МЕТОДЫ РАНЖИРОВАНИЯ В ПОЛНОТЕКСТОВОМ ПОИСКЕ ПО КОЛЛЕКЦИИ HTML-ДОКУМЕНТОВ

FULLTEXT SEARCH RANKING METHODS IN HTML DOCUMENT COLLECTION

А.Н. Федоровский *(fedorovsky@corp.mail.ru)*

М.Ю. Костин *(kostin@corp.mail.ru)*

Mail.Ru, Москва

В статье описаны алгоритмы страничного ранжирования, применяемые при расчете релевантности в системе Поиск@Mail.Ru. Приведены результаты экспериментальной проверки их эффективности. Рассмотрены также вопросы применимости данных критериев ранжирования при построении полномасштабных поисковых систем.

1. Введение

Поисковая система, используемая для проведения экспериментов по дорожкам, была разработана в компании в рамках проекта Поиск@Mail.Ru (http://go.mail.ru/) и успешно используется в настоящее время на веб-проектах компании и наших партнеров.

При разработке системы одной из главных целей было сделать ее гибкой и легко адаптируемой при помощи удобного набора настроек к самым разным поисковым задачам, значительно отличающимся друг от друга как по характеру индексируемой коллекции, так и по типичным поисковым потребностям пользователей. С другой стороны, необходимо было обеспечить высокое качество поиска по комплексной коллекции, содержащей различные типы документов, типичным примером которой является веб-коллекция по достаточно большому набору сайтов. Кроме того, важным требованием была высокая производительность, позволяющая использовать систему для поиска по большим объемам данных под высокой пользовательской нагрузкой.

Структура поискового индекса близка к классической, многократно описанной в литературе [3, 4]. Основой являются инверсные списки вхождений слов, используемые для поиска релевантных документов и прямые индексы термов для формирования фрагментов, возвращаемых пользователю (сниппетов). Применяется также ряд техник, позволяющих уменьшить требования к памяти и увеличивающих скорость работы как во время индексации, так и при обслуживании пользовательских запросов.

Для обеспечения высокого качества возвращаемых результатов в первую очередь необходим правильный выбор функции релевантности, определяющей меру соответствия документа запросу. Остановимся на этом более подробно.

2. Особенности страничного ранжирования в Поиск@Mail.Ru

При расчете релевантности нами учитывается как частота вхождения в документ единичных слов запроса, так и совместная встречаемость слов и их взаимное положение. В отличие от большинства известных систем, мы используем два различных способа учета взаимного положения слов в документе: совместное вхождение пар слов и нахождение в документе релевантных пассажей. Каждый из этих методов имеет свои достоинства: метод пар слов позволяет качественно обрабатывать не только запросы, являющиеся единым словосочетанием, но и запросы с разной связанностью групп слов внутри запроса. В то же время, наиболее близкий к запросу пассаж лучше других методов позволяет оценить наличие формального соответствия между запросом и документом, то есть наличие в документе хотя бы простого упоминания объекта, заданного в запросе. При совместном использовании эти два метода, на наш взгляд, хорошо дополняют друг друга и позволяют добиться хорошего качества поиска для максимально широкого круга запросов.

Таким образом, вес документа по запросу в нашей системе складывается из трех составляющих:

$$W = k_f W_f + k_p W_p + k_{ps} W_{ps} \quad (1)$$

где:

W_f – вес документа, вычисленный на основе TF*IDF алгоритма;

W_p – вес документа, вычисленный на основе совместных вхождений в документ пар слов, расположенных рядом в запросе;

W_{ps} – вес наиболее близкого к запросу пассажа документа;

k_f, k_p, k_{ps} – коэффициенты.

Следует также отметить, что для получения ненулевого веса в документе не обязательно должны присутствовать все слова запроса. В число ранжируемых попадают также документы, для которых отношение суммарного IDF слов запроса, встречающихся в них, к суммарному IDF всех слов запроса превышает заданный порог. Такие документы дополнительно «штрафуются» за отсутствующие слова, однако, тем не менее, вес некоторых из них в общем случае может даже превышать вес документов, содержащих все слова запроса.

Рассмотрим подробнее каждый из весов в формуле (1).

*2.1. TF*IDF вес*

Формула, используемая нами для подсчета TF*IDF веса по каждому терму запроса, является модификацией стандартной BM25 формулы [2], и выглядит следующим образом

$$TF * IDF_{term} = \frac{f_{term} \times IDF_{term}}{f_{term} + k_1(b + L(1-b))} \quad (2)$$

где:

f_{term} – вес терма в документе, вычисленный на основе количества вхождений, с учетом ряда дополнительных факторов;

IDF_{term} – обратная частотность терма в коллекции, вычисленная по стандартной логарифмической формуле;

L – нормированная длина документа;

k_1, b – коэффициенты.

