МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СУМСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ

ІНФОРМАТИКА, МАТЕМАТИКА, АВТОМАТИКА

IMA :: 2016

МАТЕРІАЛИ
та програма

НАУКОВО-ТЕХНІЧНОЇ КОНФЕРЕНЦІЇ

(Суми, 18–22 квітня 2016 року)

Суми
Сумський державний університет
2016

Применение меры tf-idf и меры странности для выделения ключевых слов при классификации текстов научных статей

Козлова Е. С., *студент*; Романов А. Ю., *старший преподаватель*
Национальный исследовательский университет
«Высшая школа экономики», г. Москва, Россия

В рамках исследования используются две меры для выделения ключевых слов в наборе текстов: tf-idf и weirdness (мера странности) [1]. В исследовании используется выборка из более чем двадцати двух тысяч научных статей из девяти тем УДК. Задача исследования состояла в выделении оптимального набора слов для быстрой классификации заданного текста.

Первой тестировалась мера tf-idf, применение которой для данной задачи наиболее оптимально при объединении всех статей одной темы в единый текст. Полученный результат показывает, что при таком подходе слишком большой вес встречаемости слова в рамках одной темы приводит к увеличению количества незначащих слов в результирующем наборе. В итоге получаем набор ключевых слов, в котором до 20 % элементов повторяются и не имеют практической пользы, что вносит значительную погрешность в эксперимент и уменьшает количество полезных слов.

Следующим этапом эксперимента стало использование меры странности. Аналогично эксперименту с tf-idf тексты объединяются в один единый, но формула позволяет повысить значимость сравнения средних значений встречаемости слова в каждой теме. Это дает возможность эффективно избавляться от равномерно распределенных по темам слов, хотя и допускает незначительное количество погрешностей для извлечения слова, случайно встретившегося в текстах другой тематики. Этот способ показывает значительно более высокую эффективность в рамках поставленной задачи.

Проведенный эксперимент показал, что наиболее оптимально мера странности применима при относительно небольшом количестве объемных текстов, в то время как tf-idf показывает наилучшие результаты при большом количестве текстов при объеме, недостаточном для эффективного применения меры странности.

1. Э.С. Клышинский, Н.А. Кочеткова, *Новые информационные технологии в автоматизированных системах* **17**, 365 (2014).