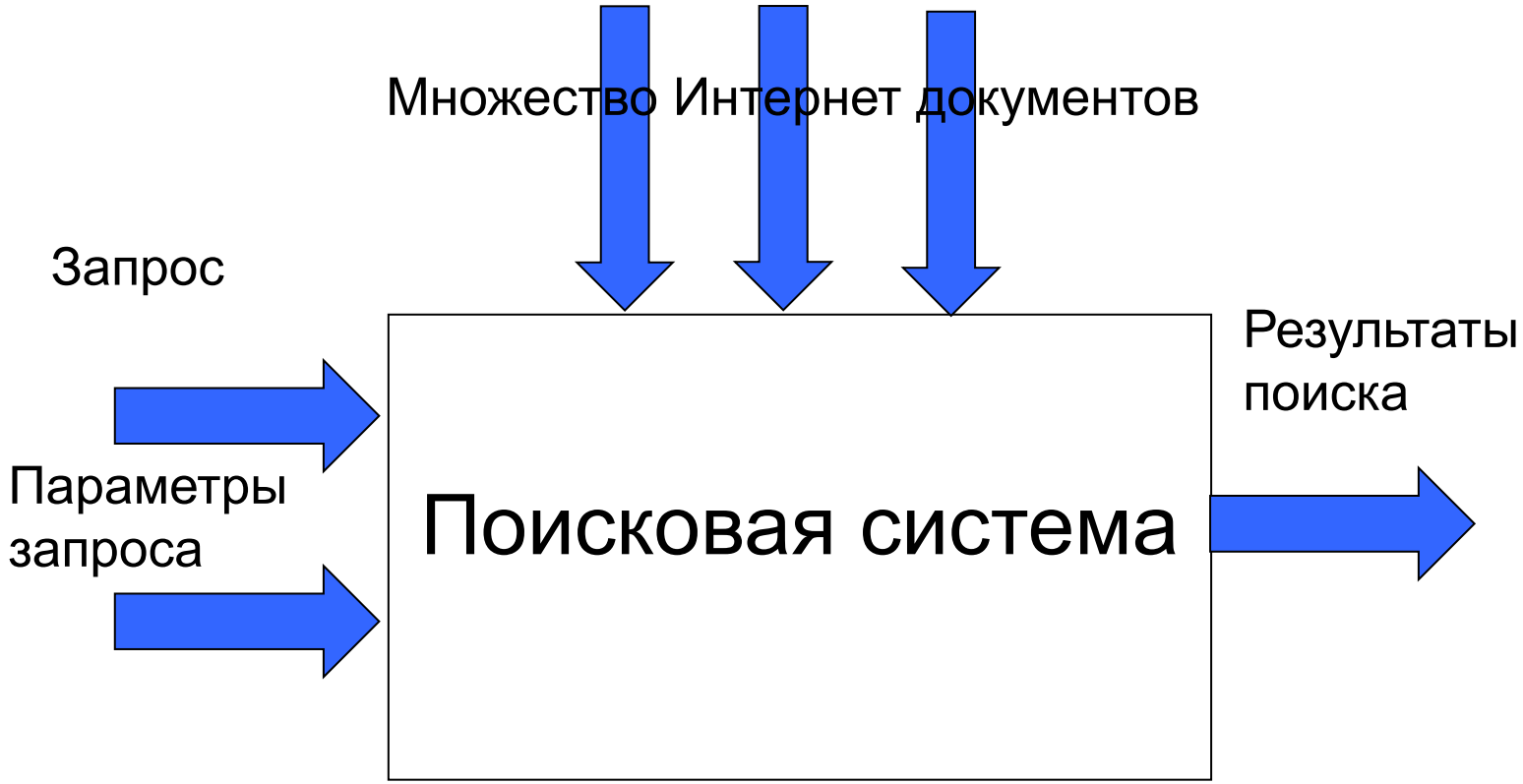
Статистические методы исследования алгоритмов текстового ранжирования поисковых систем

Зябрев Илья Николаевич

генеральный директор, AlterTrader Research Ltd.

Поисковая система как черный ящик



AlterTrader Research Ltd.
ALTERTRADER

Простейшие частотные метрики состава html-страниц

- **Абсолютная теговая частота леммы слова** – количество канонических форм слова в заданном теге html-страницы.

$N(L)=|L:L\in T|$ (1)

- количество вхождений леммы слова L в заданный тег T.

- **Относительная теговая частота леммы слова** – отношение абсолютной теговой частоты леммы слова к общему числу лемм заданного тега html-страницы.

$N\%(L)=N(L)/\sum N(li),\ li\in T$ (2)

- **Различные производные от обратной частоты документа (IDF) или обратной частоты класса ICF метрик**.

$IDF(L)=D/DF(L)$ (3),

где D-общее число документов коллекции, DF(L) - число документов, в которых встречается лемма L

$ICF(L)=TCF/CF(L)$ (4),

где TCF-общее число лемм коллекции, CF(L) - число вхождений леммы L во все документы коллекции.

