SEUS на РОМИП 2009: Оценка модели поиска системы Lucene и ее развитие с учетом семантики

© Рабчевский Е.А. Рожков М.С.

Кафедра компьютерных систем и телекоммуникаций
Пермского Государственного Университета
evgeny@rabchevsky.name demonrms@gmail.com

Аннотация

В статье описывается проект системы поиска с учетом семантики SEUS (http://seus.rabchevsky.name), и ее участие в семинаре РОМИП'2009.

1. Проект системы поиска с учетом семантики SEUS

SEUS, или search engine using semantics – проект, развиваемый на кафедре компьютерных систем и телекоммуникаций Пермского Государственного Университета. Целью проекта является разработка системы поиска по запросу для коллекций русскоязычных документов. Система ориентирована в первую очередь на поиск во всемирной паутине. Особенностью системы является ее модель поиска. Последняя представляет собой модифицированную векторную модель, в которой стандартное понятие терма заменяется RDF триплетом [1]. То есть документ представляется не набором лемм, а набором RDF триплетов, которые в совокупности отражают семантику документа.

Основной сложностью в реализации указанной методики является задача представления текста документа в виде соответствующего семантического RDF графа. Для решения этой задачи в системе используются следующие механизмы:

- извлечение заранее встроенных в документы семантических данных, например с использованием технологии микроформатов, RDF/A [2].

- лексико-синтаксические шаблоны [3], позволяющие извлекать семантику на основе характерных регулярных выражений и языковых конструкций

- использование логического вывода для расширения графа, полученного двумя предыдущими методами.

Проект SEUS – это первая попытка нашего коллектива применить механизм лексико-синтаксических шаблонов, который исследуется нами с 2006-го г., к задаче поиска по текстовым коллекциям. Поэтому за основу поисковой системы нами была взята открытая платформа для поиска по текстовым коллекциям Lucene [4].

2. Цели участия SEUS в РОМИП

На сегодняшний день нами разработан язык для формализации лексико-синтаксических шаблонов LSPL, интерпретатор (программа, применяющая шаблоны к тексту) языка LSPL, online версия семантического анализатора на основе лексико-синтаксических шаблонов, средство для разработки, хранения и валидации шаблонов, а также модифицирована векторная модель поиска под графовое представление текста. Более детально с данными результатами можно ознакомиться в материалах семинара молодых ученых, проводимого на конференции RCDL'2009 (библиографические данные на момент написания статьи не известны).

Однако, в целом, результаты решения задачи представления текста в виде семантического графа пока не достойны внимания читателя. Поэтому, участвуя в семинаре, мы ставили перед собой следующие цели:

- реализовать модель поиска библиотеки Lucene и сравнить качество ее работы в стандартном варианте с результатами других участников;

- получить данные, которые бы могли послужить «контрольной точкой» в решении задачи графового представления текста в контексте информационного поиска.

Второе является основной целью участия в семинаре.

3. Алгоритмы, используемые в прогонах семинара

Наш коллектив заявлялся на участие по дорожкам поиска по коллекции нормативно-правовых документов и Веб-коллекции, и выполнил все задания семинара.

Для решения заданий семинара нами была использована библиотека Lucene 2.4.0 и русскоязычный стеммер Snowball из той же поставки Lucene.

Во время подготовки к выполнению заданий семинара мы модифицировали стандартную поставку библиотеки Lucene под модель поиска с графовым представлением текста, указанную в первой главе. Для работы с RDF-графами мы использовали библиотеку Jena [5]. Опишем приведенную модель более детально:

- документы представлялись в виде RDF-графов, триплеты которых сохранялись в базе данных в виде Jena-моделей, имеющих уникальные идентификаторы и ссылки на документы, к которым они относятся. То есть документ представлялся набором идентификаторов триплетов, которые отражали его семантику.

- далее строился инвертированный индекс, вектора которого отражали наличие соответствующих триплетов в графе документа, принимая значения 1 или 0.

- булевы координаты векторов умножались на коэффициенты доверия соответствующих триплетов, которые отражали достоверность представления документа соответствующим графом, и хранились также в базе данных. Коэффициенты – это экспериментально получаемые данные, отражающие валидность лексико-синтаксических шаблонов и операций логического вывода.

- в итоге строился индекс, в котором вместо относительных частот TF*IDF была матрица достоверности отражения коллекции документов соответствующими графами.

Изменениям были подвергнуты методы класса Similarity библиотеки Lucene: метод обработки отдельного терма, и метод обработки коллекции термов.

