HeadHunter на РОМИП-2008

© А.В. Сафронов

HeadHunter
safronov@hh.ru

Аннотация

Статья посвящена участию компании HeadHunter в дорожках поиска по документам на семинаре РОМИП-2008. Подробно описывается алгоритм ранжирования, приводятся полученные результаты.

1. Введение

HeadHunter участвовал в РОМИП впервые. Цель нашего участия заключалась в получении независимой оценки качества разработанной в компании экспериментальной поисковой системы.

Поиск - одна из типичных функций работных сайтов, к числу которых относится hh.ru. В большинстве случаев при поиске резюме и вакансий пользователь заинтересован в получении результатов, отсортированных по дате. Однако иногда возникает необходимость в ранжировании результатов по степени соответствия поисковому запросу. Режим сортировки по релевантности уже достаточно давно имеется на сайте hh.ru. В процессе изучения необходимости дальнейшего развития этого режима нами был разработан экспериментальный алгоритм ранжирования документов. При помощи таблиц релевантности, сформированных на предыдущих семинарах РОМИП, была произведена настройка параметров алгоритма. С данным алгоритмом мы приняли участие в дорожках текстового поиска РОМИП-2008.

Далее описывается сам алгоритм ранжирования, а также рассматриваются результаты, полученные при выполнении заданий дорожек поиска по документам.

2. Описание алгоритма ранжирования

2.1 Общая формула

Для ранжирования документов используется формула, учитывающая несколько различных факторов:

$$W = W_{doc} + W_{title} + W_{begin} + W_{ps1} + W_{ps2} + W_{ps3} + W_{str} \quad (1)$$

где:

W_{doc} - вес всего документа;

W_{title} - вес заголовка;

W_{begin} - вес начальной части документа;

W_{ps1} - вес лучшего «длинного» пассажа;

W_{ps2} - вес лучшего «среднего» пассажа;

W_{ps3} - вес лучшего «короткого» пассажа;

W_{str} - вес лучшей цепочки слов.

Несмотря на то, что коллекции РОМИП позволяют использовать ссылочное ранжирование для улучшения результатов (как это следует из [3]), мы не стали включать в наш алгоритм учет гиперссылок.

Далее рассмотрим более подробно каждое из слагаемых формулы (1).

2.2 Вес всего документа

Слагаемое W_{doc} оценивает вес всего документа целиком:

$$W_{doc} = QFTFIDF(d,q) = QF(d,q) * \sum_{t \in q} TF(d,t) * IDF(t) \quad (2)$$

где:

d - оцениваемый документ ($d = title \cup body$);

t - слово из поискового запроса;

q - множество слов, входящих в поисковый запрос;

QF - функция, предназначенная для оценки доли слов запроса, встречающихся в документе. Она позволяет повышать вес тех документов, которые включают в себя все слова запроса. Функция представляет собой отношение суммы IDF слов запроса, встречающихся в документе, к сумме IDF всех слов запроса.

$$QF(d,q) = \frac{\sum_{i \in q \cap d} IDF(i)}{\sum_{j \in q} IDF(j)}$$

Для расчета TF и IDF используется модель INQUERY, описанная в [1]:

$$TF(d,t) = \frac{freq(d,t)}{freq(d,t) + k_1 + \frac{|d|}{k_2}}$$

$$IDF(t) = \frac{\log\left(\frac{|c| + 0.5}{df(t)}\right)}{\log(|c| + 1)}$$

где:

$freq(d,t)$ - количество вхождений слова t в документ d;

$|d|$ - длина документа d в словах;

k_1 = 1;

k_2 = 16384;

$|c|$ - количество документов в коллекции c;

$df(t)$ - количество документов, в которых встречается слово t.

