КМ.RU на РОМИП-2008. Оптимизация параметров поискового алгоритма

© Сергей Татевосян, Наталья Брызгалова
«КМ онлайн»
{tatevosyan, bryzgalova}@post.km.ru

Аннотация

В статье описаны модификации алгоритма информационного поиска, представленного КМ.RU на семинаре РОМИП-2007, и система оптимизации коэффициентов для определенного набора параметров, созданная с целью получения наилучших результатов в поиске и ранжировании документов. Освещаются результаты участия проекта «Поиск КМ.RU» в семинаре РОМИП-2008. Обсуждается дальнейший путь развития проекта.

1. Введение

Работа над улучшением качества поиска и ранжирования документов обычно связана с появлением новых факторов, которые учитываются при работе с документами в ответ на запрос. Мы постарались создать для себя удобный автоматический способ определения значимости определенных факторов при обработке запроса, а также способ оптимизации параметров нашего поискового алгоритма. В качестве базы для оптимизации мы использовали 1) материалы прошлогоднего семинара РОМИП: пары запрос-документ и оценки экспертов как меры соответствия документа запросу, 2) базу собственных оценок.

Задача, которая стояла перед нами в этом году: определить значимость новых факторов, а также выяснить, помогает ли оптимизация настроек улучшить результаты работы системы. Подобную работу в 2005 году уже проводили участники семинара РОМИП-2005 [4], и их результаты говорили о преимуществе оптимизированного алгоритма.

2. Характеристика коллекций

Смысл документа может быть оценен по трем составляющим:

1. Заголовок (Title).
2. Текст документа.
3. Ссылки на документ.

Исходя из этого, опишем свойства коллекций:

1. Legal – наибольшая степень структурированности. Фактически, заголовок играет наиболее существенное значение в определении релевантности документа. Ссылки достоверны.
2. KM.RU – структурирована, но меньше чем legal. Документы могут содержать информацию, не относящуюся к основной теме документа (текущие новости и т.п.). Заголовок обычно достоверен. Ссылки обычно достоверны.
3. BY.WEB – наименьшая степень структурированности. Документы могут содержать информацию, не относящуюся к основной теме документа. Заголовок может быть недостоверным. Ссылки могут быть недостоверны.

Для каждой составляющей смысла документа можно провести ряд операций, дающих наиболее эффективную оценку веса документа.

Приведем классификацию современных методов выставления веса документу.

1. Произведение частота слова из запроса в документе – TF и встречаемости слова в коллекции – IDF. Чем больше статистика, тем (в общем случае) она лучше.
2. Оценка расстояния между словами запроса. Некая функция от расстояния – F (расстояние). Пассажи (полные, неполные, с учетом порядка слов и без, с расстоянием между словами и без), пары слов (с точным порядком слов, с обратным порядком слов, со словами между парой слов).
3. Количественные оценки – вес документа и вес ссылок, рассчитанный по ссылочному рангу.

Все указанные методы могут быть применены для количественной оценки 3-х смысловых частей документа – заголовка, текста и ссылок на документ.
Соответственно, при таком подходе, например, можно отдельно оценивать пассажи в заголовках, тексте документа и ссылках на документ. Там же учитывать и пары слов.

Указанные методы гораздо лучше работают на хорошо структурированных коллекциях и недостаточно хорошо на плохо структурированных.

3. Модификация алгоритма поиска и ранжирования документов. Автоматическая оптимизация коэффициентов

3.1 Структура алгоритма

а) базовая формула;

За основу мы взяли формулу, работа которой была продемонстрирована в прошлом году [2]. Этот алгоритм с некоторыми улучшениями сейчас работает на портале KM.RU.

б) модификации формулы.

Добавление новых факторов, автоматическая оптимизация коэффициентов.

Для вычисления релевантности документа запросу мы использовали следующую зависимость:

W = k1*W1 + k2*W2 + k3*W3 + k4*W4(1),
где W – итоговое значение релевантности документа.
k1, k2, k3, k4 – коэффиценты.

