Поисковая система “mnoGoSearch” на РОМИП-2009

© Барков А.И.

bar@mnogosearch.org

Аннотация

Настоящая работа является отчетом об участии в конференции РОМИП-2009. Главной целью работы была апробация методов расчета релевантности 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. Последние версии системы можно найти на сайте http://www.mnogosearch.org/. В этом году система mnoGoSearch участвовала в конференции РОМИП в третий раз.

1. Развитие системы

В последних версиях mnoGoSearch улучшение ранжирования документов является одной из наиболее приоритетных задач. Так, в версии 3.3 формат поискового индекса был расширен, что дало возможность добавления новых важных составляющих в формулу релевантности.

Версия с новым форматом поискового индекса была впервые апробирована на РОМИП-2008, а результаты анализа данных прошлого года были использованы при подготовке к РОМИП-2009.

2. 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. Кроме непосредственно определения 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, бывает полезна при анализе заголовка (title).
- TF() – функция частоты искомых слов в секции документа.
- IDF() – обратная частота документа. Эта функция была добавлена в mnoGoSearch в 2009-м году и тестируется впервые.
- NumWords() – функция общего количества найденных слов.
- WordForm() – функция “морфологического соответствия”. Эта функция оценивает выше такие документы, слова которых встречаются в точно такой же форме, как и в запросе, а документы с другими формами слов (например, другими падежами существительных, временами глаголов, синонимами) оцениваются ниже.

Значения всех перечисленных факторов лежат в диапазоне от 0 до 1. При вычислении фактора SectionBreakDown() 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.

3. Настройка mnoGoSearch для РОМИП-2009

В 2009-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).

В прогонах, представленных для анализа, система работала только в режиме “AND - найти все слова”. Автоматический переход в режим с менее строгими критериями поиска при нулевом или малом количестве результатов режима “AND” не осуществлялся. Однако менее строгие режимы поиска были испытаны в дополнительных прогонах, проделанных уже по получении результатов оценки.

3.1. Настройка mnoGoSearch для Web-коллекции

Для коллекций By.Web и KM.RU было представлено по одному прогону. Использовалась настройка с секциями body, title, meta keywords и meta description. Вес всех секций считался одинаковым. Коэффициент функции частоты слов ID() был установлен в 200 (при возможном диапазоне 1..255). Коэффициент функции количества найденных слов NumWord() был установлен в 1 (при диапазоне 0..255). Коэффициент функции WordDistance() был установлен в 2500 (при “официальном” диапазоне 0..255, однако в реальности этот параметр позволяет задавать и большие значения). Коэффициент функции MinPos() был равен 0 (по умолчанию), то есть этот фактор не учитывался. Также был использован коэффициент по умолчанию для 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.

В дополнение к перечисленным параметрам, в 2009-м году была добавлена функция IDF() с коэффициентом 8 (при диапазоне 0..255).

3.2. Настройка mnoGoSearch для коллекции Legal

mnoGoSearch участвовала в поиске по коллекции нормативных документов второй раз. Для анализа были представлены два прогона. Прогон Khaki_1 (xxxx-2) полностью повторял настройку прошлого года. Прогон Khaki_2 (xxxx-1) дополнительно использовал новую функцию IDF() с коэффициентом 48. Коэффициенты остальных функций релевантности были аналогичны настройке для Web-коллекций.

Как и в прошлом году, заголовки между тэгами **<P ID="P000x">** и **</P>** были выделены в отдельные секции (для всех ID P0000-P0006). Веса секций, соответствующих этим заголовкам, специально не увеличивались. Однако появление слов как в body, так и в одном из P000x, делает эти слова более значимыми, поскольку увеличивает значение функции распределения слов по секциям SectionBreakdown.

Оба прогона на коллекции legal использовали синонимы, позволяющие находить нечеткие даты. Например, документ с заголовком "Закон от 1 января 2008 года" будет найден и при запросе "Закон от 01.01.2008". Также был подключен созданный по результатам 2008-го года словарь основных сокращений (рф = российская федерация, ст = статья и тд).

4. Анализ результатов

Согласно результатам оценки, mnoGoSearch показал по метрике Recall 13-е место (из 17-ти участников) в коллекции By.Web (прогон xxxx-1) и 8-е место (из 12-ти) в коллекции KM.RU (прогон xxxx-1).

В коллекции Legal заняты 13-е (из 18-ти) место прогоном xxxx-2 (с настройками прошлого года) и 4-е место прогоном xxxx-1 (с новыми настройками). Таким образом, подтверждается важность функции IDF(), которая прежде отсутствовала в формуле ранжирования mnoGoSearch. Прогон xxxx-1 показал 5%-е улучшение метрики Recall.

По метрике Precision в Web-коллекциях были заняты места: 3-е (из 17-ти) в коллекции By.Web, 5-е (из 12-ти) в коллекции KM.RU. В дорожке нормативных документов удалось занять 3-е место (прогон xxxx-2 с настройками 2008-го года) и 1-е место (прогон xxxx-1 с новыми настройками).

Высокие показатели Precision по сравнению с показателями Recall объясняются строгим режимом поиска (AND – найти все слова).

После получения результатов оценки мы проделали несколько дополнительных прогонов, отличительной особенностью которых был переход к более мягкому режиму поиска при недостаточном количестве документов, найденных в строгом режиме AND.

