

## **ОЦЕНИВАНИЕ ЗНАЧИМОСТИ ТЕРМИНОВ ПРИ АВТОМАТИЧЕСКОМ СИНТЕЗЕ ОНТОЛОГИЙ УЧЕБНЫХ ДИСЦИПЛИН**

*доцент кафедры искусственного интеллекта, с.н.с., к.т.н., Чалая Л.Э.,*

*Харьковский национальный университет радиоэлектроники,*

*г. Харьков, Украина*

*аспирант кафедры Искусственного интеллекта, Чижевский А.В.,*

*Харьковский национальный университет радиоэлектроники,*

*г. Харьков, Украина*

*аспирантка кафедры Искусственного интеллекта, Шевякова Ю.Ю.,*

*Харьковский национальный университет радиоэлектроники,*

*г. Харьков, Украина*

В данной работе предлагается метод оценки значимости слов и словосочетаний для последующего построения онтологических моделей предметной области. Метод может эффективно использоваться для задач семантического поиска в системах анализа электронных текстов и автоматического создания онтологий.

В последнее время получили широкое распространение исследования в области автоматического синтеза онтологических моделей, позволяющих повысить эффективность систем семантического поиска по запросам пользователей (в корпусе текстов, электронных библиотеках ВУЗов, в сети Интернет) [1]. К важным задачам создания онтологий учебных дисциплин следует отнести определение наиболее значимых терминов для каждого отдельного текста, а также для всего корпуса анализируемых текстов в целом. Наиболее важными словами, отображающими суть предметной области, можно считать слова, которые наиболее часто встречаются как в каждом отдельном тексте, так и в возможно большем количестве текстов из анализируемого корпуса текстов.

В настоящее время наибольшее распространение для оценки важности терминов в онтологиях получили показатели TF (term frequency) и IDF (inverse document frequency). Показатель TF определяется как отношение числа вхождения в документ некоторого слова к общему количеству слов документа [1]. Показатель IDF соответствует инвертированному значению частоты, с которой некоторое слово встречается в документах коллекции. Учёт IDF уменьшает вес широкоупотребительных слов. Эффективным методом поиска специфичных слов, характерных для анализируемого документа, является применение комбинированного коэффициента TF/IDF. По таким показателям оценивается важность слова в пределах отдельного документа или анализируемого корпуса текстов. В то же время представляется

целесообразным решить задачу выделения в тексте не только отдельных слов, но и наиболее важных словосочетаний, отражающих семантику текста.

Автоматическое создание онтологической модели данных предполагает необходимость предварительного выделения концептов (классов, понятий) будущей онтологии и определения связей между ними. Очевидно, что для выделения концептов необходимо найти слова (термины) и словосочетания, наиболее важные (значимые) для рассматриваемой предметной области [1].

В работе [2] был предложен метод поиска таких словосочетаний, основанный на количественной оценке важности элементов текста с использованием значений TF, TF/IDF и так называемых рангов слов. Исходной операцией здесь является предварительное упорядочение по убыванию важности и составление соответствующих списков важных слов для показателей TF и TF/IDF. Под частными рангами слова  $R_1$  и  $R_2$  понимаются значения величин, обратных номерам позиции этого слова в упорядоченных списках для TF и TF/IDF соответственно.

В рассматриваемом подходе предлагается оценивать в анализируемых текстах значимость слов  $w_i$  в  $i$ -м тексте корпуса по значениям коэффициента  $K(w_i)$ , рассчитываемого по следующей зависимости:

$$K(w_i) = \frac{1}{|R_1 - R_2|} * \frac{Q(w_i)}{N_{\max} - N_{\min}},$$

где  $Q(w_i)$  – количество текстов, в которых содержится слово  $w_i$ ;  $N_{\max}$ ,  $N_{\min}$  – наибольшее и наименьшее число вхождений слова  $w_i$  в корпус текстов соответственно.

В соответствии с приведенной зависимостью значимые термины, выделяемые для создания отношений между концептами, должны встречаться в большинстве текстов корпуса и при этом быть максимально равномерно распределены в каждом тексте корпуса текстов.

