Е.Б. Козеренко, Ю.И. Морозова, К.И. Кузнецов, М.М. Шарнин,
С.А. Ступников, Д.О. Брюхов, А.Е. Вовченко*

СОЗДАНИЕ СИСТЕМЫ МОНИТОРИНГА ИНТЕРНЕТ-ТЕКСТОВ ПО ТЕМЕ «СОЦИАЛЬНО-ПОЛИТИЧЕСКАЯ ЖИЗНЬ РЕГИОНОВ РОССИЙСКОЙ ФЕДЕРАЦИИ»

Аннотация. В статье описывается прототип системы мониторинга Интернет-текстов по теме «Социально-политическая жизнь регионов Российской Федерации». Основой системы является модуль лингвистической обработки текста, который производит многоуровневый анализ текстов: морфологический, синтаксический, семантический, тональный. Система работает в распределенном режиме на пяти компьютерах с использованием платформы IBM Hadoop. Результаты анализа текстов хранятся в единой базе данных. Средства работы с базами данных позволяют задавать поисковые запросы и представлять их результаты в виде графиков.

Ключевые слова: мониторинг социальных медиа, семантико-ориентированный лингвистический процессор, извлечение именованных сущностей, морфологический анализ, онтология предметной области, тональный анализ текстов, большие данные

---

* Козеренко Елена Борисовна – к. филол.н., kozerenko@mail.ru; Институт проблем информатики Российской академии наук, Москва.
Морозова Юлия Игоревна – yulia-ipi@yandex.ru; Институт проблем информатики Российской академии наук, Москва.
Кузнецов Константин Игоревич – k.smith@mail.ru; Институт проблем информатики Российской академии наук, Москва.
Шарнин Михаил Михайлович – к.т.н., mc@keywen.com; Институт проблем информатики Российской академии наук, Москва.
Ступников Сергей Александрович – к.т.н., ssa@ipi.ac.ru; Институт проблем информатики Российской академии наук, Москва.
Брюхов Дмитрий Олегович – к.т.н., brd@ipi.ac.ru; Институт проблем информатики Российской академии наук, Москва.
Вовченко Алексей Евгеньевич – к.т.н., itsnein@gmail.com; Институт проблем информатики Российской академии наук, Москва.

Анализ текстовой информации, представленной в Интернет, является необходимой составляющей профессиональной деятельности аналитиков, маркетологов, специалистов по связям с общественностью. Круг решаемых ими задач чрезвычайно разнообразен: мониторинг социальных медиа, продвижение в социальных медиа, управление репутацией в социальных медиа, клиентская поддержка в социальных сетях и др. В связи с этим существует общественная потребность в создании информационных систем, которые бы помогали специалистам решать указанные задачи. Системы мониторинга Интернет-текстов являются востребованными и активно разрабатываются во всем мире. Для английского языка разработано более 200 систем [1], для русского языка – около 30 систем [2].

В ИПИ РАН ведутся работы по созданию семантически-ориентированной системы мониторинга Интернет-текстов, относящихся к предметной области «Социально-политическая жизнь регионов Российской Федерации (на примере одного региона)». Функция данной системы – помогать заинтересованным лицам или организациям (например, органам власти) отслеживать социальное самочувствие населения и отдельных социальных групп и своевременно реагировать на возникающие социально-экономические проблемы. На текущий момент создан работающий прототип такой системы.

С точки зрения внутренней архитектуры, разрабатываемая система предназначена для решения следующих задач:

• извлечение данных из различных информационных ресурсов;

• смысловая обработка неструктурированных и слабоструктурированных данных (тексты, Веб и пр.);

• формирование структурированной базы данных;

• интеграция полученных в результате отбора и структуризации данных в массивно-параллельной архитектуре;

• реализация аналитических запросов над структурированной базой данных и визуализация результатов.

