HeadHunter на РОМИП-2009

© А.В. Сафронов

HeadHunter
safronov@hh.ru

Аннотация

Статья посвящена участию компании HeadHunter в дорожках текстового поиска на семинаре РОМИП-2009. Основное внимание уделено алгоритму ранжирования документов. Описывается формула для оценки «кучности» слов из поискового запроса в документе. Приводятся полученные результаты.

1. Введение

На семинаре РОМИП-2009 компания HeadHunter участвовала в дорожках ad hoc поиска. Для выполнения заданий семинара мы использовали экспериментальную поисковую систему собственной разработки с оригинальным алгоритмом ранжирования.

HeadHunter участвовал в РОМИП второй раз. В прошлом году мы получили неплохие результаты (см. [2]), однако представленная система не была лишена недостатков. Прежде всего, нас не совсем устраивали результаты, полученные по метрике average precision. Поэтому при участии в семинаре в 2009 году наши усилия были направлены в первую очередь на создание алгоритма, который позволил бы получить высокие оценки по метрике average precision.

Далее описывается сам алгоритм ранжирования, а также рассматриваются полученные результаты.

2. Описание алгоритма ранжирования

2.1 Общая формула

Для ранжирования документов используется формула, учитывающая несколько различных факторов:

$$Score = k_{doc} * BM25_{doc} + k_{title} * BM25_{title} + k_{begin} * BM25_{begin} + k_{proximity} * HHProximity$$

где:

$BM25_{doc}$ - вес всего документа, рассчитанный по формуле из работы [6];

$BM25_{title}$ - вес заголовка;

$BM25_{begin}$ - вес начальной части документа;

$HHProximity$ - вес «кучности» слов запроса в документе.

2.2 Кучность

Классическая формула BM25 не учитывает взаимное расположение слов из запроса в документе. Однако можно предположить, что в релевантных документах слова запроса должны быть расположены близко друг к другу. Существуют различные методы для оценки этой близости между словами запроса (см. работы [4],[5],[3]).

Мы разработали свой алгоритм для оценки взаимной близости («кучности») слов запроса в документе.

Пусть для каждого слова из запроса у нас имеется список позиций, на которых это слово встречается в документе. Для каждого вхождения слова в документ мы можем оценить, насколько оно плотно окружено другими словами из запроса. Если использовать образную аналогию, каждое вхождение слова мы рассматриваем как центр некой мишени. Тогда наша задача состоит в том, чтобы оценить, насколько кучно легли другие слова из запроса вокруг этого центра.

$$tc(d,t,p) = \sum_{t' \in q} \left(\frac{idf(t')}{lmd(d,p,t')^z} + \frac{idf(t')}{rmd(d,p,t')^z} \right) * ts(t,t')$$

где

$tc(d,t,p)$ - функция, оценивающая «кучность» расположения слов запроса вокруг слова t в позиции p документа d;

$idf(t')$ - вес слова t';

$lmd(d,p,t')$ - расстояние от позиции p до ближайшего слева вхождения слова t';

$rmd(d,p,t')$ - расстояние от позиции p до ближайшего справа вхождения слова t';

z – константа (принималась равной 1.75);

$ts(t,t')$ - функция, понижающая влияние на «кучность» вокруг слова t других вхождений этого же самого слова.

$$ts(t,t') = \begin{cases} 0.25, & t = t' \\ 1, & t \neq t' \end{cases}$$

Таким образом, наша оценка «кучности» вокруг любого вхождения слова в документ зависит от того, насколько близко все остальные слова запроса расположены к данному вхождению. При этом важна также «тяжесть» окружающих слов. Для достижения максимального значения функции tc необходимо, чтобы самые «тяжелые» слова запроса стояли «вплотную» к рассматриваемому вхождению.