Следует заметить, что W_{doc} оценивает именно весь текст документа, включая и его заголовок, несмотря на то, что в общей формуле (1) присутствует отдельная оценка веса заголовка. Такая общая оценка нужна, чтобы правильно оценить документы, в которых часть слов запроса присутствует только в заголовке, а часть – только в теле. В работе [6] подробно рассматривается эта проблема и в качестве решения предлагается включать слова заголовка в документ с удвоенной частотой. Однако наши эксперименты показали, что на коллекциях РОМИП стратегия «общий вес документа + вес заголовка» является более эффективной.

2.3 Вес заголовка

Для увеличения веса документов, содержащих слова поискового запроса в своем заголовке, используется слагаемое W_{title}. Его расчет производится по формуле, похожей на (2). Отличие состоит в том, что учитываются только вхождения слов в заголовок документа.

$$W_{title} = k_{title} * QFTFIDF(title, q)$$

где:

k_{title} - коэффициент, задающий «важность» веса заголовка в общей формуле (1);

$title$ - заголовок документа.

2.4 Вес начальной части документа

Иногда в начале документов располагают краткое описание его содержания (аннотацию). Для увеличения веса документов, содержащих слова поискового запроса в самом начале, используется слагаемое W_{begin} . Его расчет производится по формуле, похожей на (2). Отличие состоит в том, что учитываются только N первых слов из тела документа.

$$W_{begin} = k_{begin} * QFTFIDF(begin(body, N), q)$$

где:

k_{begin} - коэффициент, задающий «важность» слагаемого W_{begin} в общей формуле (1);

$begin(body, N)$ – первые N слов тела документа $body$.

2.5 Вес лучших пассажей

Наилучшие пассажи выбираются среди всех фрагментов документа, полученных с помощью скользящего окна определенной длины. В отличие от распространенной схемы, описанной в [5], в качестве шага при перемещении окна по документу используется не половина длины окна, а 1 слово. Теоретически, это может привести к более аккуратному выбору лучших пассажей для случаев, когда размер высокорелевантных фрагментов текста превышает половину длины окна. Однако практической проверки этой гипотезы мы не проводили.

В состав формулы (1) входят веса 3х пассажей, которые различаются длиной: «длинный», «средний» и «короткий» пассажи. Введение 3х пассажей вместо традиционного использования одного пассажа по нашей гипотезе должно было привести к решению проблемы определения оптимальной длины пассажа для коллекций разнородных документов. Наши эксперименты показали, что

наиболее эффективными являются длины пассажей, соотносящиеся между собой как 1:4:16.

Кроме длины, отличие между пассажами состоит в том, что к «длинному» пассажу приписывается заголовок и несколько первых слов из тела документа (подобно тому, как это описано в [4]); к «среднему» пассажу добавляется заголовок; в «коротком» пассаже используются только слова из самого пассажа.

$$W_{ps1} = k_{ps1} * \max_{0<p<|body|-L1} QFTFIDF(window(body, p, L_1) \cup title \cup begin(body, N), q)$$

$$W_{ps2} = k_{ps2} * \max_{0<p<|body|-L2} QFTFIDF(window(body, p, L_2) \cup title, q)$$

$$W_{ps3} = k_{ps3} * \max_{0<p<|body|-L3} QFTFIDF(window(body, p, L_3), q)$$

где:

L_1 = 256;

L_2 = 64;

L_3 = 16;

$|body|$ - длина тела документа в словах;

$window(body, p, L)$ – фрагмент тела документа $body$, имеющий длину L слов и начинающийся с позиции p.

2.6 Вес лучшей цепочки слов

Под цепочкой слов подразумевается группа слов из поискового запроса, стоящих в документе рядом друг с другом. Порядок следования слов не важен. Вес цепочки определяется как сумма IDF слов, входящих в группу, к сумме IDF всех слов поискового запроса.

$$W_{str} = k_{str} * \max_{str} \frac{\sum_{i \in str} IDF(i)}{\sum_{j \in q} IDF(j)}$$

где:

k_{str} - коэффициент, задающий «важность» слагаемого W_{str} в общей формуле (1);

str – цепочка слов.

