Структуризация текстовых данных

Заброда И.С., *студ.*; Ободяк В.К., *доц.*
Сумской государственный университет, г. Сумы

Text miting [1] – это одно из перспективных направлений интеллектуального анализа данных, которое используется для обработки текстовой информации. Основные задачи Text Mining:

- поиск скрытых связей и корреляций между текстами;
- автоматическое формирование аннотаций и метаданных;
- автоматическая категоризация и систематизация текстов.

Система Text Mining содержит блоки классификации (classification), кластеризации (clustering), построения семантических сетей, извлечения фактов и понятий (feature extraction), суммаризации (summarization), ответа на запросы (question answering), тематического индексирования (thematic indexing), поиска по ключевым словам (keyword searching).

В некоторых случаях набор дополняют средства поддержки и создания таксономии (taxonomies) и тезаурусов (thesauri).

В работе рассматривается задача классификации текстов на основе модифицированного Байесовского классификатора, способного оперировать динамическим набором классов при обучении системы.

Были выделены следующие подзадачи:

1. Предварительная обработка для удаления шума (стоп-слов).
2. Создание словаря лемматизированных ключевых слов.
3. Формирование и проверка гипотезы о тематической принадлежности текста.
4. Расширение множества тематических классов в случае, если ни одна из гипотез не подтвердилась.

Поиск ключевых слов выполняется среди элементов текста, которые начинаются с заглавной буквы. Когда ни одно из ключевых слов не позволяет определить тематическую принадлежность текста, выполняется динамическое формирование нового класса и проводится реструктуризация дивизионной иерархии существующих классов.

1. А.А. Барсетян, М.С. Куприянов и др. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP (СПБ.: БХВ-Петербург: 2007).