Итак, мы можем посчитать «кучность» для любого вхождения в документ любого слова из запроса. Теперь введем понятие «суммарной кучности» слова. Под ним мы будем подразумевать сумму значений функции tc для каждого вхождения слова в документ.

$$atc(d,t) = \sum_{p \in P_t^d} tc(d,t,p)$$

Где

$atc(d,t)$ – суммарная кучность вокруг слова t в документе d;

P_t^d - множество позиций, занимаемых словом t в документе d.

Наконец, общую оценку взаимной близости слов запроса в документе будем рассчитывать следующим образом:

$$HHProximity = \log\left(1 + \sum_{t \in q} atc(d,t) * idf(t)\right)$$

Функция $HHProximity$ принимает тем более высокие значения, чем больше в документе групп близко расположенных друг к другу «тяжелых» слов из запроса.

3. Результаты

Мы участвовали в следующих дорожках:

- Ad hoc поиск по коллекции BY.web;
- Ad hoc поиск по коллекции KM.ru;
- Ad hoc поиск по коллекции нормативных документов Legal-2007.

Обучение системы производилось вручную на основе прошлогодних таблиц релевантности.

Описание метрик, с помощью которых осуществлялась оценка результатов, можно найти в [1].

Для коллекции BY оценка производилась на основе 550 запросов (из них 50 было взято из прошлогодней дорожки). Глубина пула была равна 20. При оценке с сильными требованиями к релевантности наша система показала по метрике average precision самые высокие результаты. Также наш алгоритм оказался первым по метрикам precision, recall, r-precision, bpref и bpref-10. При оценке со слабыми требованиями к релевантности система продемонстрировала лучший результат по метрикам average precision, precision, recall и bpref-10.

	Prec	Prec (10)	R-prec	Avg prec	Bpref	Bpref 10	Recall
xxx-01	0,114	0,216	0,199	0,180	0,186	0,235	0,368
xxx-02	0,000	0,001	0,000	0,001	0,000	0,002	0,003
xxx-03	0,000	0,001	0,000	0,000	0,000	0,000	0,001
xxx-04	0,110	0,282	0,266	0,262	0,252	0,340	0,625
xxx-05	0,049	0,113	0,105	0,107	0,099	0,143	0,271
xxx-06	0,058	0,173	0,174	0,166	0,162	0,234	0,438
xxx-07	0,062	0,193	0,184	0,170	0,165	0,233	0,438
xxx-08	0,096	0,231	0,221	0,211	0,207	0,278	0,534
xxx-09	0,096	0,234	0,222	0,211	0,207	0,276	0,531
xxx-10	0,096	0,238	0,216	0,211	0,202	0,282	0,532
hh-alpha	0,124	0,290	0,274	0,279	0,263	0,353	0,604
hh-beta	**0,125**	0,305	**0,289**	**0,291**	**0,275**	**0,364**	**0,646**
xxx-13	0,084	0,224	0,211	0,196	0,195	0,260	0,481
xxx-14	0,109	0,300	0,276	0,267	0,258	0,348	0,612
xxx-15	0,105	0,292	0,258	0,248	0,245	0,327	0,584
xxx-16	0,113	**0,312**	0,288	0,282	0,270	0,360	0,640
xxx-17	0,061	0,184	0,129	0,105	0,120	0,150	0,227

Таблица **1**. Коллекция **BY**, оценка **AND**.