3. Результаты

Мы участвовали в 2х дорожках семинара: дорожке поиска по коллекции нормативно-правовых документов и дорожке поиска по коллекции KM.RU. В каждой коллекции использовался только один прогон нашей системы.

Описание метрик, с помощью которых осуществлялась оценка результатов, можно найти в [2].

Оценка результатов дорожки нормативной коллекции производилась экспертами с юридическим образованием. Для каждой пары документ-запрос была собрана одна оценка. Всего было оценено 95 запросов, из них 25 запросов с прошлогоднего семинара. Глубина пула составила 35 документов.

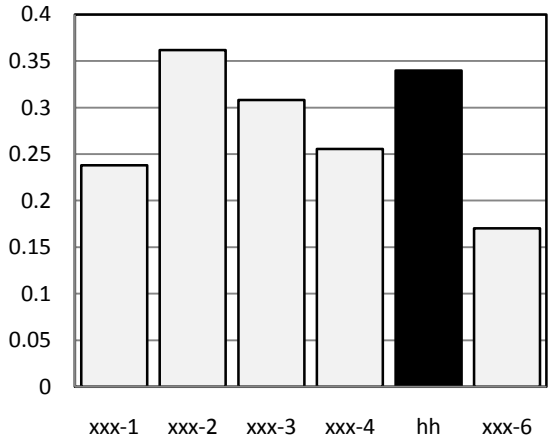

Рисунок 1. Коллекция Legal2007, метрика Average precision

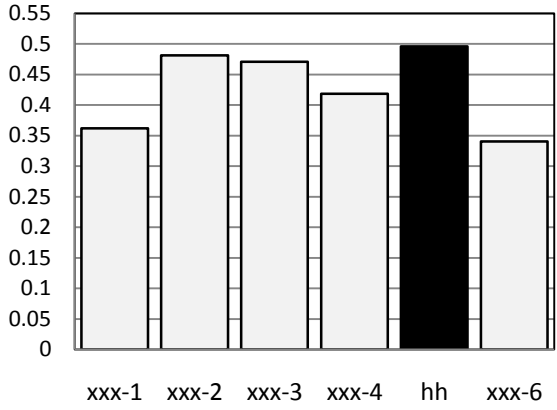

Рисунок 2. Коллекция Legal2007, метрика Precision(10)

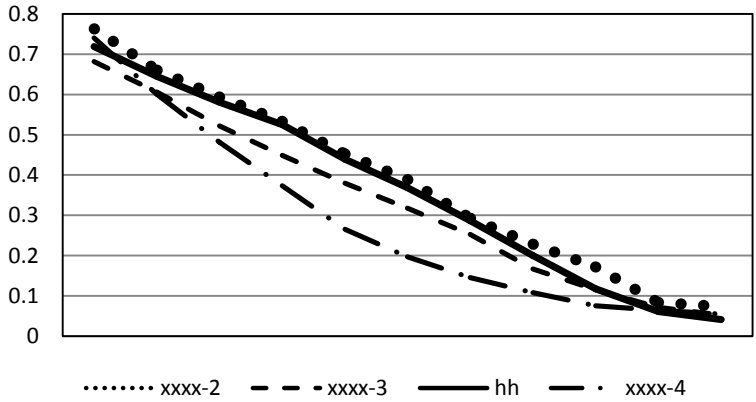

Рисунок 3. Коллекция Legal2007, 11-точечный график TREC

В дорожке поиска по коллекции KM.RU оценивалось 60 запросов, которые были отобраны случайно. При этом из случайного набора отсеивался мусор, опечатки, явно навигационные запросы на что-то, чего нет в коллекции (типа однокласники.ру). В пулы попало по 50 первых результатов в каждом прогоне. На момент написания данной статьи результаты были рассчитаны только по одной оценке, поэтому их следует считать предварительными.