В качестве материала для проверки работы системы используются различные информационные ресурсы: архивы региональных электронных СМИ, социальные сети (на примере сайта «Вконтакте»), сервисы микроблогов (на примере сайта «Twitter»). Ресурсы существенным образом различаются по структуре, а также по характеру и лексике текстов. Таким образом, программные средства должны быть настроены на каждый тип источника по отдельности. Из всех материалов, доступных на указанных информационных ресурсах, для скачивания выбираются те материалы, которые имеют наибольшую ценность в рамках решаемой задачи: опубликованные недавно (например, в текущем году) авторами, проживающими в изучаемом регионе. Такие статьи и сообщения социальных сетей извлекаются из соответствующих ресурсов, загружаются в Hadoop [3] и обрабатываются с помощью средств лингвистического анализа, развернутых на каждом из узлов кластера.

Семантико-ориентированный лингвистический процессор [4] производит многоуровневую и многоплановую обработку текстов, в которую входит морфологический анализ, синтаксический анализ, семантический анализ с выделением именованных сущностей, тональный анализ, соотнесение выделенных сущностей с онтологической базой данных.

Статьи и сообщения, обогащенные информацией, извлеченной из текстов, преобразуются к виду, удовлетворяющему единой схеме хранилища. Данные, извлеченные из профилей пользователей социальных сетей, также помещаются в хранилище. Определение отношения населения к экономическим и политическим вопросам может быть осуществлено путем различных запросов к схеме хранилища.

Литература

1. Список систем мониторинга социальных медиа на английском языке. –URL: http://wiki.kenburbary.com/social-meda-monitoring-wiki (дата обращения: 2014-09-08).
2. Список систем мониторинга социальных медиа на русском языке. – URL: http://ipiranlogos.com/ru/sm-monitoring-systems/ (дата обращения: 2014-09-08).
3. Платформа IBM для больших данных. – URL: http://www-03.ibm.com/software/products/ru/category/bigdata (дата обращения: 2014-09-08).
4. *Kuznetsov I.P., Kozerenko E.B., Matskevich A.G.* Intelligent extraction of knowledge structures from natural language texts // Proceedings of the 2011 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology – Workshops, WI-IAT 2011, P. 269-272.

E.B. Kozerenko, Yu.I. Morozova, K.I. Kuznetsov, M.M. Sharnine, S.A. Stupnikov, D.O. Briukhov, A. E. Vovchenko

SOCIAL MONITORING SYSTEM IN THE SUBJECT DOMAIN "SOCIAL AND POLITICAL LIFE OF RUSSIAN FEDERATION REGIONS"

Abstract. The article describes the prototype of the social media monitoring system developed for the subject domain « Social and political life of Russian Federation regions ». The system is based on the module of linguistic processing which performs analysis of natural language texts at different levels: morphological analysis, syntactic analysis, semantic analysis and sentiment analysis. The systems functions in the distributed mode on five computers via the IBM Hadoop Platform. The analysis results are stored in a single data base.

Data base management tools make it possible to query the data base and visualize the queries results.

Keywords: social media monitoring, semantic linguistic processor, named entities recognition, morphological analysis, subject domain ontology, sentiment analysis, big data

Kozerenko Elena B. – PhD, kozerenko@mail.ru; Institute of Informatics Problems of the Russian Academy of Sciences, Moscow.

Morozova Yuliya I. – postgraduate, yulia-ipi@yandex.ru; Institute of Informatics Problems of the Russian Academy of Sciences, Moscow.

Kuznetsov Konstantin I. – k.smith@mail.ru; Institute of Informatics Problems of the Russian Academy of Sciences, Moscow.

Sharnine Mikhail M. – PhD, mc@keywen.com; Institute of Informatics Problems of the Russian Academy of Sciences, Moscow.

Stupnikov Sergey A. – PhD, ssa@ipi.ac.ru; Institute of Informatics Problems of the Russian Academy of Sciences, Moscow.

Briukhov Dmitry O. – PhD, brd@ipi.ac.ru; Institute of Informatics Problems of the Russian Academy of Sciences, Moscow.

Vovchenko Alexey E. – PhD, itsnein@gmail.com; Institute of Informatics Problems of the Russian Academy of Sciences, Moscow.