KM.RU на РОМИП-2007

© Сергей Татевосян, Наталья Брызгалова

«КМ онлайн»
{Tatevosyan, Bryzgalova}@post.km.ru

Аннотация

Настоящая статья посвящена апробации новой поисковой системы, разработанной KM.RU. В работе дано краткое описание алгоритма и результатов, полученных по итогам его тестирования в дорожке «Поиск по запросу по веб-коллекции».

1. Введение

KM.RU первый раз принимает участие в семинаре РОМИП. На РОМИП-2007 мы ставили перед собой задачу протестировать новый алгоритм поиска и ранжирования документов в ситуации поиска по веб-коллекции и получить объективную оценку результатов.

Цель статьи – рассказать об основных принципах работы алгоритма и проанализировать итоги тестирования на материалах, предоставленных в рамках семинара.

2. Алгоритм поиска и ранжирования документов

Настоящая система используется для задач поиска релевантных документов в веб-коллекции по запросу пользователя и ранжирования найденных документов по степени соответствия обработанному запросу. В процессе работы лежащего в основе системы алгоритма производится оценка ключевых элементов ситуации информационного поиска:

- *текст запроса*: все слова из запроса пользователя считаются ключевыми;
- *веб-документ*: определяется значимость или «вес» данного документа в коллекции;

- *текст документа*: определяется присутствие в документе ключевых слов, дополнительно учитывается расстояние между словами запроса, найденными в веб-документе;
- *html-разметка*: учитываются элементы разметки, позволяющие выделить наиболее значимые и в смысловом отношении нагруженные части веб-документа;
- *гиперссылки*: учитывается наличие гиперссылок с других документов на данный.

2.1 Общая формула релевантности

Для вычисления релевантности документа запросу мы использовали следующую зависимость:

W = W1 + W2 + W3 (1),
где W – итоговое значение релевантности документа.

Остановимся подробнее на каждом из слагаемых.
W1 = TF*IDF(Doc) * F1(DocWeight)
TF*IDF(Doc) вычисляется по известной модификации BM25 [1]. Дополнительные баллы документ получает за наличие слов в заголовке (Title);
F1(DocWeight) – функция от веса документа, вычисленного по схеме, предложенной в [2].
Особенности функции:
а) F1 в том числе занимается приведением значения DocWeight до нужного диапазона;
б) Часть ссылок признаются неинформативными и в расчете не участвуют.

Фактически, W1 отвечает за информационную значимость документа и его вес по отношению к другим документам, вычисленный по [2].

W2 = Σ (TF*IDF(Link)* F2(LinkWeight))
где:
TF*IDF(Link) - TF*IDF ссылки на данный документ;
F2(LinkWeight)) – функция приведения весов ссылок на документ.
LinkWeight вычисляется аналогично DocWeight
Т.о. W2 отвечает за информационную значимость ссылок на данный документ и их веса.

W3 = F3(расст) – функция, отвечающая за учет расстояния между словами запроса в документе. Имеет ненулевое значение при прохождении кворума, вычисляющегося по сумме IDF слов.
При вычислении W1, W2, W3 учитывается ряд дополнительных параметров.
Полученную зависимость (1) мы считаем базовой. Нам было важно увидеть качество ее работы и понять, в каком направлении вести дальнейшие исследования.

3. Проведенные эксперименты и полученные результаты.

3.1. Дорожка «Поиск по запросу по веб-коллекции»

На семинаре РОМИП-2007 мы приняли участие в дорожке «Поиск по запросу по веб-коллекции». В этом году дорожка поиска по вебу проводилась по двум коллекциям, одну из которых предоставил KM.RU. Из особенностей коллекции KM.RU можно выделить такие: 1) присутствие достаточно большого количества нечетких дублей, 2) во многих документах присутствуют блоки информации, являющиеся ссылками или рекламными блоками, т.е. не относящиеся к содержанию документа, 3) многие сайты характеризуются весомой ссылочной структурой. Это же относится и к коллекции BY. Как мы полагаем, указанные свойства позволили участникам дорожки поиска по вебу, и нам в том числе, проверить эффективность работы алгоритмов на материалах веб-коллекции, более близкой к реальному Интернету, нежели коллекция документов narod.ru, на которой проводились эксперименты в предыдущие годы.