Как показали эксперименты, переход в режим OR (найти любое слово) выдавал слишком много документов, и значительная часть “хороших” документов оказывалась за границей пула 100. Поэтому мы запрограммировали новый режим поиска и назвали его AND-MINUS-HALF. В этом режиме система находит документы, в которых присутствует хотя бы половина искомых слов запроса (с округлением в большую сторону при нечетном количестве слов). Например, при запросе “социальный статус человека”, найденными считаются документы, в которых встречаются хотя бы два слова. Переход в менее строгий режим в дополнительных прогонах осуществлялся только в тех случаях, когда строгий режим выдавал менее 80-ти документов.

Также, по результатам анализа, был реализован новый режим поиска составных термов типа *ВТБ24, ГОСТ19903, Дом2*. В этом режиме, помимо поиска термов, заданных буквально, система также ищет и фразы “*ВТБ-24”*, “*ГОСТ-19903”*, “*Дом-2”*.

Дополнительные прогоны улучшили показания Recall во всех коллекциях. Так, в Web-коллекциях значение Recall улучшилось на 20-25%, что позволило подняться с 13-го до 10-го (By.Web) и с 8-го до 7-го (KM.RU) места и по абсолютным показателям подойти близко к лидерам. Значение Recall в коллекции Legal улучшилось на 10%, что позволило подняться с 4-го до 2-го места.

Улучшение метрики Recall в дополнительных прогонах, однако, привело к некоторому ухудшению показаний Precision, что было ожидаемым.

Таким образом, введение новых составляющих в функции расчета релевантности и менее строгого режима поиска в целом положительно сказалось на работе системы.

Можно сделать предположение, что лучшие показатели на коллекции Legal (по сравнению с коллекцией Web) достигнуты благодаря специфике запросов. В коллекции нормативных документов преобладают четкие запросы, имеющие фрагменты названий документов, номера статей, даты. Именно на таких запросах MnoGoSearch показывает себя лучше.

В запросах для коллекции Web документов присутствует бОльшая степень нечеткости. Можно сделать вывод о необходимости менее строгой настройки mnoGoSearch для Web коллекции в будущем.

5. Планы

Анализ результатов подсказывает возможные дальнейшие пути развития системы.

Требует улучшения формула расчета близости слов WordDistance(). Есть предположение о слишком заниженном показателе качества пассажа, состоящего из двух термов запроса, при длине запроса больше двух (т.е. при наличии под-фраз из двух слов).

Также можно повысить качество таких пассажей, как "…плоские и круглые черви…" при ответе на запрос "ПЛОСКИЕ ЧЕРВИ" (arw34551). В этом пассаже перечисляются альтернативы плоские - круглые, разделенные союзом "и", что можно считать эквивалентным (или почти эквивалентным) наличию искомой фразы "плоские черви". Такой же подход будет справедлив и в случае союза "или".

Следует отметить, что дополнительные прогоны не нашли некоторые хорошие документы в случае редких слов. Так, при ответе на запрос "алкогольная компания ООО Мозель-М г. Москва" (arw52561) слово "Мозель" встречается только в 33-х документах коллекции. Документ ID 301676, помеченный как vital и содержащий это слово, оказался только на 1270-м месте, то есть вообще не попал в пул. Система посчитала более важными документы, содержащие менее уникальные слова запроса, но при этом встречающиеся одновременно в секциях title, keywords, description. Требуются эксперименты с целью выяснения: следует ли усилить влияние функции IDF() для особо редких слов, или же необходимо в целом уменьшить влияние функции SectionBreakdown() по сравнению с функцией IDF().

Как и в прошлом году, еще раз подтвердилось несовершенство морфологического модуля. Словари от Ispell не позволяют делать переход между частями речи (Чернобыль/Чернобыльский) и между близкими словами (психология/психолог). Также использованный словарь не включает правил склонения фамилий (Бабич, Бальмонт), названий (Титаник, Сургутгазпром) и имеет ряд других недостатков, что может ухудшать значение метрики Recall на несколько процентов.

Следует улучшить показатели качества при ответе на запросы, состоящие из одного слова. На таких запросах mnoGoSearch демонстрирует низкие результаты по сравнению с другими системами. Как показывает предварительный анализ, качество поиска можно частично улучшить усилением влияния функции ID(),

отвечающей за частоту слова в секции документа. Также желательна проверка гипотезы, высказанной в работах участников РОМИП-2008, о том, что принятие во внимание близости слова к началу предложения позволяет улучшить качество поиска.

Среди возможных дополнительных путей улучшения показателей можно отметить:

- использование словаря брэндов в русской транскрипции (Ямаха = Yamaha, Шевроле = Chevrolet, Форд = Ford, Самсунг = Samsung), который бы включал популярные марки автомобилей, сотовых телефонов, бытовой техники и т.д.
- использование словаря числительных (20-летие = двадцатилетие)
- автоматическая коррекция слов с опечатками

Заключение.

Анализ результатов участия в РОМИП-2009 позволил увидеть как достоинства, так и недостатки нашей поисковой системы, что неоценимо для правильного выбора направлений дальнейшей работы. Считаем, что участие в конференции оказалось для нас плодотворным.

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

Search engine “mnoGoSearch” at ROMIP-2009

Barkov A.I.

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