Общий TF*IDF вес документа получается суммированием полученных весов по всем термам запроса.

Особенностью применения этой формулы в нашем поиске является то, что для документов, размер которых превышает константу k_2 (соответствующую в стандартной BM25 формуле средней длине документа в коллекции, а в нашей системе задаваемой в настройках) вместо нормирования по длине используется метод разбиения документа на перекрывающиеся фрагменты [1]. Применение этого метода позволяет избежать неоправданного занижения веса длинных документов, в которых имеется небольшой фрагмент с высокой релевантностью.

Фрагменты имеют фиксированный, задаваемый в настройках размер, меньший k_2, и берутся с наложением по всему тексту документа.

Вес каждого из фрагментов по каждому терму запроса оценивается по формуле (2) без нормирования по длине, то есть с $L = 1$.

В результате, выбирается фрагмент документа, имеющий наибольший вес и его вес используется в качестве W_f веса документа в (1).

В качестве особенности, не встречающейся в известных нам работах на эту тему, можно отметить, что в каждый фрагмент нами дополнительно включается небольшой отрезок текста в начале документа, существенно меньший, чем длина фрагмента, так как слова, находящиеся в самом начале длинного документа, часто описывают его содержание в целом.

Для документов, имеющих длину меньшую, чем k_2, используется нормирование по длине по формуле

$$L = \frac{L_w + k_4}{k_3 + k_4} \quad (3)$$

где:

L_w – длина документа в словах;

k_3, k_4 – коэффициенты, задаваемые в настройках.

Еще одной существенной особенностью TF*IDF ранжирования в нашем поиске является использование достаточно большого значения коэффициента k_1 в формуле (1): для прогонов использовалось значение, значительно б**о**льшее обычно принятых. Наш выбор здесь связан с тем, что небольшое значение этого коэффициента призвано дать преимущество документам с достаточно одинаковой встречаемостью в документе различных слов запроса, что актуально в случае, когда TF*IDF является единственным критерием ранжирования и совместная встречаемость слов никак иначе не учитывается. Поскольку мы учитываем совместную встречаемость слов отдельно, то здесь мы выбрали значение коэффициента, позволяющее дать достаточно высокий вес документам с высокой встречаемостью лишь некоторых (и даже, в частности, одного) слов запроса.

2.2. Вес по парам слов

При подсчете этого веса вхождение терма в документ учитывается только в том случае, если оно находится в документе на расстоянии, не превышающим заданное от хотя бы одного из стоящих рядом с ним термов запроса (особая обработка предусмотрена для стоп-слов).

Для прогона по веб-коллекции расстояние было нами выбрано как 2 (рядом или через одно) для случая, когда порядок слов в запросе и документе совпадает и 1 (только рядом) для случая, когда не совпадает.

Соответствующие этому условию вхождения слов обрабатываются по описанному выше TF*IDF алгоритму, отличается только набор коэффициентов.

2.3. Вес лучшего пассажа

Под пассажем мы понимаем фрагмент документа, размера, не превышающего заданный, в котором встречаются все термы запроса, либо значительная часть термов запроса, суммарный IDF которых превышает заданное ограничение.

При выборе лучшего пассажа документа основными факторами являются его полнота (наличие всех термов запроса), длина, порядок слов (его совпадение с порядком слов в документе), зона документа (заголовок, выделенный текст, обычный текст), в которой встретился пассаж, близость пассажа к началу документа. Учитывается также ряд дополнительных факторов.

Вес пассажа по каждому из факторов оценивается в баллах на основе специальных для каждого из них правил, после чего веса суммируются. Суммарный вес и будет весом пассажа. Из полученных весов пассажей выбирается максимальный для вычисления общего веса по формуле (1).

3. Описание экспериментов

Эффективность описываемых алгоритмов была проверена в процессе участия системы в российском семинаре методов оценки информационного поиска (РОМИП) в 2005 году [6].

На семинаре было создано 3 дорожки для сравнения алгоритмов работы поисковых систем:

- поиск по веб-коллекции (web-adhoc)
- поиск по нормативным документам (legal-adhoc)
- поиск по смешанной коллекции (mixed-adhoc)

Было заявлено по 1 экспериментальному прогону для каждой из этих дорожек. Гибкость настроек системы оказалось достаточной, чтобы по каждой из дорожек однозначно выбрать параметры, желаемые для построения гармоничной системы, настроенной на данную коллекцию и, как следствие, ограничиться одним прогоном. Конечно, глобальный оптимум мог быть и не достигнут, однако результаты в каждом случае оказались вполне удовлетворительными.