Остановимся подробнее на каждом из слагаемых.
W1 = TF*IDF(l) * F1(DocWeight),
где:
TF*IDF(l) вычисляется по

$$tf_d(l) = \text{freq}_d(l)/(\text{freq}_d(l) + 0.5 + 1.5*\text{dl}_d/\text{avg\_dl})$$

$freq_D(l)$ - частотность леммы l в документе, dl_D – мера длины документа, avg_dl – средняя длина документа (для коллекций

BY.WEB и KM.RU мы использовали значение avg_dl = 400, для коллекции Legal = 1500),

IDF - “Inverse term frequency” - форма штрафования часто используемых в коллекции слов:

$$idf(l) = \log((|c| + 0.5)/df(l))/\log(|c| + 1)$$

где $|c|$ - количество документов в коллекции, $df(l)$ - количество документов, где встретилась лемма l.

В итоговое значение *tf*idf* входят, помимо обычной встречаемости слова в документе, надбавки за присутствие слов в выделенных областях (title, заголовки типа h1-h4 и т.п.). В прошлом году для упрощения работы мы учитывали только встречаемость слов в title. В этом году ввели остальные параметры, отвечающие за форматирование слов в документе (h1-h4 и проч.).

F1(DocWeight) – функция от веса документа, вычисленного по схеме, предложенной в [1].

Особенности функции:
а) F1, в том числе, занимается приведением значения DocWeight до нужного диапазона, фактически, нормировкой. *Действие функции на вес документа сильно зависит от способа нормировки, что в итоге существенно влияет на порядок документов в выдаче*;
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 схеме ссылочного ранжирования, описанной в [1].

$$W2 = \Sigma (TF*IDF(Link)* F2(LinkWeight)),$$

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(расстояние) – функция, отвечающая за учет расстояния между словами запроса в документе. Имеет ненулевое значение при прохождении кворума. Далее об этом подробно.

В прошлом году в F3(расстояние) входил пассаж из запроса, встреченный в документе.
Свойства пассажа:
1) Пассаж может быть неполным. Он вычисляется по кворуму суммы IDF входящих в него слов и по кворуму числа слов. Оба параметра задаются в настройках. Например: пассажем считается тот, в который вошло 70% слов запроса, и их вес не менее 60% от суммы IDF всех слов запроса. Точные значения подбираются с помощью оптимизации параметров.
2) Порядок слов в пассаже не имеет значения. Мы намеренно ввели эту особенность пассажа, принимая в расчет свойства русского языка, по которому смысл часто не зависит от порядка слов (хотя и не всегда). Расчет идет на то, чтобы не исключать добавку для документов, где слова из запроса следуют в другом порядке. Плюс, что существенно для поиска по вебу, большая добавка за жесткий порядок слов дает большое поле деятельности для спама поисковых машин.

а) Пассаж вычисляется только для слов, входящих в одно предложение.
б) Расстояние между словами из запроса в документе не должно превышать максимального окна. Например, 10 или 15.
в) В этом году мы ввели дополнительное свойство пассажа: пассажем считается тот, для которого _все_ слова из запроса встретились в данном предложении +/- N предложений. За это документ получает дополнительную надбавку. В прогонах РОМИП мы использовали значение N = 1. Можно использовать бОльшее значение (например, 2), но в наших экспериментах оно дало худший результат.

Дополнительно в этом году были введен следующий параметр, зависящие от расстояния:

4) Пары слов. Вес по парам слов вычислялся как описано в [3]. При этом, понимая, что в ряде случаев точный пассаж лучше описывает смысл документа, мы парами слов учитываем в том числе и порядок слов в

пассаже. В наших экспериментах параметр дал прибавку в качестве 5%.

W4 - группа дополнительных параметров (введена в этом году):

1) Близость слов из запроса к началу предложения.

До сих пор мы не встречали описание этого фактора в работах по информационному поиску.

Основание

В русском языке тема (известная информация) предложения выражается в том числе и порядком слов. Хотя в русском языке слова могут стоять в любом порядке, прямой порядок слов является более употребительным. Таким образом, то, о чем идет речь (тема), - обозначено в начале фразы. При обработке запроса мы давали добавку документу, где ключевые слова встречаются в начале предложения. В наших экспериментах параметр дал прибавку в качестве 5%;

2) Встречаемость в документе точных словоформ из запроса.

Наиболее существенными в плане улучшения качества оказались факторы: 1. Пары слов 2. Близость слов из запроса к началу предложения.

Поскольку появление новых факторов усложняет устройство алгоритма поиска, мы задались целью создать программу автоматического подбора параметров и определения важности того или иного фактора.