Производные от ICF/IDF метрики

- IDF(L)*N(L), IDF(L)*N%(L) (5)
- $$\sum_j \frac{IDF(L)}{\sum_i IDF(l_{i,j})}, \sum_j \frac{Len_j \cdot IDF(L)}{\sum_i IDF(l_{i,j})}$$ (6)

где li,j-все леммы j-го предложения, содержащего L, Lenj-количество слов j-го предложения.

- $$\sum_j (\sum_i IDF(l_{i,j}) - IDF(L)), \sum_j (\sum_i IDF(l_{i,j}) / Len_j - IDF(L))$$ (7)
- Для каждой характеристики вместо IDF(L) можно использовать ICF(L), log(IDF(L)), log(ICF(L)). Все перечисленные выше метрики вычисляются как для каждой леммы из запроса отдельно, так и для их совокупности

Коэффициенты корреляции

- **Пирсона** (для количественных величин)

$$K = \frac{M(X \cdot Y) - M(X) \cdot M(Y)}{\sqrt{M(X^2) - M(X)^2}\sqrt{M(Y^2) - M(Y)^2}} \quad (8),$$

где - $M(X) = \sum_i M(X_i)/N$ математическое ожидание величины X.

- **Кенделла** (для ранговых величин)

$$K = 2 \cdot S / N(N-1) \quad (9),$$

где S=P-Q, P- суммарное число наблюдений, следующих за текущими наблюдениями с большим значением рангов Y, Q — суммарное число наблюдений, следующих за текущими наблюдениями с меньшим значением рангов Y

- **Спирмена** (для ранговых величин)

$$K = 1 - \frac{6\sum_i d_i^2}{N(N^2-1)} \quad (10),$$

где di=r(Yi)-r(Xi), r(X)-ранг X.

AlterTrader Research Ltd.
ALTERTRADER

Этапы исследования принципов текстового ранжирования

- **Этап 1. Формирования множества данных для анализа.** Делается подборка запросов, максимально исключающая влияние ссылочного фактора. Например, запросы из непопулярных слов или запросы, задающие поиск по одному сайту. Чем больше различных запросов используется для проведения анализа, тем выше их статистическая значимость.
- **Этап 2. Вычисление числовых характеристик.** Характеристики для исследования выбираются на основе проведенных предварительных наблюдений или возникающих в процессе исследования поисковых систем гипотез. В докладе представлены простейшие из них.
- **Этап 3. Вычисление коэффициентов корреляции.** Ранговые коэффициенты вычисляются по формулам (9) или (10), а Пирсона по формуле (8), когда ранги исследуемых характеристик равны, а анализ носит сравнительный характер.
- **Этап 4. Анализ результатов.** Если некоторая характеристика на различных запросах имеет устойчиво высокий по модулю коэффициент корреляции, то делается вывод о том, что она влияет на текстовое ранжирование.

Таблица 1. Коэффициенты корреляции для характеристик без учета тегов

Источник	**Miratools**		**ATR2009**				**-**	**-**
Исходная метрика	**IDF**		**IDF**		**ICF**		**-**	**-**
Формула / **Запрос**	**(6.1)**	**(6.2)**	**(6.1)**	**(6.2)**	**(6.1)**	**(6.2)**	**(1)**	**(2)**
гонор	**-0,52**	-0,5	**-0,64**	-0,48	**-0,67**	-0,55	**-0,5**	**-0,1**
банальность	**-0,39**	0,181	**-0,64**	0,173	**-0,69**	0,195	**-0,18**	**0,083**
клюв	**-0,71**	-0,25	**-0,65**	-0,25	**-0,67**	-0,25	**-0,24**	**-0,05**
зло	**-0,63**	0,318	**-0,66**	0,317	**-0,71**	0,345	**0,29**	**-0,55**
струпья	**-0,68**	-0,69	**-0,76**	-0,65	**-0,78**	-0,77	**-0,67**	**-0,1**
маньяк	**-0,42**	-0,64	**-0,74**	-0,6	**-0,86**	-0,69	**-0,69**	**-0,32**
подзатыльник	**-0,68**	-0,64	**-0,75**	-0,66	**-0,83**	-0,68	**-0,64**	**-0,1**
традиции	**-0,6**	-0,58	**-0,79**	-0,63	**-0,82**	-0,63	**-0,58**	**-0,47**
ученый	**-0,74**	-0,54	**-0,78**	-0,53	**-0,85**	-0,54	**-0,54**	**-0,37**
выдумка	**-0,41**	-0,66	**-0,83**	-0,67	**-0,91**	-0,71	**-0,7**	**-0,58**

