ОСНОВЫ ИНФОРМАЦИОННОЙ ТЕХНОЛОГИИ ТЕМАТИЧЕСКОГО РУБРИЦИРОВАНИЯ

Канищева О.В.
Научный руководитель – д.т.н., проф. Шаронова Н.В.
Национальный технический университет
«Харьковский политехнический институт»
(61166, Харьков, ул. Фрунзе, 21, каф. Интеллектуальных компьютерных систем, тел. (057) 707-65-05),
E-mail: olya-kanisheva@rambler.ru

In this given work the method comparator identification, as one of methods Text Mining, for the decision of problems of processing of texts in a natural language in the automated information systems is considered. Thus the problem of the decision of the following problems is put: thematic classification problems, and also definition tonalities publications concerning the certain objects (the organizations, etc.) which met in them, – classification tonalities.

До 85% новых знаний аналитики до сих пор получают, изучая тексты. В ближайшем будущем наиболее востребованными станут системы с максимально автоматизированными ETL-процессами (extract, transfer, load — «извлечение, преобразование, загрузка») структурирования контента. Важной чертой таких систем будет функция оперативного анализа информации, полученной по запросу для выбора дальнейшего направления исследования документов, выполняемая с помощью методов интеллектуального анализа текста.

Автор предлагает рассмотреть такую задачу интеллектуального анализа, как тематическое рубрицирование, т.е. автоматическое определение наличия определенных тем в документе.

Тематическая рубрикация документов в полнотекстовой базе данных основывается на тождественности текстов документов по отношению к определенной теме. Использование данного метода позволило ввести и обосновать понятие дескрипторно – текстового предиката, формально представляющего отношения между текстом и соответствующим ему ключевым термином.

Целью работы является создание информационной технологии тематического рубрицирования документов с помощью метода компараторной идентификации и аппарата алгебры конечных предикатов. В дальнейшем предполагается создание системы, предоставляющей возможность автоматического определения наличия определенных тем в документе – тематическое рубрицирование проблем, а также определение тональности публикации по отношению к определенным объектам (лицам, организациям и др.), встречающимся в них – рубрицирование тональности.