3.2 Оптимизация параметров

Мы поставили перед собой задачу создать программу, оптимизирующую коэффициенты перед слагаемыми в формуле релевантности. Программа основывается на экспертных оценках документов.

Базой для настройки системы послужили материалы прошлогоднего семинара РОМИП: запросы для дорожки поиска по веб-коллекции и документы, выданные системами-участницами на данные запросы. Ориентирами, говорящими, какой документ «хороший», а какой «плохой», стали оценки экспертов РОМИП.

Задача программы-оптимизатора – подобрать параметры, которые обеспечивают наилучшие результаты оценки выдачи.

а) принципы, лежащие в основе программы-оптимизатора;

1) Оптимизация делается на основе оценок аксессоров;

2) Для оптимизации параметров должна существовать количественная мера оценки документа – то, что говорит, чем один документ лучше другого. Мы применяли модифицированное значение показателя bpref-10.

3) Оптимизация проходила методом модифицированного координатного спуска.

б) проблемы, возникшие в процессе создания;

Оптимизируемая функция обладает следующими свойствами:

1) Функция кусочная;

2) Функция немонотонная;

3) Функция обладает заведомо бОльшим числом параметров, чем используется для ее вычисления.

Исходя из этих свойств, нахождение глобального максимума является заведомо недостижимым. В такой ситуации можно говорить только о достаточно хорошем локальном максимуме.

В этих условиях мы применили два способа оптимизации параметров:

1) Параметры оптимизируются все сразу;

2) Для наиболее значимых с нашей точки зрения параметров задается несколько начальных значений коэффициента. Для каждого значения коэффициента проводится его оптимизация (оптимизируется только этот параметр, остальные неизменны). Из нескольких «оптимизированных» значений параметра выбирается тот, в котором значение функции релевантности максимально. После чего в функцию релевантности добавляются другие параметры и оптимизируются таким же способом.

Оба способа: и 1, и 2 - построены на координатном спуске, различие в том, что в 1 спуск делается по всем параметрам «по кругу», а в 2 каждый параметр оптимизируется до максимального возможного значения функции, потом найденное оптимальное значение параметра записывается как константа и далее проводится оптимизация следующего параметра.

в) результаты оптимизации параметров

В результате работы способ 2 оказался более эффективным (эффективность в данном случае означает бОльшее значение функции релевантности).

В связи с существенно разными свойствами коллекций KM.RU, BY.WEB и Legal мы проводили оптимизацию коэффициентов для каждой коллекции отдельно.

4. Дополнительные возможности поискового механизма

Для улучшения качества работы мы ввели следующие свойства:

1) Применение словаря сокращений. Пример: по запросу «РФ» ищется «РФ» и «Российская Федерация».
2) Применение списка стоп-слов.

5. Меры по структурированию коллекции и запросов

Известные статистические методы информационного поиска лучше работают на хорошо структурированных данных.
Задача структуризации представляется нам следующей. Дано:

1. Неструктурированная коллекция.
2. Неструктурированные запросы.

Цели:

1. Структурировать коллекцию.
2. Структурировать запросы.

В результате таких преобразований из связи
|«Неструктурированные запросы» → «Неструктурирированная коллекция»| получаем связь |«Структурированные запросы» → «Структурирированная коллекция»|, работать с которой гораздо легче.

Структурирование коллекции
В условиях плохо структурированной коллекции мы предприняли меры по удалению информационного шума из документов. Под информационным шумом мы понимаем то, что не относится к

основному содержанию документов: рекламную информацию, ссылки на материалы на других сайтах и т.п. Для удаления мы применили метод, позволяющий находить оформление страниц сайта [4]. Метод не удаляет весь мусор, но довольно хорошо справляется с задачей.

Структурирование запросов

Под структурированием запросов мы понимаем следующие шаги:

1) Исправление опечаток;
2) Вычленение смысла из запросов;
3) Расширение запросов.

П.1 реализуется с помощью сервиса исправления опечаток.

П.2 удается реализовать только при определенных условиях. Например, из запроса «принтеры москва» можно получить данные для геотаргетинга, для запроса «телефон кафе на пушкинской» провести поиск по базе номеров телефонов и выдать соответствующую информацию. Но часто такую операцию произвести не представляется возможным из-за неопределенности запроса.

