Денис Бессонов,
руководитель отдела продвижения «Илма Груп»,
автор сео-блога mexboy.ru

Текстовое ранжирование в Яндексе. Особенности подхода TF*IDF.

Денис Бессонов,
руководитель отдела продвижения «Илма Груп»,
автор сео-блога mexboy.ru

Подход TF*IDF

Произведение TF*IDF определяет уровень соответствия документа запросу.

Множитель TF – прямая частота вхождения запроса в документ (отвечает за встречаемость термина в содержании документа), можем влиять

Множитель IDF – обратная частота термина в коллекции (отвечает за редкость употребления запроса во всех документах коллекции, в нашем случае базы поисковой системы), не можем влиять

Денис Бессонов,
руководитель отдела продвижения «Илма Груп»,
автор сео-блога mexboy.ru

Классический случай подхода TF*IDF

$$TF = \frac{n_i}{\sum_k n_k}$$

где n_i - количество употреблений i-го однословника, знаменатель – общая длина документа в словах

$$IDF = \log \frac{D}{DF_i}$$

где D – общее количество документов в коллекции, в нашем случае поисковой базе, знаменатель - число документов, содержащих i-й однословник

Выводы: рулит плотность вхождения

Денис Бессонов,
руководитель отдела продвижения «Илма Груп»,
автор сео-блога mexboy.ru

Подход TF*IDF в Яндексе образца 2006-2007 г.

$$TF = \frac{n_i}{\max(7, \sqrt{n_{\max}})}$$

где n_i - количество употреблений i-го однословника, $n_{\max}$ – количество вхождений в документ самого частотного однословника

$$IDF = \sqrt{\ln p_i}$$

$$p_i = \frac{TotalLemms}{CF_i}$$

где TotalLemms – общее количество терминов в коллекции, в нашем случае длина поисковой базы в словах, CF_i – количество вхождений туда i-го однословника

Денис Бессонов,
руководитель отдела продвижения «Илма Груп»,
автор сео-блога mexboy.ru

Анализ подхода TF*IDF образца 2006-2007 г.

1) рулит встречаемость однословника в документе;
2) максимальная текстовая релевантность, когда $n_i = n_{\max}$
3) плотность вхождения однословника в документ не влияет на ранжирование;
4) ресурс текстовой релевантности неограничен и растет в лучшем случае $\sim \sqrt{n_{\max}}$

Денис Бессонов,
руководитель отдела продвижения «Илма Груп»,
автор сео-блога mexboy.ru

Гипотеза текущего подхода TF*IDF в Яндексе

Предпосылки:
1) документы с огромными псевдо-естественными текстами и высокой плотностью содержания в них продвигаемых запросов;
2) небольшие тексты с высокой плотностью содержания ключевых запросов.

Формула с РОМИП 2006:

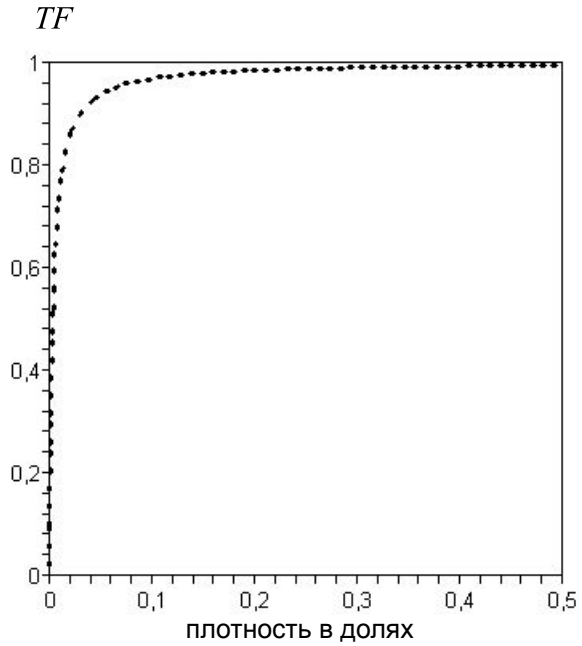
$$TF = \frac{n_i}{n_i + k_1 + k_2 \cdot Doclength}$$

где n_i - количество употреблений i-го однословника, Doclength – длина документа в словах, k_1, k_2 - некоторые постоянные числовые коэффициенты

Денис Бессонов,
руководитель отдела продвижения «Илма Груп»,
автор сео-блога mexboy.ru

Анализ формулы для TF

1) Чем выше плотность вхождения однословника в документ при фиксированной его длине, тем больше TF и выше текстовая релевантность



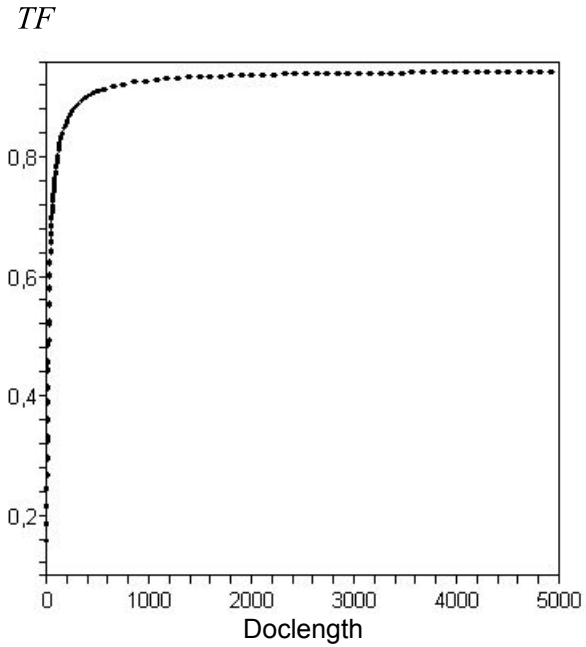
Doclength = 3000 слов,
k_1 =1, k_2 =1/350

Но TF ограничена и, начиная с некоторого значения плотности вхождения однословника, увеличивается слабо

Денис Бессонов,
руководитель отдела продвижения «Илма Груп»,
автор сео-блога mexboy.ru

Анализ формулы для TF

2) Чем больше длина документа при фиксированной плотности вхождения однословника, тем выше TF и текстовая релевантность



Плотность однословника равна 0.05 (5%), k_1 =1, k_2=1/350

Но TF ограничена и, начиная с некоторой длины документа, увеличивается слабо

Денис Бессонов,
руководитель отдела продвижения «Илма Груп»,
автор сео-блога mexboy.ru

Выводы и рекомендации

1) ресурс использования текстовой релевантности ограничен;
2) анализ формулы согласуется с предпосылками;
3) правило “один запрос – одна страница” еще более актуально в такой модели для TF;
4) рулят объемные тексты с высокой плотностью содержания ключевых запросов (но не стоит переоптимизировать)

Денис Бессонов,
руководитель отдела продвижения «Илма Груп»,
автор сео-блога mexboy.ru

Спасибо за внимание! Пожалуйста, вопросы.

Пишите на denis@ilma-group.ru или в блог www.mexboy.ru, если остались вопросы.