МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ Н. Э. БАУМАНА
Факультет информатики и систем управления
Кафедра теоретической информатики и компьютерных технологий

Лабораторная работа №6
по курсу «Информационный поиск»

«Relevance feedback»

Выполнил:
студент группы ИУ9-21М
Беляев А. В.

Проверила:
Лукашевич Н. В.

Москва 2019

1 Цель работы

Необходимо ценить, как изменится изначальный запрос пользователя при получении relevance feedback от него. Исходный запрос: *отбор кандидатов*.

Релевантный фидбек (релевантные документы):

- *Кандидат отобрать претендент*
- *Отбор выбрать претендент*

DF:

- Отбор - 70000
- Кандидат - 70000
- Отобрать - 50000
- Претендент - 30000
- Выбрать - 70000

Число документов: 1000000
$a = 0.7, b = 0.3$

2 Ход работы

Воспользуемся формулой Роккио:

$$Q_{mod} = (a * Q_0) + (b * \frac{1}{|D_{rel}|} * \sum_{D_j \in D_{rel}} Dj) - (c * \frac{1}{|D_{irrel}|} * \sum_{D_k \in D_{irrel}} Dk)$$

$$D_{rel} = D_1 + D_2$$

Найдем нормализованное векторное представление документа D_1.

$$D = \frac{u(d)}{\|u(d)\|}$$

,

где $u(d) =< u(w_1, d), u(w_2, d), \cdots >$, а $u(w, d) = TFIDF$.

В свою очередь $TFIDF$ вычисляется следующим образом:

$$TFIDF(w, d) = \ln(TF(w, d) + 1) * \log_{10}(\frac{|D|}{DF(w)})$$

Вычислим значения для Документа D_1 (*Кандидат отобрать претендент*). (Примечание: в формулах используется транслитерация):

- $u(otbor, d_1) = 0$
- $u(kandidat, d_1) = \ln(1 + 1) * \log_{10}(\frac{1000000}{70000}) = 0.8$

- $u(otobrat, d_1) = \ln(1+1) * \log_{10}(\frac{1000000}{50000}) = 0.9$
- $u(pretendent, d_1) = \ln(1+1) * \log_{10}(\frac{1000000}{30000}) = 1.1$
- $u(vybrat, d_1) = 0$

$$D_1 =< 0, 0.8, 0.9, 1.1, 0 >$$

$$normD_1 =< 0, 0.5, 0.55, 0.67, 0 >$$

Проведем аналогичные вычисления для документа 2 (*Отбор выбрать претендент*):

- $u(otbor, d_2) = \ln(1+1) * \log_{10}(\frac{1000000}{70000}) = 0.8$
- $u(kandidat, d_2) = 0$
- $u(otobrat, d_2) = 0$
- $u(pretendent, d_2) = \ln(1+1) * \log_{10}(\frac{1000000}{30000}) = 1.1$
- $u(vybrat, d_2) = \ln(1+1) * \log_{10}(\frac{1000000}{70000}) = 0.8$

$$D_2 =< 0.8, 0, 0, 1.1, 0.8 >$$

$$normD_2 =< 0.51, 0, 0, 0.7, 0.51 >$$

Таким образом у нас есть релевантые документы, прдставленные нормализованными вертоками $TFIDF$.

Количество релевантых документов: $|D_{rel}| = 2$

Далее следует представить изначальный запрос (*отбор кандидатов*) вектором частот:

$$Q_0 =< 1, 1, 0, 0, 0 >$$

Подставим все в формулу Роккио:

$$Q_{mod} = 0.7* < 1, 1, 0, 0, 0 > +0.3*\frac{1}{2}*(< 0, 0.5, 0.55, 0.67, 0 > + < 0.51, 0, 0, 0.7, 0.51 >) =$$

$$< 0.7, 0.7, 0, 0, 0 > + < 0.07, 0.07, 0.08, 0.2, 0.07 >=$$
$$< 0.77, 0.77, 0.08, 0.2, 0.07 >$$

Таким образом, в оригинальном запросе имели веса лишь слова **отбор** и **кандидат**, а после фидбека появлось еще одно слово со «значащим» весом - **претендент**.

3 Выводы

В лабораторной работе было нагляно продемонстрировано, как вектор запроса смещается к «идеальному» положению при получении релевантной обратной связи от пользователя.