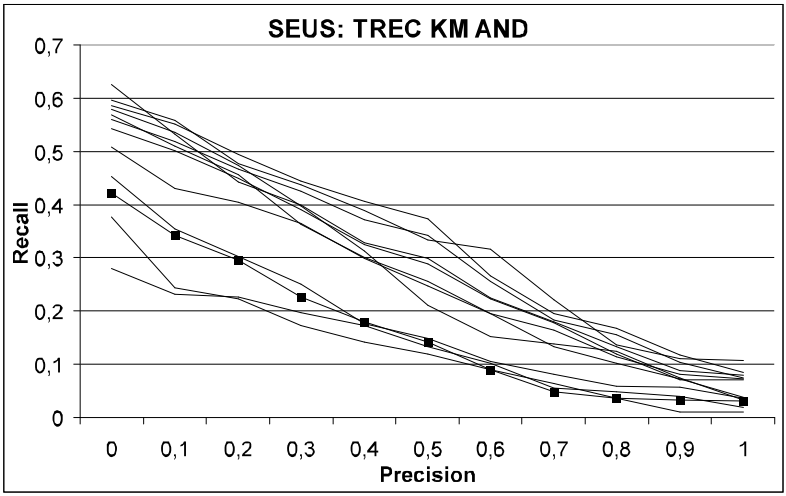
Указанная модель была реализована, однако при выполнении заданий семинара мы столкнулись с такой проблемой: ~~того, что~~ механизм представления текста в виде графа давал очень плохие результаты. Так, указанная модель поиска была реализована, но так и не использовалась, потому что исходные для нее данные не были пригодны.

Поэтому результаты, полученные с помощью указанной методики, на семинар представлены не были. В связи с этим указанная модель приведена здесь очень сжато. Более детально с указанной методикой можно ознакомиться на сайте проекта [6] или в ~~—~~последствии ~~е~~ в материалах семинара молодых ученых, проводимого на конференции RCDL’2009 (библиографические данные на момент написания статьи не известны).

Таким образом, в прогонах системы участвовала только стандартная версия библиотеки Lucene. О которой следует сказать пару слов.

Механизм ранжирования библиотеки Lucene основан на инвертированном индексе, состоящем из векторов документов, и метрике TF*IDF, с помощью которой определяется угол между вектором запроса и векторами документов коллекции. Ввиду того, что векторная модель широко известна в кругах специалистов, здесь она раскрываться не будет. Подробно ознакомиться с алгоритмами библиотеки Lucene можно в ее документации [7].

Здесь лишь приведем результаты оценки работы нашей системы на примере дорожки по поиску по коллекции KM.RU в виде 11-ти точечного графика TREC.



На данном графике линией, отмеченной черными квадратиками, обозначены результаты оценки нашей системы.

4. Заключение

Обе цели, поставленные перед участием в семинаре, были достигнуты:

- выполнен поиск по коллекциям семинара с помощью библиотеки Lucene;
- и получены данные, которые в последствие можно будет использовать для оптимизации решения задачи поиска по текстовым коллекциям данных с учетом их семантического описания.

В результате участия в семинаре мы сделали следующий вывод: исходя из результатов, представленных на 11-ти точечном графике, видно, что библиотека Lucene без модернизации под конкретную коллекцию дает не самые худшие результаты. Хотя мы предполагали, что получим самые низкие оценки ввиду того, что на семинар была представлена только стандартная поставка библиотеки Lucene, над которой мы не производили ни–каких операций.

В дальнейшем планируется улучшить результат решения задачи графового представления текста. Имея коллекции данных, запросы и таблицы релевантности, полученные при решении заданий семинара, и по мере улучшения качества решения задачи графового представления текста и определения показателей достоверности соответствующих триплетов, мы сможем применить указанную модель поиска и оценить ее~~го~~ качество.

Также планируется выбрать из общего количества запросов лишь те, которые принципиально могут отражать какую-то семантику. Например, в заданиях семинара присутствовали запросы, состоящие из цифр, скобок и других не–буквенных символов. Поиск с учетом семантики текста не может успешно выполняться для таких запросов априори. Таким образом, в дальнейшем поиск и его оценка будут проводиться только для запросов, которые принципиально могут представлять какой-то смысл.

Литература

[1] Среда Описания Ресурса (RDF): Понятия и Абстрактный Синтаксис. Рабчевский Евгений. http://www.w3.org/2007/03/rdf_concepts_ru/Overview.html

[2] Начальное руководство по RDFa. Сергей Щербак. http://shcherbak.net/translations/ru_rdfa_primer_shcherbak_net.html

[3] Рабчевский Е.А., Автоматическое построение онтологий // Научно-технические ведомости СПбГПУ № 4, часть 1-ая стр. 22-26, 2007 . – Санкт-Петербург: Издательство Политехнического Университета 2007.

[4] Домашняя страничка проекта Lucene. http://lucene.apache.org

[5] Домашняя страница библиотеки для работы с онтологиями Jena. http://jena.sourceforge.net

[6] Применение технологий Semantic Web в задаче поиска по коллекциям текстовых документов. Рабчевский Евгений. http://rabchevsky.name/semantic_web_in_IR

[7] Apache Lucene – Scoring.
http://lucene.apache.org/java/2_4_0/scoring.html

SEUS on ROMIP 2009: Lucene framework evaluation and his advance in semantics context

Rabchevsky Evgeny, Rogkov Michail

In the paper we describe project of search engine which uses semantics SEUS (http://seus.rabchevsky.name), and our participation in ROMIP'2009.