3.1. Web-adhoc

Для дорожки поиска по веб-коллекции (web-adhoc) участникам была предложена достаточно обширная коллекция веб-страниц, представляющая собой часть сайтов домена narod.ru (более 700000 страниц, 6.3ГБ). По условиям дорожки по этой коллекции необходимо было выполнить большое количество (более 24000) запросов, специально отобранных из поисковых логов. Первые 100 документов из поисковой выдачи считались ответом системы на запрос, упорядоченный по мере убывания значимости документов в выдаче. Для итоговой оценки после получения результатов из всего множества запросов были отобраны 75. Достаточно подробное описание способов производимых оценок и стандартных параметров приведено, например, в описании семинара РОМИП [5].

Ниже приведена таблица результатов участников дорожки для способа оценки web-adhoc-or-pd50-all (хотя бы одна оценка экспертов превышает минимальный порог релевантности; рассматриваются ответы систем, суженные до глубины пула – 50 документов; по запросам 2004 и 2005 годов) и 11-точечный график TREC по тому же способу оценки.

Прогон / участник	1	2	**Mail.Ru**	4	5	6
Recall	0,4023	0,2471	**0,5443**	0,4548	0,4265	0,4627
Precision(5)	0,3707	0,2000	**0,5840**	0,4853	0,5013	0,5013
Average precision	0,2027	0,0933	**0,3178**	0,2488	0,2401	0,2585
Precision(10)	0,3507	0,2133	**0,5147**	0,4467	0,4373	0,4560
R-precision	0,2733	0,1509	**0,3521**	0,3074	0,2925	0,3138
Precision	0,2701	0,1797	**0,3568**	0,3045	0,2999	0,3231

Таблица 1. Сравнительные результаты участников дорожки веб-поиска с оценкой web-adhoc-or-pd50-all.

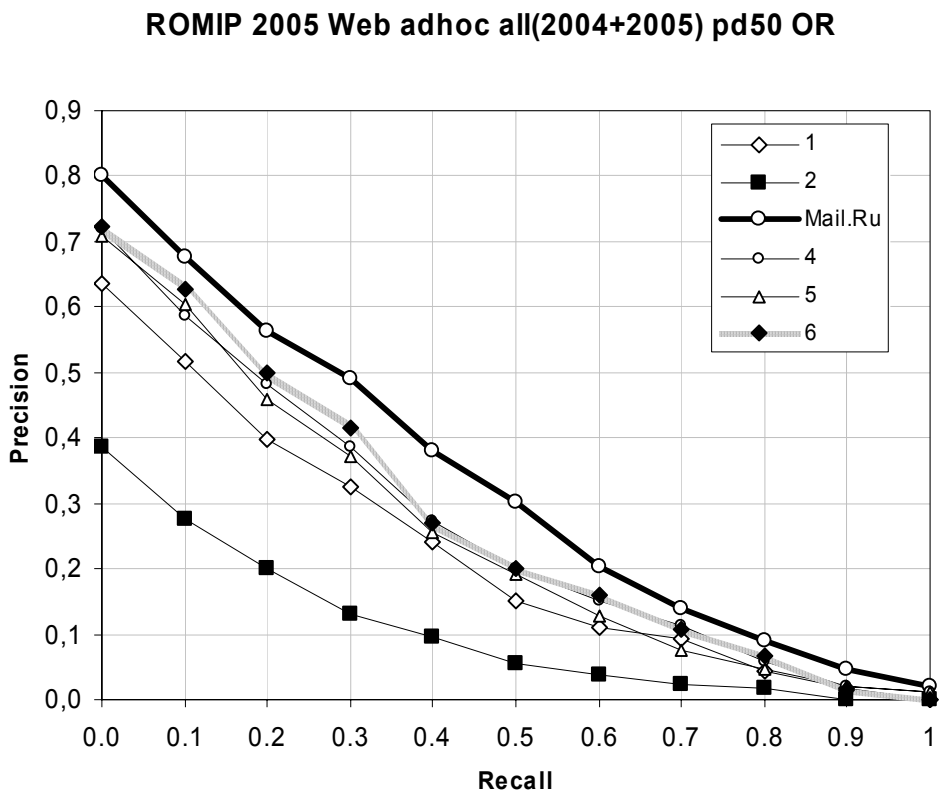

Рисунок 1. 11-точечные графики TREC для участников дорожки веб-поиска с оценкой web-adhoc-or-pd50-all.