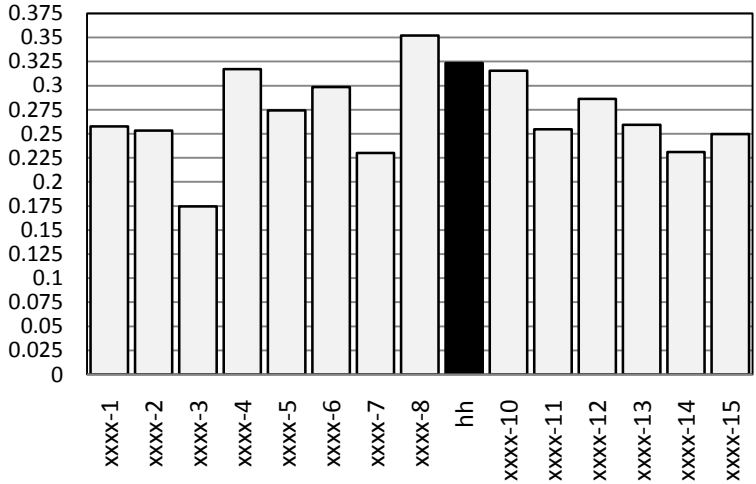

Рисунок 4. Коллекция KM.RU, метрика Average precision

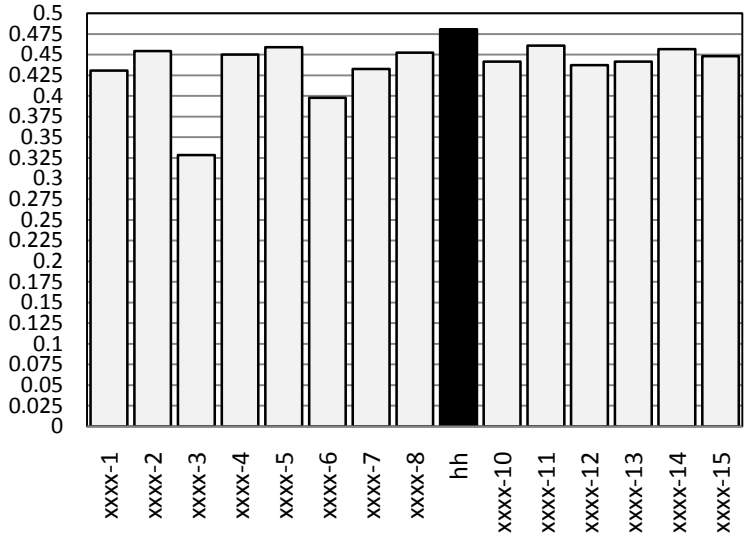

Рисунок 5. Коллекция KM.RU, метрика Precision(10)