П.3 - для прогонов РОМИП мы реализовали словарь сокращений.

6. Участие в семинаре и полученные результаты.

6.1 Дорожки, в которых мы приняли участие

а) дорожки поиска

В этом году, как и в прошлом, мы участвовали в дорожке поиска по веб-коллекции. Поиск осуществлялся отдельно по набору документов KM.RU и отдельно по документам белорусского Интернета (BY.WEB). Мы попробовали себя в новых для нас дорожках: поиска по нормативно-правовой коллекции (Legal) и дорожке поиска по смешанной коллекции, где собраны вместе сразу три набора документов: коллекция KM.RU, коллекция белорусского Интернета и нормативно-правовая коллекция.

б) прогоны

Для дорожки поиска по веб-коллекции (и для коллекции KM.RU, и для коллекции BY.WEB) мы делали два прогона. В первом прогоне использовался алгоритм, который сейчас работает на поиске по порталу KM.RU. Второй прогон осуществлялся с помощью алгоритма с новыми поисковыми факторами и оптимизированными коэффициентами.

Для дорожки поиска по нормативно-правовой коллекции мы делали один экспериментальный прогон. Нам было интересно, как поисковый механизм, оптимизированный для веба, справится со специфической коллекцией.

Для дорожки поиска по смешанной коллекции мы делали один прогон. Мы выясняли, документ из какой коллекции будет на 1-м месте в выдаче по соответствующему запросу. Другие показатели мы в расчет не принимали, т.к. задача была именно такой.

6.2 Полученные результаты, их анализ

Для применения новых факторов мы создали новый механизм поиска и ранжирования документов. К сожалению, часть опытов к срокам сдачи результатов провести не успели. Тем не менее, продолжили исследования, результаты которых, наряду с официальными данными, представляем ниже.

Коллекция BY.WEB

Основной упор в этом году мы делали на качество поиска по коллекции BY.WEB

У нас получились следующие результаты:

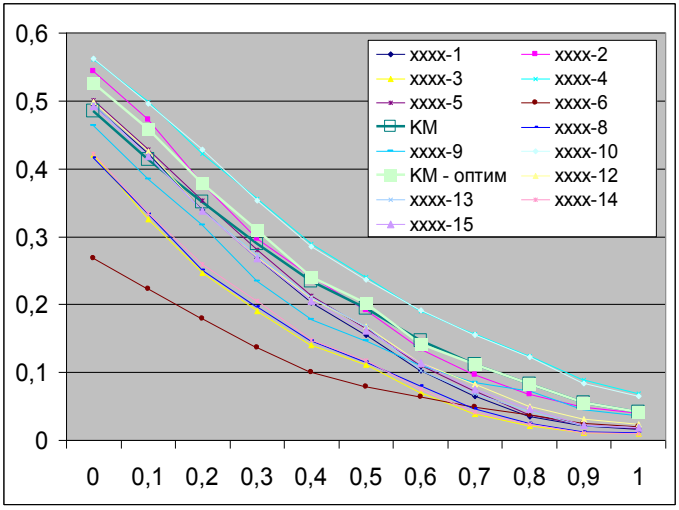

График TREC, оценка OR для коллекции BY.WEB

«KM» - результаты работы алгоритма, работающего на портале KM.RU, «KM оптим» - результаты нового алгоритма.

Приведем результаты по основным показателям:

Участник	Prec(5)	Prec(10)	Bpref-10	Bpref	Recall
xxxx-1	0,3	0,26	0,24	0,19	0,37
xxxx-2	0,33	0,3	0,28	0,22	0,43
xxxx-3	0,24	0,2	0,17	0,14	0,24
xxxx-4	0,33	0,31	0,33	0,26	0,55
xxxx-5	0,31	0,26	0,24	0,2	0,38
xxxx-6	0,17	0,15	0,13	0,11	0,2
КМ	**0,31**	**0,28**	**0,26**	**0,22**	**0,38**
xxxx-8	0,25	0,21	0,17	0,15	0,25
xxxx-9	0,29	0,27	0,22	0,19	0,31
xxxx-10	0,34	0,32	0,33	0,26	0,55
КМ-оптим	**0,34**	**0,3**	**0,27**	**0,23**	**0,4**
xxxx-12	0,31	0,26	0,24	0,2	0,39
xxxx-13	0,29	0,26	0,24	0,19	0,38
xxxx-14	0,25	0,21	0,17	0,15	0,25
xxxx-15	0,3	0,27	0,24	0,19	0,39

По показателю Precision(5) у нас результат на уровне лидера (прогон xxxx-10), что соответствует нашим целям. По Precision(10), Bpref и Bpref-10 - достаточно хорошие значения. По показателю Recall находимся в основной группе. Отставание в Recall от лидеров предположительно основывается на неиспользовании нами в прогонах поиска по кворуму.