Примерно такое же соотношение как параметра Average Precision, так и точек графика TREC между нашими результатами и другими участниками сохраняется и для остальных способов оценки.

В таблице 2 приведены значения параметров, достигнутые системой Поиск@Mail.Ru по различным видам оценок. В двух первых колонках – показатели рассчитываются для всех документов из поисковой выдачи, в последних двух – только для первых 50 (pd50). Соответственно, в первой и третьей – сильные требования к релевантности (все оценки экспертов превышают минимальный порог релевантности), во второй и четвертой – слабые требования к релевантности (хотя бы одна из оценок экспертов превышает минимальный порог релевантности).

	вся выдача		pd50	
	AND	OR	AND	OR
Recall	0,7634	0,6847	0,6309	0,5443
Precision(5)	0,4149	0,5840	0,4149	0,5840
Average precision	0,3330	0,3704	0,3061	0,3178
Precision(10)	0,3179	0,5147	0,3179	0,5147
R-precision	0,3331	0,3826	0,3207	0,3521
Precision	0,1235	0,2460	0,1824	0,3568

Таблица 2. Результаты прогона Поиск@Mail.Ru для оценок вида web-adhoc-*-*-all.

3.2. Legal-adhoc, Mixed-adhoc

Для дорожки поиска по нормативным документам (Legal-adhoc) участникам компанией «Кодекс» была предоставлена коллекция размером около 67000 документов. Полностью аналогично дорожке веб-поиска, задача состояла в выполнении большого количества (12900) запросов на этой коллекции. Ответом системы также считалось до 100 наиболее релевантных документов в выдаче.

Для дорожки смешанного поиска необходимо было на объединенной коллекции веб- и нормативных документов выполнить объединенное множество запросов, предоставленных для дорожек Web-adhoc и Legal-adhoc. Целью являлась проверка возможности систем функционировать в сильно разнородной коллекции, так как веб- и нормативные документы значительно отличаются по своим характеристикам.

Для выполнения заданий этих двух дорожек использовалась та же поисковая система, отличие состояло только в настройке параметров для конкретной задачи.

Так, например, в нормативных документах заголовок, скорее всего, будет содержать слова, напрямую относящиеся к тематике документа. В то же время в веб-коллекции мусор в заголовках страниц – вполне обычное дело. Соответственно, был изменен вес для слов и пассажей, входящих в важные зоны документа.

Также в правовых документах чаще встречаются сложные синтаксические структуры, в результате чего связанные по смыслу слова оказываются разделенными б**о**льшим количеством сторонних слов. Для учета этого были изменены максимально возможная длина пассажа и ряд бонусов и штрафов, начисляемых за связность слов запроса в документе. И т. п.

Результаты этих дорожек, по большому счету, похожи на результаты дорожки web-adhoc. Подробно с ними можно ознакомиться в работе [6].

4. Применимость описанных алгоритмов для поисковых систем

Попробуем проанализировать актуальность описанных выше алгоритмов и полученных результатов к задаче поиска по Интернету, решаемой большими поисковыми системами.

Приведем основные отличия между задачей поиска по Интернету и задачей полнотекстового поиска в ее классической постановке, использованной (с некоторыми модификациями) на семинаре РОМИП:

- высокие требования к производительности системы
- большой объем данных
- возможность применения внестраничных критериев ранжирования
- проблема поискового спама
- отсутствие формализованных критериев релевантности

Этот список, безусловно, неполон, мы перечислили только наиболее важные различия.

Посмотрим теперь, как каждое из этих отличий повлияет на ценность использованных нами алгоритмов и репрезентативность полученных экспериментальных результатов.

4.1. Производительность

Необходимость обеспечить высокую производительность системы при большом объеме проиндексированной информации часто приводит к невозможности применить в поиске по Интернету алгоритмы, хорошо зарекомендовавшие себя в экспериментальных исследованиях. Однако в нашем случае высокие требования по производительности предъявлялись к системе изначально и были соблюдены при ее разработке. Как и в других поисковых системах, обеспечивается это как непосредственно оптимизацией описанных алгоритмов, так и оптимизацией, учитывающей ранжирование, то есть позволяющей выполнять ресурсоемкие операции точного ранжирования только для документов, имеющих шанс оказаться на достаточно высоких местах в выдаче по результатам более грубой оценки их релевантности..