Таблица 2. Коэффициенты корреляции для характеристик тега body

Источник	**Miratools**		**ATR2009**					
Исходная метрика	**IDF**		**IDF**		**ICF**			
Формула / **Запрос**	**(6.1)**	**(6.2)**	**(6.1)**	**(6.2)**	**(6.1)**	**(6.2)**	**(1)**	**(2)**
гонор	**-0,51**	-0,5	**-0,66**	-0,5	**-0,67**	-0,51	**-0,49**	**-0,19**
банальность	**-0,34**	0,181	**-0,61**	0,186	**-0,65**	0,188	**-0,19**	**0,032**
клюв	**-0,7**	-0,23	**-0,68**	-0,23	**-0,78**	-0,25	**-0,21**	**-0,04**
зло	**-0,63**	0,311	**-0,66**	0,302	**-0,69**	0,325	**0,291**	**-0,65**
струпья	**-0,42**	-0,44	**-0,52**	-0,45	**-0,53**	-0,49	**-0,44**	**-0,07**
маньяк	**0,151**	-0,62	**-0,79**	-0,59	**-0,8**	-0,68	**-0,68**	**-0,37**
подзатыльник	**-0,7**	-0,73	**-0,78**	-0,69	**-0,86**	-0,75	**-0,68**	**-0,16**
традиции	**-0,61**	-0,62	**-0,78**	-0,62	**-0,84**	-0,64	**-0,59**	**-0,54**
ученый	**-0,73**	-0,55	**-0,79**	-0,59	**-0,84**	-0,59	**-0,55**	**-0,59**
выдумка	**-0,41**	-0,66	**-0,84**	-0,67	**-0,95**	-0,73	**-0,69**	**-0,56**

Таблица 3. Коэффициенты корреляции для характеристик тега title

Источник	**Miratools**		**ATR2009**					
Исходная метрика	**IDF**		**IDF**		**ICF**			
Формула / **Запрос**	**(6.1)**	**(6.2)**	**(6.1)**	**(6.2)**	**(6.1)**	**(6.2)**	**(1)**	**(2)**
гонор	**0,026**	0,499	**-0,36**	0,483	**-0,36**	0,536	**0,416**	**-0,45**
банальность	**-0,35**	0,128	**-0,73**	0,129	**-0,73**	0,131	**0,178**	**-0,54**
клюв	**-0,07**	-0,05	**-0,17**	-0,05	**-0,18**	-0,06	**-0,35**	**-0,63**
зло	**0,179**	0,002	**-0**	0,002	**-0**	0,002	**-0**	**-0,76**
струпья	**-0,17**	0,333	**-0,43**	0,34	**-0,48**	0,367	**0,174**	**-0,55**
маньяк	**-0,21**	-0,54	**-0,44**	-0,57	**-0,46**	-0,55	**-0,18**	**-0,57**
подзатыльник	**-0,01**	0,122	**-0,33**	0,124	**-0,35**	0,127	**0,226**	**-0,45**
традиции	**0,227**	0,193	**-0,36**	0,182	**-0,4**	0,195	**0,256**	**-0,44**
ученый	**0**	0,748	**-0,29**	0,723	**-0,3**	0,751	**0**	**-0,4**
выдумка	**-0,04**	0,159	**-0,34**	0,166	**-0,35**	0,169	**-0,15**	**-0,43**

Двухфакторная линейная регрессионная модель для оценивания позиции оптимизируемой страницы: Y(X1, X2)=a2X2+a1X1+a0

- Система уравнений по МНК

$$N \cdot a0 + a1\sum_i X1_i + a2\sum_i X2_i = \sum_i Y_i$$

$$\sum_i X1_i \cdot a0 + a1\sum_i X1_i^2 + a2\sum_i (X1_i \cdot X2_i) = \sum_i (Y_i \cdot X1_i)$$

$$\sum_i X2_i \cdot a0 + a1\sum_i (X1_i \cdot X2_i) + a2\sum_i X2_i^2 = \sum_i (Y_i \cdot X2_i)$$

- Решение системы

$$a2 = \frac{(N\sum_i (Y_i \cdot X2_i) - \sum_i X2_i \sum_i Y_i)(N\sum_i X1_i^2 - (\sum_i X1_i)^2) - (N\sum_i (Y_i \cdot X1_i) - \sum_i X1_i \sum_i Y_i)(N\sum_i (X1_i \cdot X2_i) - \sum_i X1_i \sum_i X2_i)}{(N\sum_i X2_i^2 - (\sum_i X2_i)^2)(N\sum_i X1_i^2 - (\sum_i X1_i)^2) - (N\sum_i (X1_i \cdot X2_i) - \sum_i X1_i \sum_i X2_i)^2}$$

$$a1 = \frac{N\sum_i (Y_i X1_i) - \sum_i X1_i \sum_i Y_i}{N\sum_i X1_i^2 - (\sum_i X1_i)^2} - a2\frac{N\sum_i (X1_i \cdot X2_i) - \sum_i X1_i \sum_i X2_i}{N\sum_i X1_i^2 - (\sum_i X1_i)^2}$$

$$a0 = \frac{\sum_i Y_i - a1\sum_i X1_i - a2\sum_i X2_i}{N}$$

Ваши вопросы