АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ АКТУАЛЬНОСТИ ТЕМ НАУЧНЫХ РАБОТ

Рахиля Гасанова

Институт Информационных Технологий Академии Наук Азербайджана
ул. Ф.Агаева 9, AZ-1141 Баку, Азербайджан, тел.: (99412) 439-72-13 E-mail: rahasanova@gmail.com

Аннотация

В данной статье для оценки диссертационных работ предлагается метод определяющий близость между разделов в диссертации. Одновременно предлагается метод, позволяющий автоматически определить актуальности тем диссертаций. Здесь рассматривается близость между краткой аннотации введения и ссылок, соответствующих введению с помощью метрикой косинуса. Далее предлагается оценка актуальности по распределению литературных ссылок по годам.

Известно, что оценивать диссертационных работ требует от ученых много усилий и времени. Объективность один из важнейших условий в оценивание. Осуществимо упрощать – автоматизировать этот процесс с сохранением объективности. Для этого требуется описать формальную структуру диссертационных работ.

Обозначим через D элементы множества диссертации и представим формальную структуру следующим образом:

$$D = \{I, C, R, L\}$$

Здесь, I – введение, C – множества глав, R – выводы, L – список литературы.

Предположим, что задана совокупность глав $C = \{c_1, c_2, ..., c_n\}$. $T = (t_1, t_2, ..., t_m)$ термины встречающихся в совокупности $C = \{c_1, c_2, ..., c_n\}$. Первая наша задача состоит в определение близости между главами. Для текстовых данных одной из наиболее распространенных мер близости является метрика косинуса [1]. С этой целью в рамках модели векторного пространства (Vector Space Model) каждому термину t_t сопоставляется некоторая положительная взвешенная функция w_{it} [2]. Таким образом, каждое глава будет представлено в виде m-мерного вектора $c_i = (w_{i1}, w_{i2}, ..., w_{im}),\ \ i = \overline{1, n}$. Вес w_{it} термина t_t зависит от частоты его появления в конкретной главе, который определяется формулой TF*IDF (Term Frequency * Inverse Document Frequency):

$$w_{it} = tf_{it} \log\left(\frac{n}{n_t}\right), \tag{1}$$

tf_{it} – частота появления термина t_t в главе c_i, n – общее число глав в диссертации, n_t – количество глав, в которых присутствует термин t_t.

Для определения смысловой близости между главами в диссертационной работе целесообразно выбрать метрику косинуса [3]. Согласно этому метрику мера подобия глав c_i и c_j вычисляется по формуле:

$$\cos(c_i, c_j) = \frac{\sum_{t=1}^{m} w_{it} w_{jt}}{\sqrt{\sum_{t=1}^{m} w_{it}^2} \sqrt{\sum_{t=1}^{m} w_{jt}^2}},\ \ i, j = \overline{1, n}. \tag{2}$$

От зависимости значения $\cos(c_i, c_j)$ мы сможем говорить насколько близки по тематике глава между собой.

На следующем этапе постараемся оценить актуальность тем диссертационных работ. Для этого введем следующие обозначения: A_I – краткая аннотация введения, L_I – список ссылок соответствующих видению.

Пусть $T = (t_1, t_2, ..., t_m)$ будет набор терминов встречающихся в A_I и L_I.

Определим близость между множествами A_I и L_I. Для текстовых данных одной из наиболее распространенных мер близости является метрика косинуса [1]. С этой целью в рамках модели векторного пространства (Vector Space Model) каждому термину t_k сопоставляется некоторая взвешенная функция w_k [2]. Таким образом, множества A_I и L_I будут представлены в виде m-мерных векторов $A_I = (w_1^A, w_2^A, ..., w_m^A)$ и $L_I = (w_1^L, w_2^L, ..., w_m^L)$ соответственно.

$$w_k^{A,L} = tf_k^{A,L} \log\left(\frac{n}{n_k}\right),$$

здесь, $tf_k^{A,L}$ – число появлений термина t_k в множествах A_I и L_I соответственно,

n – общее число множеств (в нашем случае 2),

n_k – число множеств, в которых встречается термин t_k.

Близость между A_I и L_I определяется следующей формулой:

$$\cos(A_I, L_I) = \frac{\sum_{k=1}^{m} w_k^{A_I} w_k^{L_I}}{\sqrt{\sum_{k=1}^{m} \left(w_k^{A_I}\right)^2} \sqrt{\sum_{k=1}^{m} \left(w_k^{L_I}\right)^2}}.$$

Далее в списке литературы ведется разделение по годам и множество L_I разделяется на 3 подмножества: $L_I = (L_I^5, L_I^{6-10}, L_I^{11-})$.

Здесь, L_I^5 – список литературы за последние 5 лет, L_I^{6-10} – список литературы за последние 10 лет с вычетом 5 последних лет, L_I^{11-} – список литературы, предшествующим последним десяти годам.

На следующем этапе определяется близость между этими подмножествами и множеством A_I, то есть вычисляются $\cos(A_I, L_I^5)$, $\cos(A_I, L_I^{6-10})$ и $\cos(A_I, L_I^{11-})$. Близость с наибольшим результатом называется актуальной, относительно актуальной и неактуальной соответственно.

Литература:

[1] Р.М. Алгулиев, Р.М. Алыгулиев, «Аннотирование текстовых документов с определением скрытых тематических разделов и информативных предложений», Автоматика и вычислительная техника, 2007

[2] G. Salton, A. Wong, and C. S. Yang, “A vector space model for automatic indexing”, Communication of the ACM, vol. 18, no. 11, November 1975, pp. 613-620.

[3] Alguliev R.M., Aliguliev R.M. Effective summarization method of text documents // Proc. of the 2005 IEEE/WIC/ACM International Conf. on Web Intelligence (WI’05). – France. – September 19-22, 2005. – Compiegne University of Technology. – P. 264-271.