Построение поисковой системы для узкоспециализированной цифровой коллекции

© А.Н. Талбонен

Петрозаводский государственный университет (ПетрГУ)
perhetal@onego.ru

Аннотация

В данной статье описываются основные подходы к построению поисковой системы для узкоспециализированной текстовой коллекции, допускающей наличие текстовых ошибок, на примере коллекции фотографий со строительства Беломорско-Балтийского канала. Предложенные в статье методы можно использовать для других проектов, связанных с работой над коллекциями исторических документов.

1. Описание задачи

В качестве исходных данных выступает текстовая коллекция, полученная в результате распознавания коллекции изображений объемом около 6500 фотографий и обработанная текстовыми методами [3]. Вследствие низкого качества изображений текстовая коллекция также содержит ошибки и неточности. Ниже перечислены основные особенности задачи:

1. Наличие ошибок распознавания
2. Наличие орфографических ошибок в правильно распознанных словах
3. Наличие ключевых слов, которые отсутствуют в каких-либо словарях.
4. Наличие орфографических ошибок в неизвестных словах
5. Наличие различных сокращений, распознанных как мусор.

Цель данной работы – построить систему поиска по данной коллекции.

2. Решение задачи

2.1 Обнаружение и исправление ошибок распознавания

Наличие ошибок распознавания, особенно в ключевых словах, значительно снижает качество поискового индекса. Более того, отсутствие слова в словарях означает, что оно будет распознано как неизвестное слово или слово с ошибкой. Для обнаружения и исправления данных ошибок применяется автоматизированный метод.

Прежде всего, каждое слово текстовой коллекции необходимо проверить на отсутствие ошибок. Для этого достаточно, чтобы оно содержалось в одном из словарей, либо было распознано морфологическим анализатором. В противном случае, можно считать, что слово содержит ошибку.

Далее для каждого слова с ошибкой необходимо найти наиболее близкое слово. В качестве словаря для проверки может выступать база данных словоформ, а мерой близости может служить функция Левенштейна [2]. В случае, если расстояние до «ближайшего» слова превышает определенную границу (порог распознавания), слово будет считаться неизвестным.

Исправление ошибок можно осуществлять как в ручном режиме, так и в автоматическом. В последнем случае будет существовать ненулевая вероятность ошибки. Однако, учитывая основную цель, можно допустить замену слова с ошибкой на нормальную форму правильно распознанного слова. Если при этом разрешить автоматическое исправление только для незначительного расстояния между словами, то вероятность ошибки будет низкой.

2.2 Построение полнотекстового индекса

Индекс можно построить на основе текстовой коллекции с помощью любого морфологического анализатора, например, Mystem [1], позволяющего находить для каждого слова его нормальную форму, часть речи и другую морфологическую информацию. Однако в данном случае, часть слов и, в особенности, часть ключевых слов, будет оставаться нераспознанной, т.к. не эти слова не содержатся ни в одном словаре. Для решения этой проблемы был разработан человеко-машинный метод дополнения индекса.

Суть метода заключается в том, что человек добавляет новое слово в один из нескольких тематических словарей, определенных заранее и зависящих от предметной области, либо в общий словарь. При этом указывается морфологическая информация: часть речи, нормальная форма слова и возможные словоформы. В итоге, при

Труды 13[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

индексировании можно записывать в индекс только нормальную форму добавленного слова.

Метод также предполагает, что по мере добавления новых слов поступающая информация должна учитываться при выявлении и исправлении ошибок, в том числе ошибок в неизвестных словах. Например, в случае, когда вследствие объективных причин (например, опечатки) одному и тому же слову может соответствовать несколько вариантов написания, притом, что само слово отсутствует в каком-либо словаре.

Решение проблемы заключается в сочетании метода дополнения индекса и метода обнаружения и исправления ошибок, что позволяет в словарь нового слова индексировать все его возможные словоформы, обнаруженные в коллекции. При этом оба метода могут чередоваться и применяться неограниченное количество раз, тем самым постепенно сокращая процент ошибок. Поэтому для экономии времени и ресурсов разумно ограничивать количество итераций неким пороговым значением процента ошибок, ниже которого проводить дальнейшее уточнение индекса бессмысленно.

2.3 Построение онтологии

К сожалению, полностью автоматических методов построения онтологии по узкой предметной области не существует, однако есть возможность дополнять содержимое онтологии автоматически при наличии различных тематических словарей. Так, например, в рамках данной работы можно выделить несколько основных тематических словарей: географические названия, строительные объекты, строительная техника, люди и др. При этом нераспознанные слова можно отнести к одному из словарей с помощью вышеописанного метода дополнения индекса. Таким образом, можно получить по одной ветки в отологической структуре для каждого из тематических словарей, после чего можно уточнять полученную онтологию.

Другой способ дополнения онтологии заключается в нахождении в текстовой коллекции ключевых словосочетаний таких, которые в случае замены данного словосочетания уникальным ключевым словом будут повышать точность поиска. Например, словосочетание вида «<объект> <номер>» можно заменить ключевым словом «<объект>_<номер>», которое станет уникальным для данной коллекции, при этом его можно добавить в таксономический узел онтологии как самостоятельный объект. Данного результата можно достичь применением контекстного анализатора.

2.4 Контекстный анализ текстовой коллекции

Суть данного метода заключается в поиске по определенным наборам правил и выполнении соответствующих действий над найденными словами. При этом каждое правило представляет собой ориентированный набор элементов определенного типа. В данной работе были выделены следующие типы элементов правил:

1. Константа
2. Лексема (токен), например, слово, состоящее из букв алфавита.
3. Лексическая группа, т.е. слову соответствует определенная часть речи.
4. Тематическая группа, т.е. слово относится к одному из тематических словарей.
5. Онтологическая группа, т.е. слово принадлежит определенному таксономическому узлу.

Данные правила позволяют выполнять поиск ключевых словосочетаний не только в текстовой коллекции, но и в поисковом запросе, что в свою очередь позволяет уточнять как сам поисковой индекс, так и результат поиска.

3. Заключение

Принципиальной разницы в выборе движка поисковой системы по построенному индексу и онтологии нет. Требование к поисковому движку одно – возможность полнотекстового поиска по индексу и учет метаинформации, извлеченной из онтологии. Основная отличительная идея данной работы – уточнение индекса с помощью человеко-машинного метода дополнения словарей, поиск ключевых словосочетаний с помощью контекстного анализатора с последующим повышением веса найденных комбинаций, а также дополнение онтологии найденными ключевыми словосочетаниями.

Литература

[1] О программе mystem — Компания Яндекс. http://company.yandex.ru/technology/mystem/

[2] Расстояние Левенштейна - Википедия. http://ru.wikipedia.org/wiki/Расстояние_Левенштейна

[3] Талбонен А.Н., Рогов А.А. Анализ машинописных подписей к фотографиям в цифровом альбоме // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XII Всероссийской научной конференции RCDL'2010.- Казань: Казан. ун-т, 2010. С. 422-429.

Creating Search System for Highly Specialized Digital Collection

© A.N. Talbonen

This article contains description of basic approaches to building a search system for highly specialized digital collection, allowing text errors, and is based on image collection of White Sea - Baltic Canal’s construction. Techniques proposed here can be used in other projects related to historical document collection processing.