После сдачи результатов в РОМИП мы продолжили исследования. Предпринимались следующие шаги:

1. Убрали использование стоп-слов.

Интересно, что, проверив этот режим на разных запросах и коллекциях (BY, KM.RU с запросами 2007 и 2008 годов) везде мы получили результат лучший, чем при исключении стоп-слов из поиска.

2. Включили режим исправления опечаток и использования кворума. При этом кворум по умолчанию не работал, а подключался только при следующих условиях: а) отсутствие документов в выдаче; или б) отсутствие пассажа в документах выдачи.

Результаты нашей работы представлены на графике ниже.

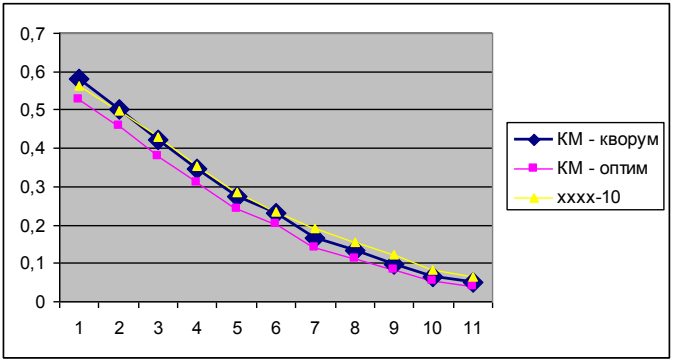

График TREC по BY.WEB после отработки модулей исправления опечаток и поиска по кворуму (делался вне РОМИП после сдачи основных результатов)

После применения экспериментальных модулей мы улучшаем график, который первые шесть точек идет наравне с лидером, потом становится несколько хуже, но лучше графика «КМ – оптим». Новые модули нуждаются в оптимизации, и обкатка их на данных РОМИП, пусть и в неофициальном режиме, позволила нам увидеть их перспективность.

Изучая поведение графика TREC и показателей Precision, мы сделали несколько выводов относительно работы нашего алгоритма на запросах 2007 и 2008 года к коллекции BY.WEB:

1) Показатели практически не зависят от применения или нет ссылочного ранжирования (как весов документов, так и текстов ссылок).
2) Применение в ссылках пар слов и пассажей не дает улучшения качества.
3) Применение пассажа в заголовке (Title) не дает улучшения качества.

Коллекция KM.RU

У нас получились следующие результаты:

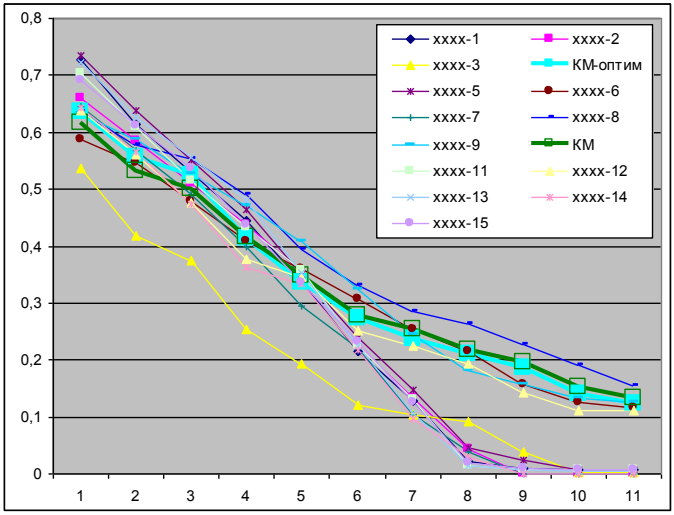

График TREC, оценка OR для коллекции KM.RU
«КМ» - результаты работы алгоритма, работающего на KM.RU, «КМ оптим» - результаты нового алгоритма.

