Поисковая система “mnoGoSearch”

© Барков А.И.

bar@mnogosearch.org

Аннотация

Настоящая работа является отчетом об участии в конференции РОМИП-2008. Главной целью работы была апробация методов расчета релевантности документа запросу при поиске по Web-страницам и коллекции нормативных документов.

Введение

MnoGoSearch является свободно распространяемой поисковой системой, работающей в операционных системах семейства Unix, предназначенной для организации поиска на одном или многих Web-серверах. Первая версия mnoGoSearch была выпущена в ноябре 1998 под названием UDMSearch. В октябре 2001 года появились коммерческие модификации системы, реализованные для операционных систем Windows. Последние версии системы можно найти на сайте http://www.mnogosearch.org/.

1. Краткое описание системы

mnoGoSearch состоит из двух частей. Первая часть - индексирующий механизм (**indexer**). **Indexer** пробегает по ссылкам и сохраняет в базе данных информацию о документах, терминах и ссылках. Вторая часть состоит из CGI-программы, предоставляющей возможность поиска в данных, собранных **indexer’**ом.

Основные возможности mnoGoSearch включают:

- Поддержку основных протоколов Интернета (HTTP, HTTPS, FTP, NNTP) и работа с локальными файлами;
- Встроенную поддержку документов формата txt, html, xml, а так же возможность подключения внешних программ-конверторов для любых других типов документов, таких как doc, pdf, rtf, xls, ppt, и т.д.;

- Нечёткий поиск на основе синонимов, подстрок, а так же генерации словоформ (падежи, склонения, и т.д.) с использованием словарей грамматического анализатора ispell;

2. Направления развития системы

Последние версии mnoGoSearch позволяют индексировать несколько миллионов документов на одном компьютере, а в версии 3.3 был добавлен модуль кластеризации, распределяющий данные и процессы их обработки между несколькими компьютерами, что позволило создавать поисковые системы по коллекциям, состоящим из нескольких десятков и даже сотен миллионов документов. На таких больших объемах документов задача ранжирования выдаваемых на запрос документов является одной из самых важных, и в настоящее время при разработке mnoGoSearch именно ей уделяется особое внимание. Так, в версии 3.3 формат поискового индекса был расширен, что дало возможность добавление новых важных составляющих в формулу релевантности.

3. Формула релевантности mnoGoSearch

Формула расчета релевантности mnoGoSearch состоит из следующих частей (факторов):

- SectionBreakdown() – функция распределения слов по секциям документа. Эталонным считается документ, где каждое слово из поискового запроса встречается в каждой секции документа (например, в случае HTML документов типовая настройка включает секции title, body, meta keywords, и т.д., которые задаются перед индексацией). При расчете функции распределения слов составляется вектор длиной *количество_секций*количество_слов_в_запросе*. Вектор эталонного документа заполняется единицами. Вектор анализируемого документа заполняется нулями там, где слово не найдено в секции и единицами там, где слово найдено в секции. Затем, вычисляется математическая корреляция между двумя векторами и возвращается в качестве значения фактора SectionBreakdown(). Так, например, в случае запроса из двух слов в поисковой системе, настроенной для работы по трем секциям, размеры векторов будут равны 6. Если оба слова запроса найдены только в title и нигде больше, то в качестве

результата вернется число ~0.57 – величина математической корреляции между векторами (1,1,0,0,0,0) и (1,1,1,1,1,1).

- WordDistance() – функция близости слов. Документы, где слова запроса стоят рядом друг с другом, оцениваются выше, нежели те, где слова “разбросаны” по разным частям документа. Кроме определения непосредственного расстояния между словами, в расчет также берется порядок слов и полные вхождения поисковых фраз.
- MinPos() – функция степени близости первого найденного слова к началу секции документа.
- WordDensity() – функция частоты искомых слов в документе.
- NumWords() – функция общего количества найденных слов.
- WordForm() – функция “морфологического соответствия”. Этой функцией выше оцениваются те документы, в котором слова встречаются в точно такой же форме, как и в запросе, чем документы с другими формами слов запроса (например, другими падежами существительных, временами глаголов, синонимами).

Значения всех перечисленных факторов лежат в диапазоне от 0 до 1. При вычислении каждого фактора используется дополнительный настроечный вектор wf, который позволяет менять веса различных секций документа (например, можно сделать секцию title более значимой, по сравнению с секцией body). Для получения единого численного показателя релевантности значения перемножаются. Степень влияния каждого фактора задается настроечными коэффициентами, а при указании нулевого коэффициента – соответствующий ему фактор в расчете не учитывается.

4. Настройка mnoGoSearch для участия в РОМИП 2008

В 2008-м году мы участвовали в дорожках “поиск по web-коллекции” (коллекции by.web и km.ru) и “поиск по коллекции нормативных документов” (коллекция legal). При настройке системы во всех коллекциях для генерации словоформ был использован словарь русского языка Александра Лебедева (изначально предназначенный для системы грамматической проверки ispell, но с успехом применяемый и в поиске). Система работала в режиме “AND - найти все слова”, автоматический

переход в режим “OR - найти хотя бы одно слово” при нулевом или малом количестве результатов режима “AND” не осуществлялся.