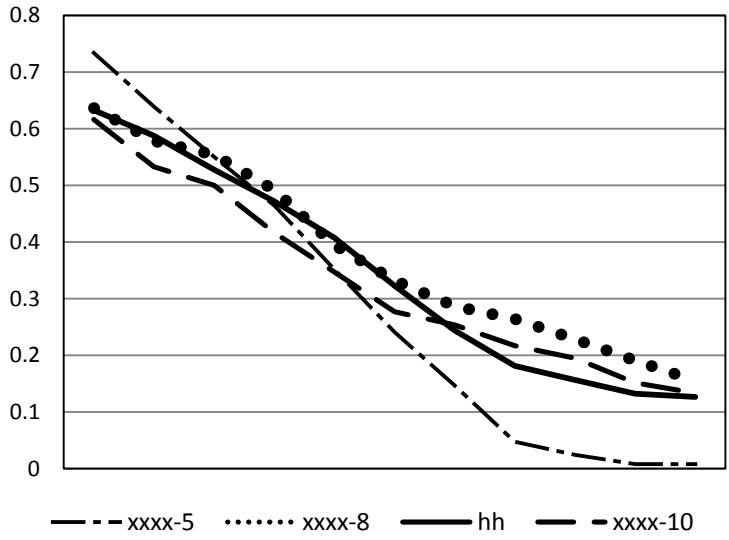

Рисунок 6. Коллекция KM.RU, 11-точечный график TREC

Как видно из диаграмм и графиков, наш алгоритм ведет себя схожим образом на коллекции нормативно-правовых документов и коллекции KM.RU. Уступая по метрике average precision лишь прогону 8 на KM.RU и прогону 2 в правовой коллекции, наш алгоритм является лучшим по метрике precision(10) в обеих коллекциях.

Высокая оценка по метрике precision(10) для нас была несколько неожиданной, поскольку оптимизация параметров алгоритма производилась на основе метрики average precision. Также мы не ожидали высоких оценок по коллекции KM.RU, поскольку из-за нехватки времени оптимизация параметров для этой коллекции не производилась, и для выполнения заданий этой дорожки использовались параметры, оптимизированные под коллекцию правовых документов.

Отказ от использования ссылочного ранжирования не привел к отставанию от большинства других участников. Это может косвенно свидетельствовать как о том, что в остальных прогонах также не учитывались гиперссылки, так и о том, что ссылочное ранжирование в рассматриваемых коллекциях не дает сильного преимущества.

Обращает на себя внимание характерный провал нашей системы в правой части 11-точечного графика TREC по коллекции KM.RU. В результате анализа поисковых запросов, по которым наша система показала слабые результаты, мы выяснили, что причина кроется в несовершенном алгоритме отсечения предположительно нерелевантных документов (pruning). Наша система производила оценку документов в 2 этапа: сначала находились все документы, предположительно соответствующие запросу, и для них вычислялось слагаемое W_{doc} из формулы (1); затем рассчитывались остальные слагаемые только для топ-N документов, полученных на первом этапе, а прочие документы отбрасывались. Оказалось, что в такой насыщенной дубликатами коллекции, как KM.RU, использование слишком маленького значения для N (100) приводит к потере большого количества релевантных документов. Простое увеличение параметра N в наших последующих экспериментах привело к значительному улучшению результатов по метрике average precision и, соответственно, к поднятию 11-точечного графика TREC.

4. Заключение

Участие в РОМИП-2008 оказалось для нас чрезвычайно полезным. Сравнение с прогонами других участников семинара позволило выявить достоинства и недостатки разработанного нами алгоритма ранжирования. В целом мы удовлетворены результатами нашей системы, хотя у нас и есть основания полагать, что в будущем эти результаты можно улучшить.

Семинар действительно является замечательной возможностью получить объективную оценку качества работы поисковой системы. Кроме того, мы считаем, что ежегодная публикация трудов участников РОМИП служит очень полезному делу – созданию качественного русскоязычного ресурса по информационному поиску.

Литература

[1] *М.С. Агеев, Б.В. Добров, Н.В. Лукашевич, А.В. Сидоров*. Экспериментальные алгоритмы поиска/классификации и сравнение с "basic line". Труды второго российского семинара по оценке методов информационного поиска. Под ред. И.С. Некрестьянова, стр. 214, Санкт-Петербург: НИИ Химии СПбГУ, 2004.

[2] *М. Агеев, И. Кураленок, И. Некрестьянов*. Официальные метрики РОМИП 2006.
[3] *С. Татевосян, Н. Брызгалова*. KM.RU на РОМИП-2007.
[4] *А. Федоровский, М. Костин, А. Проскурин*. Mail.Ru на 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.
[5] *J.P. Callan*. Passage-retrieval evidence in document retrieval. Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pages 302-310, Dublin, Ireland, 1994.
[6] *S. Robertson, H. Zaragoza, M. Taylor*. Simple BM25 Extension to Multiple Weighted Fields. Thirteenth ACM international conference on Information and knowledge management, pages 42-49, 2004.