Видно, что графики участников разделились на две группы. У первой группы графиков резко задрано начало, соответственно, первые точки у них выше. Вторая группа графиков – горизонтальная. Начальные точки на их графиках ниже. Группы пересекаются после второй точки, и далее графики первой группы стремительно падают, а второй плавно снижаются. Думаем, что деление на группы связано с особенностями алгоритмов участников.

Приведем результаты по основным показателям:

№	Pr(5)	Pr(10)	Bpref-10	Bpref	Recall
xxxx-1	0,54	0,43	0,29	0,27	0,42
xxxx-2	0,54	0,45	0,28	0,27	0,33
xxxx-3	0,40	0,33	0,20	0,19	0,28
КМ-оптим	**0,53**	**0,45**	**0,36**	**0,34**	**0,43**

xxxx-5	0,55	0,46	0,31	0,29	0,45
xxxx-6	0,42	0,40	0,34	0,32	0,50
xxxx-7	0,50	0,43	0,27	0,26	0,31
xxxx-8	0,49	0,45	0,39	0,37	0,58
xxxx-9	0,52	0,48	0,37	0,34	0,46
KM	**0,49**	**0,44**	**0,36**	**0,34**	**0,43**
xxxx-11	0,60	0,46	0,29	0,27	0,40
xxxx-12	0,48	0,44	0,34	0,30	0,40
xxxx-13	0,56	0,44	0,31	0,29	0,39
xxxx-14	0,51	0,46	0,27	0,26	0,33
xxxx-15	0,50	0,45	0,30	0,28	0,41

После сдачи результатов мы испытали на коллекции KM.RU поиск по кворуму с соблюдением описанных выше условий. Видно, что характер графика не изменился, но сам график плавно поднялся вверх. В целом график TREC оказался очень чувствителен к ситуациям, когда система не находит документов. Применив кворум к шести запросам из шестидесяти, заданных к коллекции KM.RU, мы получили существенно лучший график. В ситуациях, когда кворум не требуется, модификация системы, представленная на РОМИП, отрабатывает хорошо.

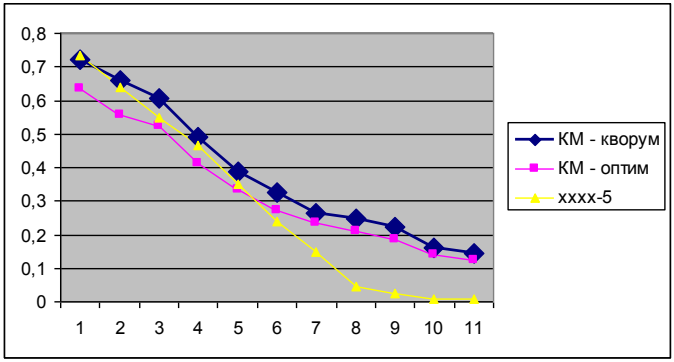

График TREC с механизмом кворума для коллекции KM.RU (делался вне РОМИП после сдачи основных результатов)

Коллекция Legal

У нас получились следующие результаты:

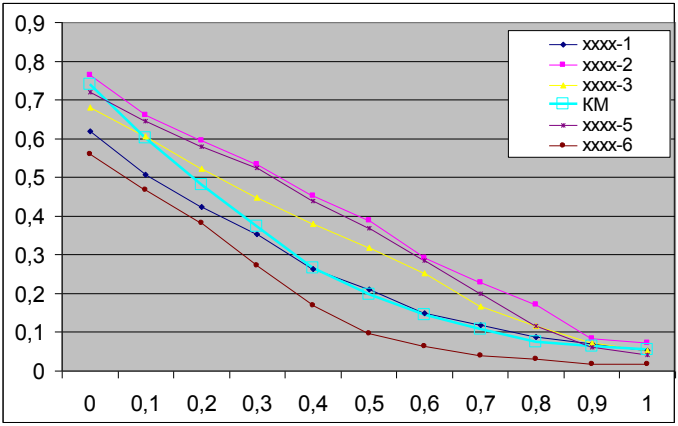

График TREC, оценка OR для коллекции Legal

Приведем результаты по основным показателям:

№	Pr(5)	Pr(10)	Bpref-10	Bpref	Recall
xxxx-1	0,4	0,36	0,27	0,29	0,44
xxxx-2	0,53	0,48	0,37	0,42	0,63
xxxx-3	0,5	0,47	0,33	0,37	0,55
KM	**0,5**	**0,42**	**0,28**	**0,33**	**0,51**
xxxx-5	0,52	0,5	0,36	0,41	0,62
xxxx-6	0,38	0,34	0,2	0,24	0,34

Мы сделали один экспериментальный прогон. Нам было интересно, как поисковый механизм, оптимизированный для веба, справится со специфической коллекцией. Результат показал, что наше направление – лучший результат на максимально высокой позиции – выдерживается. 1-я точка на графике TREC находится достаточно высоко. Показатель precision(5) находится на хорошем уровне по сравнению с другими участниками. Далее мы идем вслед за основной группой. Поскольку поиск по специфичным

коллекциям не является нашим приоритетным направлением, мы в целом довольны результатом.

Тем не менее, мы решили выяснить причину падения графика и возможность улучшения качества поиска по произвольной коллекции (в данном случае, нормативных документов). Несмотря на то, что в коллекции присутствует большое число ссылок, решили выяснить, как поиск отработает без ссылочного ранжирования. Тесты показали, что в этом случае результаты получаются лучше. Возможно, это означает, что к подобной коллекции неприменим традиционный подход, использующий ссылочный граф. Дополнительно мы выяснили, что качество ранжирования существенно зависит от учета пассажей в заголовке документа. Дальнейшее повышение качества поиска может быть связано с использованием кворума и словаря специфических для коллекции сокращений.

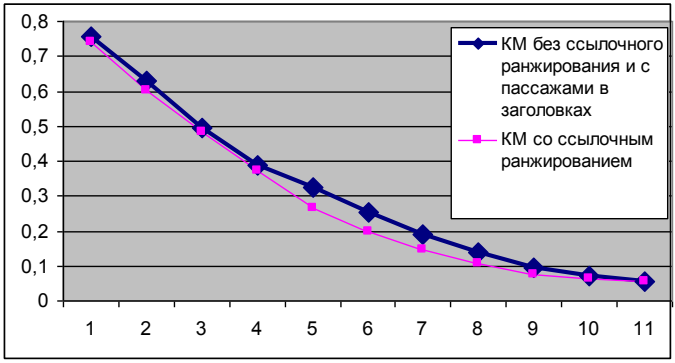

График TREC без ссылочного ранжирования и с пассажами в заголовках для коллекции Legal (делался вне РОМИП после сдачи основных результатов)

Выводы

Эксперименты на РОМИП показали, что оптимизация коэффициентов в формуле релевантности дает прирост качества поиска. Новые параметры - пары слов и близость слов к началу предложений – тоже зарекомендовали себя хорошо. Это особенно заметно в поиске по веб-коллекции.

VII. Возможные пути дальнейшего развития

Основные методы улучшения качества выдачи поисковых машин в Интернете до сих пор были связаны с анализом документов в веб-коллекции, а анализу запросов пользователей не уделялось должного внимания. Однако сейчас все больше исследователей обращается к этой теме (поскольку нам интересен информационный поиск в сети Интернет, мы говорим об исследовании запросов и текстов документов в рамках веба).

К анализу запроса можно отнести анализ его синтаксического устройства, морфологических особенностей слов в запросе, а также их семантического значения.

Синтаксический анализ запроса отчасти представляется труднопроизводимым, поскольку формулировка многих веб-запросов не соответствует строению фразы в естественном языке (пример: *Москва принтеры заказать*). Синтаксический анализ запроса может быть интересен для создания вопросно-ответного поиска (на вопрос: *В каком году родился Пушкин* – ищем ответ: *Пушкин родился в X году*, где *X* – составное числительное или четырехзначное число). Однако организация такого вида поиска с собственной базой проиндексированных документов требует намного больше усилий по сравнению с «традиционной» поисковой машиной, поскольку вопросно-ответный поиск предполагает написание большого количества правил, отработка которых будет замедлять процесс обработки запроса, и поискового индекса из размеченных веб-документов, который будет значительно превышать по объему индекс традиционной поисковой машины, что может позволить себе не каждая команда разработчиков. Поэтому сейчас на рынке нет вопросно-ответной системы, способной наравне конкурировать с промышленными системами информационного поиска в Интернете.