Мы сделали два прогона по коллекциям KM и BY с модификацией формулы (1). В первом прогоне слагаемое W3 всегда равно нулю, т.е. учет расстояния между словами запроса в документе не проводился. Во втором прогоне слагаемое W3 вычислялось, как описано выше. Целью первого прогона было выяснить, насколько хорошо базовая функция ранжирует документы. Целью прогона №2 – как учет расстояния повышает качество ранжирования.

3.2. Результаты прогонов

Ниже представлены таблицы и графики, иллюстрирующие результаты наших прогонов на дорожке web ad-hoc для коллекций

KM и BY (оценка аксессоров AND, без ограничения глубины пула). В таблицах приведены значения оценок, посчитанных отдельно для прогонов по каждой коллекции, а не оценки с учетом «сужения» результатов, поскольку мы осуществляли прогоны только по каждой коллекции отдельно, а не по обеим одновременно.

В этом году к стандартным метрикам оценки результатов экспериментов по данной дорожке добавилась метрика bpref и ее модификация bpref-10 [3], более подходящие для работы с неполностью оцененными коллекциями.

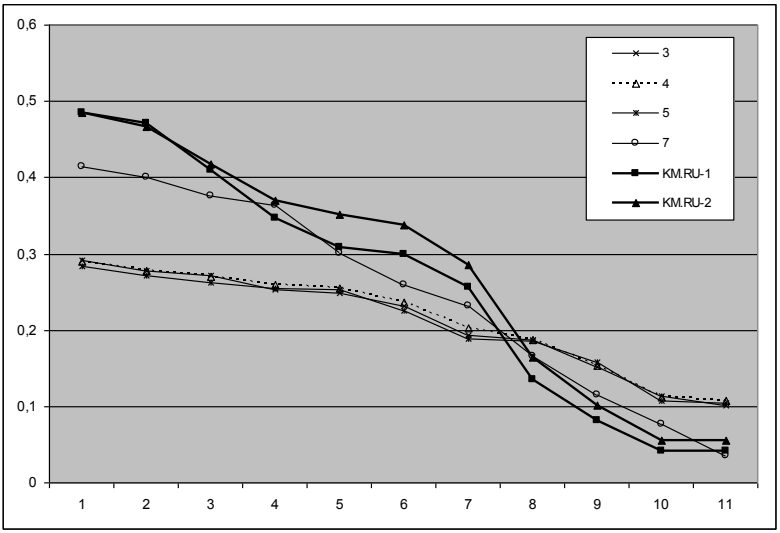

Рисунок 1. Результаты участников дорожки web-adhoc, коллекция km.ru, оценка AND.
KM.RU-1 – модификация алгоритма без учета расстояния между словами (1 прогон)
KM.RU-2 – модификация алгоритма с учетом расстояния между словами (2 прогон)

	3	4	5	7	KM.RU-1	KM.RU-2
Prec(5)	0,18	0,18	0,17667	0,27	**0,33667**	**0,36333**
Prec(10)	0,18333	0,17833	0,16667	0,22	**0,26**	**0,27333**
Recall	0,743	0,7457	0,73027	0,65796	**0,72177**	**0,72177**
Bpref	0,56856	0,58403	0,55651	0,50648	**0,5148**	**0,54349**
Bpref-10	0,62069	0,6329	0,60335	0,53475	**0,57455**	**0,58865**

Таблица 1. Оценки precision(5), precision(10), recall, bpref и bpref-10 для участников дорожки web-adhoc, коллекция km.ru, оценка AND.

