Алгоритм поиска Яндекс

© Михаил Маслов

ООО Яндекс
maslov@yandex-team.ru

Аннотация

Описан алгоритм поиска: фильтрации и ранжирования для поисковой системы Яндекс.

1. Процесс поиска

Процесс поиска Яндекс разбит на две фазы: «фильтрацию», то есть принятие решения о включении документа в список найденных, и ранжирование найденных документов

2. Основные факторы, влияющие на "фильтрацию"

2.1 Вид ограничения контекста, или "способ трактовки пробела в запросе"

Можно выделить два основных ограничения контекста - "слова должны быть в одном предложении" и "слова должны быть в одном документе". Можно задавать и другие ограничения - посредством указания "допустимого расстояния" в словах или предложениях.

2.1 Параметр нестрогости поиска - число от 0 до 100

Основной режим поиска Яндекса — взвешенный поиск по кворуму. Чем более весОм термин, тем он дает большее "количество голосов за документ". Число 0 соответствует *наименьшей* нестрогости, т.е. оператору AND. Число 100 соответствует *наибольшей* нестрогости, т.е. оператору OR.

3. Основные факторы, влияющие на ранжирование найденных документов

1. Наиболее весомый фактор — приоритет степени соответствия запросу. Различаются три приоритета — "совпадение фразы", "строгое соответствие" и "нестрогое соответствие". Документы в ранжированном списке найденного делятся на соответствующие три непересекающихся группы.
3. 2. Соответствие порядка и близости искомых терминов в тексте документа порядку и близости терминов в запросе.Вес термина, вычисляемый по tf*idf. Похож на классический, [1] за исключением того, что вместо документной частоты используется частота термина в корпусе; кроме того, величина tf модифицируется степенной функцией с показателем меньше 1 - одна из мер по борьбе с т.н. накачкой релевантности.
4. "Разметочный коэффициэнт" для веса термина, зависящий от наличия термина в заголовке, в тэге "meta keywords" и т.п.
5. "Антиспамовый" коэффициэнт – еще один механизм подавления накачки релевантности.

4. Результаты тестирования

К сожалению, из-за досадной технической ошибки при постановке эксперимента, поиск производился по текстам с обрезанным восьмым битом и поэтому разумно оценить работу алгоритмов не удалось.

Литература

1. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval – *Information Processing and Management*, 24:513-523, 1988