	Prec	Prec (10)	R-prec	Avg prec	Bpref	Bpref 10	Recall
xxx-01	0,198	0,333	0,203	0,174	0,203	0,224	0,282
xxx-02	0,008	0,015	0,004	0,003	0,004	0,006	0,007
xxx-03	0,005	0,012	0,005	0,002	0,004	0,005	0,009
xxx-04	0,202	0,482	0,339	0,304	0,341	0,384	0,539
xxx-05	0,096	0,224	0,157	0,141	0,158	0,178	0,247
xxx-06	0,121	0,364	0,240	0,198	0,241	0,278	0,386
xxx-07	0,122	0,368	0,241	0,205	0,242	0,283	0,400
xxx-08	0,179	0,412	0,290	0,249	0,288	0,327	0,470
xxx-09	0,178	0,412	0,291	0,251	0,288	0,327	0,469
xxx-10	0,179	0,424	0,291	0,250	0,290	0,328	0,468
hh-alpha	**0,230**	0,462	0,333	0,305	0,335	0,372	0,507
hh-beta	0,226	0,474	0,347	**0,320**	0,348	**0,389**	**0,550**
xxx-13	0,155	0,405	0,268	0,228	0,265	0,302	0,432
xxx-14	0,202	0,486	0,341	0,305	0,342	0,376	0,518
xxx-15	0,193	0,468	0,318	0,283	0,321	0,360	0,500
xxx-16	0,210	**0,488**	**0,350**	0,317	**0,349**	0,388	0,543
xxx-17	0,107	0,318	0,155	0,114	0,145	0,160	0,200

Таблица **2**. Коллекция **BY**, оценка **OR**.

Для коллекции KM всего было оценено 50 запросов, глубина пула составила 50 документов. При оценке с сильными требованиями к релевантности по average precision наша система показала второй результат, отстав от лидера на 1,8%. При оценке со слабыми требованиями к релевантности система продемонстрировала самый высокий результат по average precision, превысив результат следующего участника на 13%.

Результаты по коллекции нормативно-правовых документов были получены на основе 75 запросов, оцененных юристами. По метрике average precision наш алгоритм показал лучшие результаты. Также он оказался лучшим по метрикам precision(5), bpref-10 и recall, а также разделил первое место по метрике precision(10) с другим участником.

	Prec	Prec (10)	R-prec	Avg prec	Bpref	Bpref 10	Recall
xxx-01	0,172	0,300	0,246	0,231	0,229	0,270	0,434
xxx-02	0,090	0,169	0,131	0,120	0,117	0,154	0,293
xxx-03	0,185	0,383	0,272	0,276	0,257	0,310	0,642
xxx-04	0,107	0,176	0,147	0,117	0,121	0,146	0,355
xxx-05	0,118	0,252	0,199	0,156	0,168	0,195	0,374
xxx-06	0,178	0,321	0,274	0,248	0,238	0,281	0,647
xxx-07	0,185	0,326	0,292	0,257	0,258	0,293	0,625
hh-beta	0,184	0,381	0,302	0,308	0,286	0,346	0,634
xxx-09	0,158	0,386	0,265	0,247	0,253	0,283	0,494
xxx-10	**0,198**	0,379	**0,342**	**0,313**	**0,318**	**0,351**	0,638
xxx-11	**0,198**	**0,388**	0,328	0,291	0,293	0,342	**0,638**
xxx-12	0,078	0,290	0,168	0,152	0,158	0,178	0,269

Таблица **3**. Коллекция **KM**, оценка **AND**.

	Prec	Prec (10)	R-prec	Avg prec	Bpref	Bpref 10	Recall
xxx-01	0,346	0,543	0,286	0,257	0,275	0,294	0,353
xxx-02	0,255	0,485	0,230	0,187	0,212	0,224	0,302
xxx-03	0,368	0,594	0,334	0,308	0,324	0,344	0,474
xxx-04	0,284	0,477	0,235	0,184	0,210	0,237	0,331
xxx-05	0,285	0,530	0,268	0,220	0,250	0,263	0,336
xxx-06	0,327	0,574	0,344	0,289	0,309	0,345	0,465
xxx-07	0,342	0,574	0,352	0,294	0,318	0,344	0,463
hh-beta	**0,385**	**0,677**	**0,402**	**0,383**	**0,392**	**0,424**	**0,540**
xxx-09	0,310	0,602	0,305	0,263	0,287	0,300	0,391
xxx-10	0,379	0,585	0,377	0,338	0,356	0,381	0,487
xxx-11	0,381	0,579	0,377	0,332	0,353	0,381	0,492
xxx-12	0,219	0,538	0,210	0,181	0,201	0,211	0,250

Таблица **4**. Коллекция **KM**, оценка **OR**.