4.2. Объем данных

Основные следствия огромного различия в объеме данных между Интернетом (даже только российским) и тестовыми коллекциями (в частности коллекцией, использованной в РОМИП) это:

- увеличение доли запросов, по которым количество релевантных документов велико
- увеличение доли многословных, подробно сформулированных запросов

Первая из этих особенностей в основном компенсируется применением на этих запросов внестраничных факторов ранжирования. Влияние второй проявляется прежде всего в повышении важности алгоритмов учета взаимного расположения слов. Поскольку именно им мы уделили особое внимание, то и рассчитывали изначально на то, что значительное увеличение объема данных не скажется на качестве работы нашей системы отрицательно. В то же время, до экспериментальной проверки об этом обычно нельзя говорить с полной уверенностью.

4.3. Внестраничные факторы

Ссылочное ранжирование, индекс цитирования, описания сайтов в каталогах, релевантность запросу сайта в целом и другие внестраничные критерии имеют решающее значение для релевантности по значительной доле запросов в поиске по Интернету. Релевантность текста страницы для таких запросов также имеет значение, однако при этом бывает достаточно ее грубой оценки, тонкие различия практически не влияют на релевантность результатов по подобным запросам. В то же время, не менее значительна и доля запросов, для которых внестраничная информация практически отсутствует и решающим оказывается страничное ранжирование. Таким образом, можно говорить, что хотя релевантность результатов в поиске по Интернету определяется не только качеством алгоритмов страничного ранжирования, их влияние на качество поиска достаточно велико.

Конечно, здесь надо иметь в виду, что релевантность результатов поиска в Интернете зависит не только от качества ранжирования, но и от других факторов. Объем и частота обновления базы, отслеживание нечетких дубликатов, фильтрация спама - все это также оказывает значительное влияние на качество поиска.

4.4. Спам

Страничные критерии релевантности сравнительно легко подделать, ведь страница полностью находится во власти ее автора. Однако, все страничные факторы подвержены спаму примерно в равной степени из-за равной же степени контроля автором текстовой информации, на основе которой эти критерии строятся. Следовательно, есть смысл не задаваться вопросом о мере этой подверженности, а вынести основную тяжесть борьбы со спамом на специализированные антиспамовые алгоритмы, которые не имеют прямого отношения к ранжированию.

4.5. Отсутствие формализованных критериев релевантности

Значимой мерой релевантности в реальных поисковых системах является степень удовлетворенности пользователя полученными результатами. Естественно, этот критерий не поддается точному формальному определению, в отличие от критериев, используемых в экспериментах по информационному поиску. Вопрос о степени применимости традиционных формальных критериев к реальному поиску в Интернете остается малоисследованным. Например, такие значимые в экспериментальных исследованиях критерии как Precision, Recall, Average Precision ориентированы на ситуацию, когда пользователя интересуют все релевантные документы, и он просматривает всю поисковую выдачу. В реальном же поиске по Интернету подобная модель поведения пользователя является всего лишь одной из многих и встречается не столь уж часто. Возможно, в будущем будут разработаны системы оценки релевантности, учитывающие вероятную модель поведения пользователя для оцениваемого запроса и выбирающие адекватный критерий ранжирования, хотя и это будет лишь частичным решением проблемы.

С другой стороны, наличие корреляции между формальными критериями и качеством поиска с точки зрения пользователя несомненно. Кроме того, как показали, в частности, наши исследования, различные формальные критерии релевантности, как правило, достаточно хорошо коррелируют друг с другом, что тоже является аргументом в пользу их применимости к поиску по Интернету, несмотря на многообразие моделей поведения пользователя.

Таким образом, в целом можно говорить о том, что описанные алгоритмы применимы, в том числе, для построения полноценной системы поиска по Интернету. В то же время, многие вопросы не могут считаться ясными до их экспериментальной проверки.

Список литературы

1. J. P. Callan. Passage-level evidence in document retrieval. // In The 17 Conference on Research and Development in Information Retrieval, pages 302-309, Dublin, Ireland, 1994. ACM
2. S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. // In TREC-3, 1994.
3. G. Salton, C. Buckley. Term-weighting approaches in automatic text retrieval. // In Information Processing and Management: an International Journal, Volume 24, Issue 5, pages: 513 – 523, 1988.
4. S. Brin, L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. // In Computer Networks and ISDN Systems, Volume 30, number 1-7, pages 107-117, 1998.
5. Под ред. И.С. Некрестьянова. Труды РОМИП'2004 // Санкт-Петербург: НИИ Химии СПбГУ, 2004.
6. Федоровский А.Н, Костин М.Ю. Mail.ru на РОМИП-2005. // в сб. "Труды РОМИП'2005" Труды третьего российского семинара по оценке методов информационного поиска. Под ред. И.С. Некрестьянова, стр. 106-124, Санкт-Петербург: НИИ Химии СПбГУ, 2005.