ВЫБОР ПРОТОТИПА ДЛЯ СПОСОБА ТЕМАТИЧЕСКОГО РАНЖИРОВАНИЯ ТЕРМИНОВ

Ходенева М.А.*, Кудрявцев А.Г.

Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, г. Екатеринбург, Россия
E-mail: masha.hodeneva@mail.ru

CHOOSE A PROTOTYPE FOR THE METHOD OF THEMATIC RANKING OF TERMS

Hodeneva M.A.*, Kudryavtsev A.G.

Ural Federal University, Yekaterinburg, Russia

Annotation. The article deals with the problem of choosing a prototype for the method of thematic ranking of terms. It is analyzed several analogues and the selected compilation prototype on the basis of two of them. In the text of article describes the algorithmic model of the prototype and given the criticism, which consists in the incompleteness of the resulting terminology list, as well as the proposed a solution that will overcome this criticism.

В ходе проведенного литературно-аналитического обзора нами были отобраны четыре аналога.

Первый аналог – тематическое ранжирование предложений. Суть данного способа заключается в отборе предложений, наиболее соответствующих заданной теме с формированием матрицы, у которой столбцы соответствуют векторам ранжирования предложений относительно темы. Для предварительного ранжирования может быть использован любой алгоритм. Данный способ подходит также для терминов и рубрик [1].

Второй аналог – ранжирование терминов по частоте встречаемости, либо производных от нее характеристик (например, суммы коэффициентов ассоциативности для данного термина) [2].

Третий аналог – ранжирование терминов на основе таксономии. В основе данного способа лежит предположение о том, что указанные пользователем в таксономии термины, расположенные на «нижних» уровнях древовидной структуры, в большей степени определяют для него «ценность» публикации, чем термины, расположенные на «верхних» уровнях этого дерева [3].

Последний из аналогов – ранжирование терминов на основе рубрик. На начальном этапе из исходного текста выделяют термины. Далее, задавшись конкретными терминами, проводят поиск библиотечных текстов, содержащих каждый из них. В этих текстах идет поиск рубрики. Далее проводят частотное оценивание вероятности $P(G_l)$ события, связанного с обнаружением l-й рубрики без дополнительных условий, а также вероятности $P(G_l \mid T_k)$ события, связанного с обнаружением указанной рубрики по k-му термину. Применяя далее формулы

Байеса и полной вероятности можно рассчитать сначала вероятности $P(T_k | G_l)$ использования k-го термина при нахождении l-й рубрики, а затем $P(T_k)$ использования k-го термина в процессе нахождения рубрик [4].

Сравнение аналогов по выбранным критериям приведено в таблице.

Таблица 1

Сравнение аналогов

№ п/п	**Оценки по характеристикам:**			
	возможность тематического ранжирования	**однозначность критерия ранжирования**	**простота процедуры ранжирования**	$\sum$
I	1	0	1	2
II	0	0	0	0
III	0	0	0	0
IV	0	1	1	2

В результате выбран компилятивный прототип на основе первого и четвертого аналогов. Алгоритмическая модель прототипа предполагает наличие исходных данных (анализируемых текстов, библиотечных текстов, рубрикатора, тем и их приоритетов), а также выполнение следующих действий: 1. Формирование матрицы значимости рубрик относительно тем; 2. Извлечение терминов из текущего текста; 3. Автоматическое числовое рубрицирование текста; 4. Формирование взвешенного словника. После выполнения всех действий имеем взвешенный словник анализируемого текста.

Недостаток данного прототипа - неполнота терминологического списка: часть информации по терминам пропадает за счет того, что для них не находится рубрик. В связи с этим, данный прототип нуждается в доработке. В качестве решения предложено добавление в структуру системы дополнительного блока восполнения данных по терминам.

1. Тарасов С.Д. Метод тематического ранжирования в задачах автоматического сводного реферирования [Электронный ресурс] / С.Д. Тарасов. Режим доступа: http://aidt.ru/images/documents/2010-02/36_41.pdf
2. Браславский П.И. Автоматическое извлечение терминологии с использованием поисковых машин интернета [Электронный ресурс] / П.И. Браславский, Е.А. Соколов. Режим доступа: http://www.dialog-21.ru/digests/dialog2007/materials/html/14.htm
3. Вдовицын В.Т. Ранжирование документов в системе поиска, основанной на применении онтологии [Электронный ресурс] / В.Т. Вдовицын, В.А. Лебедев. Режим доступа: http://ceur-ws.org/Vol-934/paper19.pdf
4. Болбаков Л.Г. Теорема Байеса в когнитивной семантике образовательных информационных систем [Электронный ресурс] / Л.Г. Болбаков. Режим доступа: http://cyberleninka.ru/article/n/teorema-bayesa-v-kognitivnoy-semantike-obrazovatelnyh-informatsionnyh-sistem