Изначально мы тренировали систему на коллекции документов KM.RU, что хорошо видно по результатам. Базовая функция вычисления релевантности дает нам устойчивое первое место вверху графика, функция с учетом расстояния увеличивает разрыв. Завал графика в конце объясняется тем, что мы хотим добиться максимального значения оценки Precision первых документов, для чего намеренно завышаем некоторые параметры, т.е. для нас главными показателями являются Precision(5) и Precision(10).

Новая модификация алгоритма, которую мы тестировали на момент написания статьи, с нашей точки зрения, дает лучшие результаты по всей кривой, в том числе и в конце графика.

Особенно интересным для нас было испытание разработанного алгоритма на незнакомой коллекции. В данном случае это коллекция BY. Ниже приведен график результатов KM.RU по коллекции by, оценка AND. Наша базовая функция (прогон 1) обеспечивает серединное положение результатов, на которое мы и рассчитывали. Функция с учетом расстояния повышает качество ранжирования. В первой половине графика, наверху, мы делим первое-второе места с участником № 2, далее, во второй половине графика, уступаем первое место участнику № 1 и идем по-прежнему рядом с участником №2.

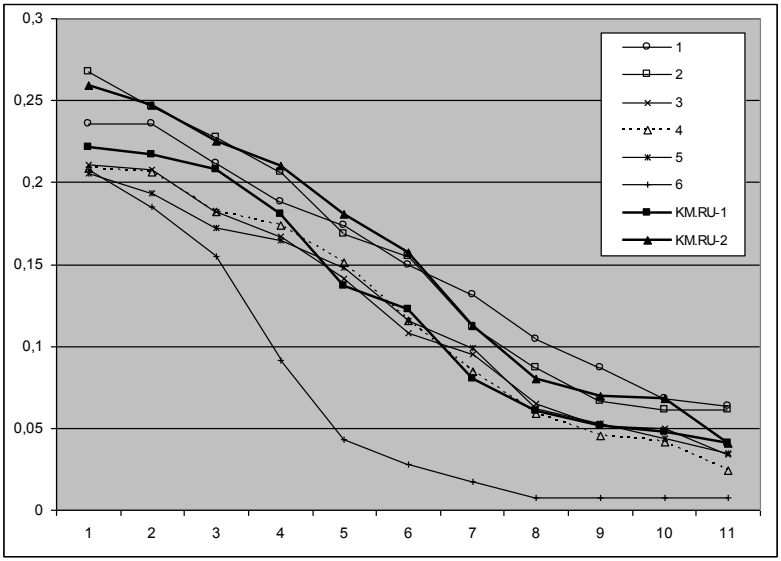

Рисунок 2. Результаты участников дорожки web-adhoc, коллекция by.web, оценка AND.

	1	2	3	4	5	6	KM.RU-1	KM.RU-2
Prec(5)	0,123	0,133	0,087	0,083	0,093	0,097	**0,127**	**0,13**
Prec(10)	0,105	0,095	0,083	0,082	0,085	0,065	**0,1**	**0,113**
Recall	0,865	0,77	0,788	0,793	0,791	0,598	**0,782**	**0,782**
Bpref	0,508	0,52	0,508	0,501	0,509	0,461	**0,488**	**0,517**
Bpref-10	0,558	0,583	0,583	0,58	0,597	0,5	**0,56**	**0,579**

Таблица 2. Оценки precision(5), precision(10), recall, bpref и bpref-10 для участников дорожки web-adhoc, коллекция by.web, оценка AND.

Анализируя графики TREC и таблицы, мы можем сделать следующие выводы:
- как видно на графиках TREC, алгоритм удачно позволяет находить и поднимать в начало выдачи высокорелевантные документы;
- значения Precision(5) и Precision(10) являются наибольшими или делят первое-второе место с другими участниками дорожки;
- графики TREC и таблицы наглядно демонстрируют, что модификация алгоритма, учитывающая расстояние между ключевыми словами при определении релевантности документа, оказалась более эффективной;
- хочется отметить, что на графиках TREC для прогонов по коллекции КМ явно выделяются две группы участников, в каждой группе графики демонстрируют сходные тенденции. Очевидно, такая закономерность может говорить о сходных особенностях алгоритмов, используемых системами из одной группы. Также отметим, что в таблицах с графиками TREC для поиска по белорусскому Интернету деление на группы менее заметно или почти незаметно. Можно предположить, что та особенность устройства алгоритмов, которая в поиске по КМ стала причиной своеобразного разделения участников на группы, связана с каким-либо свойством, которое отличает коллекцию КМ от коллекции белорусского Интернета. Полагаем, что этим свойством может быть число ссылок внутрь коллекции.