Учет морфологических особенностей слов в запросе потребует анализа морфологических особенностей слов и в веб-документах, что, очевидно, значительно увеличит объем поискового индекса. Процесс, скорее всего, будет тесно связан с синтаксическим анализом слов в запросе (например, использование информации о падежах для определения главного слова в словосочетаниях «браслет из золота», «карта города»).

Наиболее привлекательной пока остается перспектива проведения семантического анализа. Информацию о семантических значениях можно использовать для определения тематики запроса, подбора синонимов для слов из запроса, создания базы связанных

понятий и для переформулировки запроса другими словами. Уже сейчас поисковые системы умеют расшифровывать сокращения (*МГУ = Московский государственный университет*) и сопоставлять русский и английский вариант написания слов (*БМВ = BMW*).

Для нас представляется интересным построение базы связанных понятий для слов русского языка, и мы начали экспериментировать в этом направлении. Эксперименты показали, что построение связей между словами для выделенной тематической области - это вполне решаемая задачей, однако использование выбранных нами методов вряд ли позволит создать сеть связанных понятий для всех тематических областей сразу. Поэтому на ближайшее будущее мы ставим перед собой две задачи:

1) научиться устанавливать связи между словами для произвольной выборки текстов и за приемлемое время;

2) научиться корректно использовать базу связанных понятий для переформулировки веб-запроса.

Тогда как при простых переформулировках запроса (расшифровка сокращений, замена русского написания на английское) можно максимально оградить себя от ошибок, то при более сложных вариантах переформулировки приходится сталкиваться с рядом трудностей. До сих пор основным этапом обработки запроса Интернет-пользователя является поиск слов из запроса в веб-документах, и при замене слов из запроса на синонимы, которые представляются нам очевидными, поиск будет производиться уже по другим словам, что может ухудшить качество выдачи. Например, в настоящее время связь понятий *Медведев* и *президент России* для нас безусловна, и если на запрос *Медведев* мы выдадим сайт президента России, то этот документ пользователь, скорее всего, сочтет хорошим как один из вариантов ответа, однако мы не можем выдачу для запроса *Медведев* полностью подменить выдачей для запроса *президент России*, т.к. в этом случае пользователь, возможно, не получит документы с биографией Медведева, с его интервью и проч., и это еще не самое плохое, хуже, если при такой подмене выдачи мы отдадим документы про президента Путина или президента Ельцина, тогда как пользователь, может быть, хотел узнать личные факты биографии президента Медведева, и его совсем не интересовал политический аспект.

Тем не менее, несмотря на множество трудностей, связанных с проведением семантического анализа запроса, этот путь развития представляется достаточно перспективным, и мы планируем

проводить эксперименты в этом направлении. Если такие способы оценки релевантности документа, как расчет tf*idf, веса документа, расстояния между ключевыми словами, можно считать количественными показателями, то анализ семантического значения запроса – это оценка уже другого плана. Несомненно, мы не хотим отказываться от учета количественных показателей при обработке запроса, но надеемся, что использование новой информации поможет нам сделать качественный шаг вперед.

Литература

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

[2]. «KM.RU на РОМИП-2007», С.Татевосян, Н.Брызгалова - http://romip.ru/romip2007/romip2007_KM.RU.pdf

[3]. «Алгоритм текстового ранжирования Яндекса на РОМИП-2006» Андрей Гулин, Михаил Маслов, Илья Сегалович - http://download.yandex.ru/company/03_yandex.pdf

[4]. «Извлечение значимой информации из web-страниц для задач информационного поиска» М.С. Агеев, И.В. Вершинников, Б.В. Добров - http://www.cir.ru/docs/ips/publications/2005_yandex_obraml.pdf

[5]. «Оптимизация параметров алгоритма поиска на основе анализа оценок экспертов», М.С. Агеев, Б.В. Добров - http://romip.ru/romip2005/07_uirussia.pdf

KM.RU at RIRES-2008. Parameter optimization

S. Tatevosyan, N. Bryzgalova

The paper describes a new modification of the information retrieval algorithm, introduced by KM.RU at RIRES-2007, we also talk about an optimizing system aimed at getting the most efficient coefficients for algorithm's parameters for obtaining better results in information retrieval and document ranking. The article reports on results of KM.RU at RIRES-2008 and our future plans.