Яндекс на РОМИП-2006

А. Гулин

Яндекс на РОМИП-2006

- Яндекс участвовал в web-adhoc и legal поисковых дорожках РОМИП
- Использовался стандартный и экспериментальный алгоритм текстового ранжирования

Я

Факторы, учтенные в ранжировании

- Встречаемость слов
- Встречаемость пар слов
- Все слова запроса в документе
- Запрос целиком в документе
- Pseudo relevance feedback
  - По теме страницы
  - По используемой лексике

Встречаемость слов

- Перепробовали много вариантов, включая BM3, PL2 итп, лучше всего работает BM25
- В качестве IDF хорошо работает вероятность “выделенности” слова в модели 2-х Пуассонов
- Форматирование учитывается в виде отдельного слагаемого

Встречаемость слов

$$W_{single} = \log(p) * (TF_1 + 0.2 * TF_2)$$

$$TF_1 = \frac{TF}{TF + k_1 + k_2 * DocLength}, k_1 = 1, k_2 = 1/350$$

$$TF_2 = \frac{Hdr}{1 + Hdr}$$

$$p = 1 - \exp(-1.5 * \frac{CF}{D})$$

Я

Встречаемость пар слов

- Положительно влияют слова подряд, через одно и в обратном порядке
- В качестве IDF хорошо работает сумма IDF слов пары
- Учет форматирования пар слов не дает преимущества

Много слов запроса в документе

- Наличие всех слов запроса в документе – важнейший фактор
- Дилемма показывать ли документы с малым количеством слов запроса
- Учет наличия текста запроса в документе
- Учет количества предложений похожих на запрос

Я

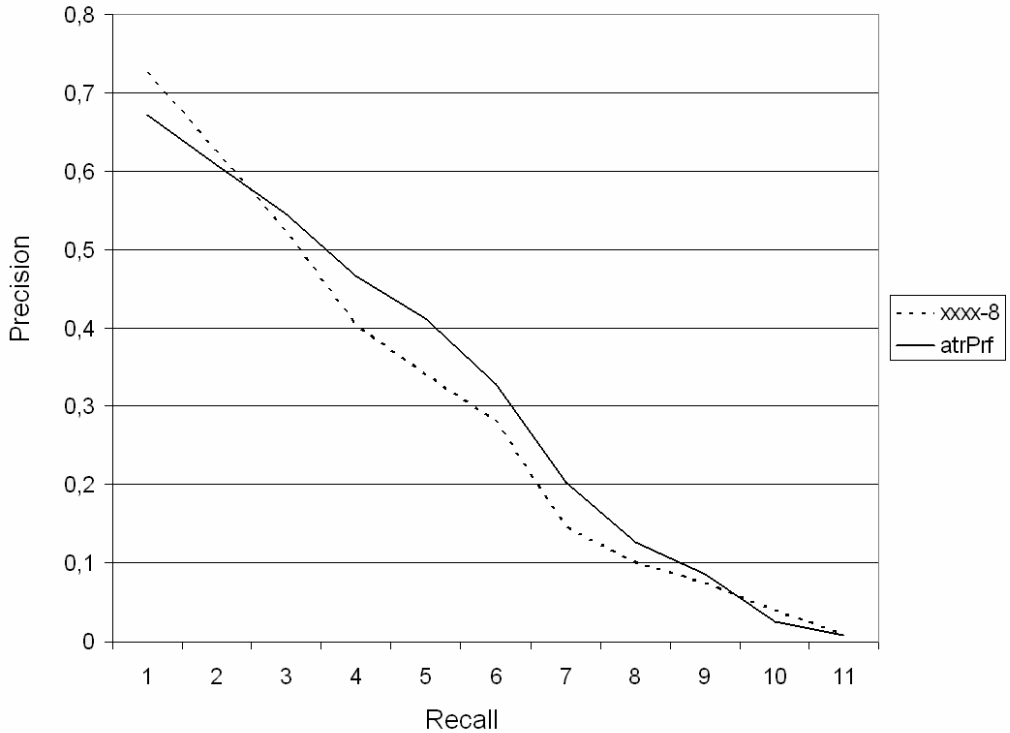
Pseudo relevance feedback

- Считаем найденное в верхушке релевантным, похожее поднимаем
- Мера похожести может быть любой
  - Тема документа
  - Лексика

Похожесть лексики

- Строим граф “слово-документ”
- Строим компактное представление этого графа с помощью групп слов
- Типичная группа: дружба, понимание, тепло, нежность, страсть, забота, узы, брак, семейный, диван, тысячелетие, уставать, бессмертие …

Результат


0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Precision
1
2
3
4
5
6
7
8
9
10
11
Recall
xxxx-8
atrPrf

Использование в Интернете

- Объем данных
- Тип данных
  - Интернет-магазины
  - Доски объявлений, форума
  - Новостные ленты, блоги
- Спам
- Оптимизированные сайты