Для коллекций by.web и km.ru использовалась настройка с секциями body, title, meta keywords и meta description. Вес всех секций считался одинаковым. Коэффициент функции частоты слов WordDensity был установлен в 200 (при возможном диапазоне 1..255). Коэффициент функции количества слова NumWord был установлен в 1 (при диапозоне 0..255). Коэффициент функции WordDistance был установлен в 2500 (при официальном диапазоне 0..255, однако в реальности этот параметр позволяет задавать и большие значения без переполнения разрядной сетки при расчетах). Коэффициент функции MinPos был равен 0 (по умолчанию), то есть этот фактор не учитывался. Также, был использован коэффициент по умолчанию у функции WordForm (255), то есть система не делала предпочтения точным формам слов запроса перед падежными, временными формами (и т.д.). Синонимы не использовались. Такая настройка является типовой настройкой mnoGoSearch для поиска по web-коллекции, за исключением увеличенного влияния функции расстояния между словами.

Участие в поиске по коллекции нормативных документов – наш первый опыт. Мы попытались произвести более тонкую настройку с учетом особенностей коллекции. Так, заголовки между тэгами **<P ID=”P0000”>** и **</P>** помимо body были выделены и в отдельные секции, то же самое было проделано с заголовками с ID P0001-P0006. Веса секций, соответствующих этим заголовкам специально не увеличивались, однако факт нахождения слов как в body, так и в одном из P000? делает эти слова более значимыми, поскольку увеличивают значение функции распределения слов по секциям (SectioBreakdown). Еще следует отметить, что в коллекции legal были подключены синонимы, позволяющие находить нечеткие даты, чтобы, например, документ с заголовком “Закон от 1 января 2008 года” был найден при запросе “Закон от 01.01.2008”. В коллекции legal были использованы коэффициенты функций-факторов релевантности, аналогичные web-коллекциям.

5. Анализ результатов

mnoGoSearch показал разные результаты на разных коллекциях. На коллекции WEB.BY наш результат был стабильно на 6-ом и 7-м местах по значению различных метрик (среди 15 участников), а метрика Precision заняла 3-е место. На коллекции KM.RU был

получен лучший результат по метрике Precision(5) и второй результат по метрике Precision(10) среди 15-ти предоставленных результатов. Однако результаты по остальным метрикам были слабыми и колебались между 10-м и 11-м местом.

Хуже всего система показала себя на коллекции Legal. Это было ожидаемым, поскольку, во-первых, мы первый раз участвуем в этой дорожке, а во-вторых, поиск среди нормативных документов является необычным применением mnoGoSearch. По большинству метрик был показан 5-й результат из 6-ти предоставленных, лишь по одной метрике (Precision) удалось подняться до 3-го места.

По совокупности результатов из трех дорожек можно сказать, что система выступила в целом неплохо. Поскольку на коллекции legal был получен худший результат, мы, прежде всего, провели детальный анализ для поиска причин неудачи именно на этой коллекции. Наша система вообще не смогла найти 722 документа из 3601 помеченных как "vital" (ни среди 100 лучших, ни даже среди остальных результатов, выданных системой).

87 документов (12%) были потеряны по причине аббревиатур, например, ГК = ГРАЖДАНСКИЙ КОДЕКС, ФЗ - ФЕДЕРАЛЬНЫЙ ЗАКОН, и т.д. Из этого можно сделать вывод, что для успешного участия в следующих сезонах нам, безусловно, понадобится словарь аббревиатур из соответствующей предметной области.

81 документ (11%) был потерян в результате ошибки в функции расчета близости слов - в некоторых ситуациях получался нулевой результат, и такие документы вообще отбрасывались как нерелевантные. Причем, ошибка в большинстве случаев произошла на запросах с двумя словами, а при более длинных запросах таких сбоев практически не возникало.

Следующая причина потери - 71 документ (9%) - упрощенная реализация генератора словоформ. mnoGoSearch подключает файлы от системы ispell, предназначенной для проверки орфографии. Используя словари ispell, нельзя получать разные части речи. Так, mnoGoSearch не нашел документы с прилагательным "Ленинградский" при запросе "Ленинград". Это не является проблемой самого ispell, поскольку при проверке орфографии перехода между частями речи не требуется, но, как показали результаты - это важно для поиска. Сделан вывод о необходимости подключения более сложных систем для генерации словоформ.

61 документ (8%) не был найден в результате, как оказалось, неправильного использования файлов от ispell. После консультации с авторами ispell проблему удалось устранить.

Заключение.

Анализ результатов участия в РОМИП-2008 позволил увидеть как достоинства, так и недостатки нашей поисковой системы, что неоценимо для правильного выбора направлений дальнейшей работы. Поэтому считаем, что участие в конференции оказалось для нас плодотворным.

Хотим выразить благодарность оргкомитету за предоставленную возможность участия в конференции РОМИП-2008, а также за быструю помощь при возникновении текущих вопросов и затруднений. В частности, хотим поблагодарить Игоря Некрестьянова и Марину Некрестьянову.

Search engine “mnoGoSearch”

Barkov A.I.

This article presents a report on experiments in full text retrieval made as a part of ROMIP’2008. The main goal of these experiments was to approbate methods of document ranking implemented in mnoGoSearch throughout the last years.