При оценке результатов мы, в первую очередь, учитывали те задачи, для которых используется система, а именно: задачи информационного поиска в условиях «живого» Интернета. Рассматривая результаты с этой точки зрения, мы можем сказать,

что в целом удовлетворены проведенными экспериментами: разработанный алгоритм позволяет поднять на первые места высокорелевантные документы, что при условии большой коллекции и значительного числа релевантных документов может обеспечить хорошее начало поисковой выдачи. Полагаем, что не очень высокие показали по метрикам bpref и bpref-10 и невысокие оценки полноты объясняются особенностью алгоритма, который настроен не на абстрактно хорошие результаты, а на хорошую выдачу в ситуации особого типа - информационный поиск в сети Интернет. Таким образом, из всех оценок наших прогонов наиболее важными для нас были оценки Precision(5), Precision(10) и 11-точечный график, построенный по методике TREC, а не оценки полноты и bpref, bpref-10, характеризующие выдачу целиком.

4. Актуальность результатов экспериментов для информационного поиска в сети Интернет

В настоящей статье мы говорили об эффективности разработанного алгоритма в отношении поиска по коллекциям, представленным в рамках дорожки поиска по вебу на семинаре РОМИП. Коллекция документов всей сети Интернет отличается от коллекции РОМИП не только размерами, но и присутствием разнообразных типов сайтов, которые, к сожалению, не представлены в наборе сайтов КМ и белорусского Интернета, как-то: Интернет-магазины, сайты, «раскрученные» под конкретные поисковые системы, поисковый спам – об этом уже говорили в предыдущие годы участники семинара [4]. Безусловным плюсом семинара в этом году мы считаем присутствие в коллекциях дублей и развитой ссылочной структуры. К сожалению, все запросы, по которым отрабатывали системы на дорожке поиска по вебу, относились скорее к информационным, поэтому у нас не было возможности проверить, насколько успешно алгоритм находит релевантные документы с опорой в основном на ссылочную структуру коллекции. Мы предлагаем в будущем добавить к списку запросов запросы-маркеры: это могут быть, например, сайты компаний, которые находятся поисковыми системами по ссылкам, содержащим название компаний, или сайты, посвященные определенной тематике, - «кино», «почта», «хостинг».

5. Заключение

Участие KM.RU в семинаре РОМИП'2007 можно считать результативным: мы получили объективную оценку работы нашего поискового алгоритма, смогли сравнить результаты работы системы с результатами других участников и определить для себя те элементы системы, которые оказались эффективными, и те, которые требуют доработки. Не сомневаемся, что итоги экспериментов и их анализ помогут нам в будущем усовершенствовать существующий поисковый алгоритм.

Литература

[1] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3 // In TREC-3, 1994.

[2] S. Brin, L. Page. The anatomy of a large-scale hypertextual Web search engine. http://infolab.stanford.edu/~backrub/google.html

[3] Ch. Buckley, El. Voorhees. Retrieval evaluation with incomplete information. http://www.nist.gov/itl/iad/IADpapers/2004/p102-buckley.pdf

[4] А. Гулин, М. Маслов, И.Сегалович. Алгоритм текстового ранжирования Яндекса на РОМИП-2006, 2006. http://www.romip.ru/romip2006/03_yandex.pdf

KM.RU at RIRES-2007

S. Tatevosyan, N. Bryzgalova

The present article is devoted to the introduction of a new information retrieval system at RIRES-2007. The paper contains a brief description of the system and reports on the results of experiments run in Web adhoc track.