УДК 004.83

А.Г. Подвесовский, Д.В. Будыльский

ПРОБЛЕМЫ И ОСОБЕННОСТИ АВТОМАТИЗАЦИИ МОНИТОРИНГА СОЦИАЛЬНЫХ СЕТЕЙ И ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ПОЛЬЗОВАТЕЛЬСКИХ СООБЩЕНИЙ

Разработана модель мониторинга мнений пользователей социальных сетей с поддержкой интеллектуального анализа данных. Рассмотрен метод анализа тональности как один из подходов к интеллектуальной обработке текстовых сообщений пользователей социальной сети. Сформулированы требования к программному комплексу мониторинга социальных сетей и описаны результаты его практического применения, на основании чего определены направления дальнейших исследований.

Ключевые слова: социальная сеть, мониторинг, анализ тональности, нейронные сети, глубокое обучение.

В современном мире количество пользователей сети Интернет постоянно увеличивается, следовательно, увеличиваются объемы информации, передаваемой этими пользователями в сеть. В нашей стране, согласно данным ФОМ[1], доля активных пользователей сети Интернет (пользователей, выходящих в сеть хотя бы раз в сутки) летом 2014 года составила 50,1 % (58,4 млн человек). Годовой прирост числа пользователей российского сегмента Интернета, выходящих в сеть хотя бы раз за месяц, составил 9 %, а для суточной аудитории данный показатель равен 12 % [8]. При этом 63 % суточной интернет-аудитории пользуется Интернетом для общения в социальных сетях [7]. Таким образом, около 36,8 млн человек пользуются социальными сетями ежедневно.

Интерес к исследованию социальных сетей подтверждается множеством научных работ последних лет [2; 4-6]. В работе [6] отмечается, что социальные сети «помимо выполнения функций поддержки общения, обмена мнениями и получения информации их членами в последнее время все чаще становятся объектами и средствами информационного управления и ареной информационного противоборства». Крупные компании и политические партии активно используют социальные сети для мониторинга и анализа общественного мнения [14; 15; 17; 19]. При этом практически неисчерпаемым источником данных для мониторинга и анализа служит огромное, постоянно пополняющееся количество текстовых сообщений, оставляемых пользователями. В то же время объемы этих данных являются главным минусом – невозможно обработать столько информации вручную. Единственный выход – использование средств автоматизации, основанных на применении моделей и алгоритмов интеллектуальной обработки текстовой информации.

С учетом изложенного можно сделать вывод об актуальности автоматизации мониторинга мнений пользователей социальных сетей с поддержкой интеллектуальной обработки текстовых сообщений.

Модель мониторинга мнений пользователей социальных сетей. Под мониторингом мнений пользователей социальных сетей понимается непрерывное во времени отслеживание пользовательских публикаций, находящихся в открытом доступе. Публикации могут иметь различный формат: текст, изображения, аудио- и видеоматериалы. В настоящей статье ограничимся рассмотрением мониторинга текстовых сообщений.

За промежуток времени T в социальных сетях появляется M новых сообщений. Обозначим через F частоту поступления новых сообщений в социальных сетях:

$$F = \frac{M}{T}.$$

[1] Фонд «Общественное мнение» – российская организация, занимающаяся проведением социологических исследований. Создана в 1991 году в структуре ВЦИОМ (Всероссийский центр изучения общественного мнения).

На практике, ввиду ряда ограничений, обеспечить мониторинг социальной сети с заданной частотой F не представляется возможным. Количество сообщений m, которое реально извлечь за промежуток времени t, позволяет осуществить мониторинг с частотой

$$F' = \frac{m}{t},$$

где $F' \to F$, однако на практике величина F' значительно меньше F.

На рис. 1 в общем виде представлена схема мониторинга социальной сети с частотой F' сообщений в секунду. Алгоритм предполагает циклическое взаимодействие программного обеспечения мониторинга с социальной сетью посредством запроса на получение m последних сообщений социальной сети каждые t секунд. При этом необходимо сделать следующие важные замечания:

- данные, полученные в результате i-го запроса, не должны дублироваться в (i+1)-м запросе;
- $m_i \le m$, т.е. реальное количество сообщений, полученных из социальной сети, может быть меньше заданного (например, это может наблюдаться в периоды низкой активности пользователей социальных сетей).

