в текущем блоке к предыдущему и т.д.). Дополнением к этим подходам могут служить составленные вручную словари атрибутов. В этих словарях содержатся значения атрибутов тех блоков, которые не могут содержать полезной информации и наоборот – значения атрибутов «хороших» тэгов. Такие вспомогательные словари могут существенно улучшить качество очистки, если производится обработка страниц одного веб-сайта, где все страницы размечены по одним и тем же шаблонам при использовании одних и тех же таблиц стилей.

Подробный обзор существующих подходов к очистке от «зашумляющих» блоков приводится в статье [8].

В нашем случае производилась обработка страниц сайта hp.com; верстка этого сайта выполнена в одном стиле с применением одних и тех же CSS-классов и идентификаторов. Это позволило составить словари «зашумляющих» тэгов, которые использовались при последующей фильтрации элементов. После удаления этих тэгов для последующей фильтрации применялась библиотека *boilerpipe,* использующая статистические метрики для выделения основного содержания страницы; эти метрики подробно описаны в [8].

4. Распознавание именованных сущностей

Описанный в разделе 2 алгоритм предполагает использование автоматического распознавателя именованных сущностей. В данном случае речь идет о названиях продуктов компании Hewlett-Packard. Чаще всего это наименования оргтехники, имеющие сложные наименования. Например, «*HP Color LaserJet Q5952A Yellow Print Cartridge with ColorSphere Toner*». Специфичность и сложность таких именованных сущностей не позволяют использовать для их распознавания существующие натренированные распознаватели. Для того, чтобы извлекать такие названия из текста было принято решение натренировать собственный распознаватель.

Распознавание именованных сущностей (Named Entity Recognition или NER) – типичная задача разметки последовательностей (sequence labeling), для которой существует множество алгоритмов. Для нашей задачи был выбран алгоритм на основе модели CRF (Conditional Random Field) [9]. Системы NER на основе CRF демонстрируют высокое качество работы и превосходят системы, использующие скрытые марковские модели [10].

Для тренировки распознавателя нужно составить тренировочное множество, в котором были бы уже размечены продукты HP. В качестве такого набора данных было решено использовать часть корпуса html-страниц hp.com. Для автоматической разметки тренировочного набора данных нужно иметь словарь продуктов HP, чтобы выделить их в тексте. Этот словарь не обязательно должен быть исчерпывающим и иметь полное покрытие всех сущностей: в процессе тренировки и тестирования распознавателя этот словарь пополняется новыми строками.

Для получения словаря были использованы каталоги продукции HP в виде pdf-файлов. Эти файлы были преобразованы в html-разметку и из содержащихся в ней таблиц были извлечены все строки, соответствующие названиям продуктов HP. В данном случае pdf-файлы были хорошо

структурированы, что позволило автоматически произвести извлечение всех необходимых строк.

Далее, имея в наличии словарь, была произведена разметка тренировочного корпуса. Одной из основных проблем на этом этапе была малая степень покрытия, то есть в тексте встречалось много названий, которые не были размечены. Пересекающиеся названия также приводят к некорректной разметке. Например, возможна ситуация, когда в предложении вместо «HP Proliant G1» может выделиться только часть «HP Proliant», потому что словарь не содержит «HP Proliant G1», а только «HP Proliant» или «HP Proliant G2». При этом зачастую распознаватель правильно определял границы названия, несмотря на некорректную разметку. Это делало невозможной правильную оценку работы распознавателя.

4.1. Расширение словаря

Чтобы повысить степень покрытия словаря, было необходимо дополнить его новыми строками. Для этого был предложен следующий алгоритм. После работы распознавателя составлялся список названий, которые были распознаны некорректно. В информационном поиске для таких случаев употребляется термин «false positive» (FP). Далее каждый FP сравнивался со строкой из словаря. Для сравнения использовалась специально созданная метрика, использующая принцип, похожий на подсчет расстояния Левенштейна, но не на уровне символов, а на уровне токенов. Иными словами, если FP отличается от строки из словаря на один токен, и этот токен имеет категорию «модель», то FP добавляется в словарь (рис. 2). Определение категории происходит с помощью регулярных выражений. В категорию «модель» попадают, например, такие токены как: «BL450c», «DL365», «G1», «d9214» и т.д. Таким образом словарь пополняется названиями, которые отличаются только номером модели. Такое расширение словаря позволило существенно повысить качество работы распознавателя и провести корректную оценку распознавания.

