ОПРЕДЕЛЕНИЕ НОВИЗНЫ ИНФОРМАЦИИ В НОВОСТНОМ КЛАСТЕРЕ

Алексеев А.А.

Московский Государственный Университет им. М.В.Ломоносова,
ф-т Вычислительной Математики и Кибернетики, каф. Алгоритмических Языков
Россия, 119991, г. Москва, ул. Ленинские горы 1, стр. 52, 2-й учебный корпус, ВМК.
Телефон: (495)939-30-10, факс: (095)939-25-96,
E-mail: bmw-motors@mail.ru

Растущие информационные потоки делают невозможным ручные анализ и извлечение необходимой информации из информационных источников. Одной из важных задач при автоматической обработке потока новостей является задача автоматического распознавания новой информации, то есть той информации, которая еще не поступала до текущего момента [1, 2].

В данной работе предложено два различных подхода к задаче определения новизны информации в новостном кластере.

Первый подход основан на представлении предложений в виде вектора идентификаторов в векторно-пространственной модели и последующем их сравнении по косинусовой мере угла между векторами. Каждое пространство соответствует отдельному терму, входящему в данное предложение, а значение идентификатора определяется лексическими характеристиками самого терма.

Второй подход заключается в анализе частотных характеристик слов в новостных коллекциях, их весов и значимости, и дальнейшем ранжировании предложений исследуемых новостных кластеров в соответствии с полученными характеристиками, так что вес (новизна) предложения определяется как сумма найденных характеристик новых слов, входящих в это предложение.

Для оценки качества предложенных методов была сделана ручная разметка предложений реальных новостных кластеров на предмет содержания новой информации. Были установлены веса и пороги, при которых данные методы достигают наилучших результатов с точки зрения человека. Получены сравнительные характеристики двух методов, установлены преимущества и недостатки каждого метода. Приведены описание экспериментальной программы, тестовых данных, формулы, лежащие в основе методов, и анализ полученных результатов.

Литература

1. *Soboroff I.* Overview of the TREC 2004 Novelty Track. – National Institute of Standards and Technology (USA), 2004.
2. *Schiffman B., McKeown K.R.* Machine Learning and Text Segmentation in Novelty Detection. – Columbia University (USA), 2004.