Важность словосочетаний  $M(W_i)$  предлагается оценивать по произведению рангов соответствующих слов. При этом, поскольку ранги слова нормированы от нуля до единицы, то с увеличением длины словосочетания уменьшается результат такого произведения. В связи с этим его необходимо умножить на количество слов в рассматриваемом словосочетании. Следует также учитывать, что слова, для которых  $R_1 > R_2$ , определяют специфику анализируемого текста, то есть слова из первого списка с большой вероятностью попадут в большинство словосочетаний, характерных для этого текста. Наиболее же специфичными для конкретного текста являются слова второго списка, для

которых  $R_1 > R_2$ , следовательно, именно за счет этого можно повысить релевантность критерия оценки важности словосочетания.

Таким образом, в концептах высшего уровня могут быть использованы словосочетания с наибольшей оценкой  $M(W_i)$  либо отдельные слова из этих словосочетаний, соответственно с большей оценкой важности слова  $K(w_i)$ .

Следующим шагом создания дерева иерархии будущей онтологии является нахождение дочерних концептов. Для этого в концептах высшего уровня должны быть выделены главные смысловые слова (зачастую это термины предметной области) и соответствующие словосочетания. Далее в качестве дочерних концептов словосочетаний выделяются словосочетания, имеющие наибольшее совпадение слов с концептами высшего уровня и, соответственно, наибольшую оценку  $M(W_i)$ . На следующем этапе необходимо выявить связи между найденными концептами. Всего существует три основных типа отношений: «целое-часть» (part-of), «класс-подкласс» (is-a) и остальные (экземпляр-класс, причина-следствие, отношение зависимости и др.). Для отыскания первых двух типов связей должен быть составлен специальный словарь (например, для связи «класс-подкласс» – «является», «в частности», «относится» и др.; для связи «целое-часть» – «входит», «состоит из» и др.). Если в анализируемый текст входят слово или словосочетание из словаря, то следует проверка нахождения правее и левее по тексту ранее выделенных концептов.

По предложенному методу был разработан программный модуль «Termin-O», который может эффективно использоваться для задач семантического поиска в системах анализа электронных текстов и автоматического создания онтологий.

### ***Перечень ссылок***

1. Чалая, Л. Э. Метод автоматического построения онтологии с древовидной структурой концептов [Текст] // Л. Э. Чалая, А.В. Чижевский, Ю.Ю. Шевякова // Материалы Международной научно-технической конференции «Информационные системы и технологии» ИСТ-2012. – Морское: 2012. – С.75.

2. Чалая, Л. Э. Меры важности концептов в семантической сети онтологической базы знаний [Текст] / Л.Э. Чалая, Ю.Ю. Шевякова, А.Ю. Шафроненко// Матеріали другої міжнар. наук.-техн. конф. «Сучасні напрями розвитку інформаційно-комунікаційних технологій та засобів управління». – Киев : КДАВТ, 2011. – С. 51.

## ESTIMATION OF THE SIGNIFICANCE OF TERMS FOR ONTOLOGY AUTOMATIC SYNTHESIS OF ACADEMIC DISCIPLINES

В данной работе предлагается метод оценки значимости слов и словосочетаний для последующего построения онтологических моделей предметной области. Метод может эффективно использоваться для задач семантического поиска в системах анализа электронных текстов и автоматического создания онтологий.

This paper proposes a method for assessing the significance of words and phrases for the subsequent construction of the ontological domain models. The method can be used effectively to problems of semantic search systems analysis of electronic texts and automatically generate of ontologies.

### **Данные об авторах:**

*Чалая Лариса Эрнестовна, доцент кафедры искусственного интеллекта,  
С.Н.С., к.т.н.,*

*Харьковский национальный университет радиоэлектроники,  
г. Харьков, Украина*

*Чижевский Антон Валериевич, аспирант кафедры Искусственного  
интеллекта, Харьковский национальный университет  
радиоэлектроники,  
г. Харьков, Украина*

*Шевякова Юлия Юрьевна, аспирантка кафедры Искусственного  
интеллекта, Харьковский национальный университет  
радиоэлектроники,  
г. Харьков, Украина*

### **Контакты:**

*Чалая Лариса Эрнестовна: тел. +380506360603, (057) 7021471, (057) 7021337, e-mail: kovalivnich@yahoo.com*