

Секция «Вычислительная математика и кибернетика»

Анализ социальной активности вокруг Интернет-новостей и блогов

Зайцев Владимир

Студент

Югорский государственный университет, Институт (НОЦ) Систем Управления и Информационных Технологий, Ханты-Мансийск, Россия

E-mail: vladimir@zvt.me

В последнее десятилетие социальные сети и Интернет-СМИ стали важным источником обратной связи для бизнеса и, отчасти, электронной коммерции. В тоже время, контент генерируемый этими сайтами остается в большой степени неиспользованным. При этом, анализ таких данных может быть применен для повышения эффективности рекламных и маркетинговых компаний, оптимизации контента под социальные сети (SMO) и оптимизации взаимодействия с аудиторией сайтов.

Целью моей работы является создание прототипа системы для анализа социальной активности вокруг Интернет-новостей и блогов. Система собирает контент из определенных веб-сайтов, предоставляющих данные о количестве просмотров Интернет-страниц и реакции на эти страницы из социальных сетей. Собранные данные используются для анализа и выявления связей между содержанием и популярностью контента [1]. Полученные результаты предоставляются пользователю в удобном для восприятия виде.

Доклад будет посвящен обзору архитектуры системы и используемым методам анализа данных. Архитектура системы состоит из четырех основных частей: подсистемы сбора ссылок, подсистемы извлечения данных из полуструктурированных источников, подсистемы анализа данных и веб-интерфейса. Анализ полученных данных происходит в несколько этапов: преобразование текста в векторную модель [2]; уменьшения размерности модели с помощью статистических методов, SVD [3] и д.р.; построение модели для прогнозирования популярности контента с помощью различных методов машинного обучения; настройка и выбор моделей [3]. Также в докладе будут представлены результаты анализа таких сайтов, как: Business Insider, Forbes Blogs и LookAtMe.ru.

Литература

1. K. Lerman and T. Hogg, Using a Model of Social Dynamics to Predict Popularity of News // In Proceedings of 19th International World Wide Web Conference. 2010.
2. C. Manning, P. Raghavan and H. Schütze. Introduction to Information Retrieval. Cambridge University Press. 2008.
3. T. Hastie, R. Tibshirani and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. Springer. 2009.