	Prec	Prec (10)	R-prec	Avg prec	Bpref	Bpref 10	Recall
xxx-01	**0,308**	0,476	0,356	0,331	0,340	0,382	0,532
xxx-02	0,275	0,466	0,357	0,321	0,343	0,364	0,504
xxx-03	0,275	0,503	0,345	0,311	0,334	0,357	0,517
xxx-04	0,208	0,437	0,260	0,222	0,251	0,274	0,424
xxx-05	0,215	0,413	0,345	0,298	0,328	0,356	0,470
xxx-06	0,113	0,218	0,162	0,123	0,151	0,168	0,244
xxx-07	0,049	0,084	0,055	0,037	0,047	0,065	0,138
hh-alpha	0,273	**0,515**	0,359	**0,344**	0,349	**0,412**	**0,579**
xxx-09	0,239	0,513	0,371	0,334	**0,356**	0,386	0,515
xxx-10	0,255	0,492	0,368	0,323	0,350	0,380	0,513
xxx-11	0,238	**0,515**	**0,369**	0,333	0,354	0,385	0,515
xxx-12	0,229	0,490	0,312	0,302	0,296	0,361	0,500
xxx-13	0,231	0,503	0,318	0,303	0,309	0,361	0,507
xxx-14	0,239	0,503	0,343	0,325	0,327	0,385	0,518
xxx-15	0,235	0,497	0,335	0,318	0,320	0,378	0,511
xxx-16	0,232	0,503	0,318	0,302	0,308	0,360	0,506
xxx-17	0,238	0,503	0,350	0,323	0,339	0,375	0,515
xxx-18	0,211	0,427	0,269	0,244	0,257	0,300	0,494

Таблица **5**. Коллекция **Legal**.

4. Заключение

Поставленную в этом году задачу получения высоких оценок по метрике average precision мы считаем в целом выполненной. Главным итогом участия в семинаре для нас стало подтверждение предположения, что формула оценки «кучности» позволяет получать хорошие результаты. Алгоритм ранжирования на основе функции HHProximity кажется нам довольно перспективным для дальнейших исследований.

Как и в прошлом году, участие в семинаре для нас было чрезвычайно увлекательным. Автор благодарен организаторам семинара за неизменную готовность оперативно отвечать на все возникающие вопросы, а также за проявленную терпимость к нашему отставанию от графика сдачи результатов. Также хочется сказать «спасибо» остальным участникам семинара, в особенности участникам дорожек ad hoc поиска. Полученными результатами мы в немалой степени обязаны публикациям семинара за прошлые годы, поэтому довольно успешное выступление нашей системы мы

в первую очередь рассматриваем как успех самой инициативы РОМИП.

Литература

[1] *М. Агеев, И. Кураленок, И. Некрестьянов*. Официальные метрики РОМИП'2007. Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2007-2008. Стр. 237-247.

[2] *А. Сафронов*. HeadHunter на РОМИП-2008. Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2007-2008. Стр. 33-42.

[3] *S. Buttcher, C. Clarke, B. Lushman*. Term proximity scoring for ad-hoc retrieval on very large text collections. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. Pages 621-622.

[4] *C. Clarke, G. Cormack, E. Tudhope*. Relevance ranking for one to three term queries. Information processing & management, vol. 36, 2000. Pages 291-311.

[5] *Y. Rasolofo, J. Savoy*. Term Proximity Scoring for Keyword-Based Retrieval Systems. Proceedings of the 25th European Conference on IR Research (ECIR 2003). Pages 207-218.

[6] *S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford*. Okapi at TREC-3. In Proceedings of the Third Text REtrieval Conference (TREC 1994).