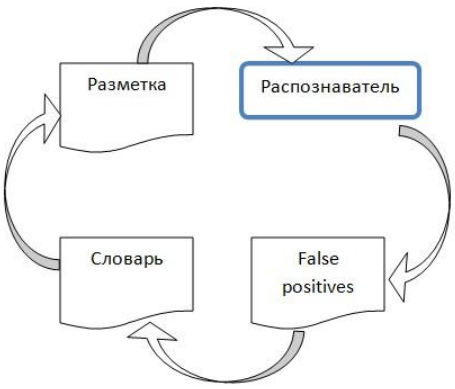

Рис. 2. Цикл пополнения словаря

Результаты работы, полученные в результате перекрестной проверки (cross-validation) на корпусе в 11,500 предложений следующие:

Точность: 0.924
Полнота: 0.919
F-мера: 0.921

4.2. Выделение синонимов

Достигнутая точность и полнота распознавания позволяют перейти к следующему шагу – разметке паттернов синонимов и выделению троек «сущность-паттерн-сущность». Список паттернов синонимов на данный момент уже сгенерирован. Эта работа была выполнена на базе англоязычной Википедии и описана в [11].

5. Заключение

На следующем этапе будут размечены паттерны синонимов и сгенерирован словарь синонимов для продуктов компании HP.

Литература

[1] Маннинг, К.Д. Введение в информационный поиск / Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце. М.: Вильямс, 2011. С. 168-169.

[2] Frei, Y. Concept Based Query Expansion / Frei, Y. Qiu and H.P.,// Proceedings of the 16th ACM International Conference on Research and Development in Information Retrieval. Pittsburgh, 1993.

[3] Gibson, D. The volume and evolution of web page templates / D. Gibson, K. Punera, A. Tomkins // Proceedings of the 14th international conference on World Wide Web. New York, NY, USA, 2005.

[4] Baluja, S. Browsing on small screens: recasting web-page segmentation into an efficient machine learning framework // Proceedings of the 15th international conference on World Wide Web. New York, NY, USA, 2006.

[5] Chakrabarti, D. A graph-theoretic approach to webpage segmentation / D. Chakrabarti, R. Kumar, K. Punera // Proceeding of the 17th international conference on World Wide Web, New York, USA, 2008.

[6] Chakrabarti, D. Page-level template detection via isotonic smoothing / D. Chakrabarti, R. Kumar, K. Punera // Proceedings of the 16th international conference on World Wide Web, 2007.

[7] Lan, Y. Eliminating noisy information in Web pages for data mining / Lan Yi, Bing Liu, and Xiaoli L // Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '03), New York, NY, USA, 2003.

[8] Kohlschütter, Ch. Boilerplate detection using shallow text features / Christian Kohlschütter, Peter Fankhauser, Wolfgang Nejdl // Proceedings of the third ACM international conference on Web search and data mining (WSDM '10), New York, NY, USA, 2010.
[9] Lafferty, J. Conditional random fields: Probabilistic models for segmenting and labeling sequence data / Lafferty, J., McCallum, A., Pereira, F. // Proceedings of the 18th International Conference on Machine Learning, 2001.
[10] McCallum, A. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons / Andrew McCallum, Li Wei // Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, Edmonton, Canada.
[11] Simanovsky, A. Mining Text Patterns for Synonyms Extraction / Andrey Simanovsky, Alexander Ulanov // Accepted for publication at the 1st International Workshop on Exploiting Large Knowledge Repositories, 2011.

Automatic generation of thesaurus extension for enterprise search

Dmitriy O. Dontsov

Article describes approach to automatic generation of synonyms dictionary for Hewlett-Packard products. These dictionaries are used for query extension in search engine of enterprise web-site. Article provides description of the algorithm of synonyms dictionary generation.