Рис. 1. Общая схема мониторинга социальной сети с частотой F'

Множество w сообщений, полученных в результате очередного запроса, можно представить в виде объединения двух подмножеств:

$$w = w_{полез} \cup w_{шум},$$

где $w_{полез}$ – подмножество полезных сообщений (представляющих интерес в рамках задачи анализа); $w_{шум}$ – подмножество заведомо несодержательных сообщений (например, спам). В идеальном случае $w = w_{полез}$, т.е. при каждом запросе порции сообщений из социальных сетей ответ содержит только полезные сообщения.

Рассматривая задачу оптимизации выборки w (сокращения множества $w_{шум}$ вплоть до пустого), можно предложить два подхода к ее решению:

- оптимизация запросов к социальным сетям, в результате которой набор данных, извлекаемых из сети, должен содержать как можно меньше шумовых данных, что достигается заданием особых параметров при запросе;
- обработка ответа социальной сети, заведомо содержащего шумовые данные; в этом случае программное обеспечение берет на себя задачу отбраковки несодержательных сообщений (например, распознавание и удаление спама) и допускает к анализу только заведомо пригодные данные.

Первый подход сильно зависим от социальной сети, с которой приходится работать. Параметры запросов могут в значительной мере варьироваться от одной сети к другой, но в общем виде можно выделить два основных приема:

- задание подмножества подверженных мониторингу агентов социальной сети: $a \subset A$, где A – множество всех агентов;
- задание множества p_m параметров запроса на выборку сообщений, позволяющего, в зависимости от социальной сети, минимизировать количество шумовых сообщений.

Второй подход представляет собой классический этап преобразования данных в ETL-процессе[2], в частности фильтрацию данных. В основе фильтрации лежит использование набора условий f, которые играют роль фильтров, позволяющих оставлять в выборке одни данные и исключать другие. Зачастую фильтрация достаточно проста с вычислительной точки зрения, но в общем случае алгоритмы фильтрации могут отличаться по своей сложности. Простейший пример: фильтрации можно подвергнуть сообщения, количество слов в которых не превышает заданного. Более сложным примером может служить классификация сообщений по одному или нескольким критериям (анализ спама).

Обобщив изложенное и собрав все зависимости воедино, получим следующую модель мониторинга мнений пользователей социальных сетей S:

$$S = < w, m, t, a, p_m, f >,$$

где w – множество сообщений, извлекаемых из сети за период времени t; m – число таких сообщений (мощность множества w); a – множество агентов социальной сети, подверженных мониторингу ($a \subset A$); p_m – множество параметров запроса на выборку сообщений (секретные ключи, токены доступа и т.д.); f – множество фильтрационных условий (фильтр по количеству слов, фильтр спама и др.).

Мониторинг является первым этапом получения и частичной обработки данных о мнениях пользователей. Полученные данные необходимо сохранить, а затем подвергнуть анализу в соответствии с поступившим от пользователя запросом. Схематически данный процесс изображен на рис. 2.

Анализ тональности как метод интеллектуальной обработки пользовательских мнений. В литературе можно встретить разные подходы к формализации модели мнений. В англоязычных публикациях данную область исследований обычно называют opinion mining and sentiment analysis (дословно – «поиск мнений и анализ чувств»). В русскоязычных статьях обычно употребляется термин «анализ тональности». Несмотря на то что тональность является лишь одной из характеристик пользовательского мнения, именно задача классификации тональности наиболее часто рассматривается в настоящее время. Это можно объяснить следующими причинами.

Во-первых, определение автора и темы – гораздо более трудные задачи, чем классификация тональности, поэтому имеет смысл сначала решить более простую задачу, а затем уже переключиться на остальные.

Во-вторых, во многих случаях достаточно определить лишь тональность сообщения, поскольку другие его характеристики уже известны. Так, при анализе мнений пользователей социальной сети не требуется определять авторов сообщений. Кроме того, зачастую бывает известна тема. Например, при анализе сообщений, содержащих название определенного бренда, требуется лишь определить тональность отношения пользователя к этому бренду.

[2] ETL (от англ. extraction, transformation, loading – извлечение, преобразование, загрузка) – комплекс методов, реализующих процесс переноса исходных данных из различных источников в аналитическое приложение или поддерживающее его хранилище данных [11].

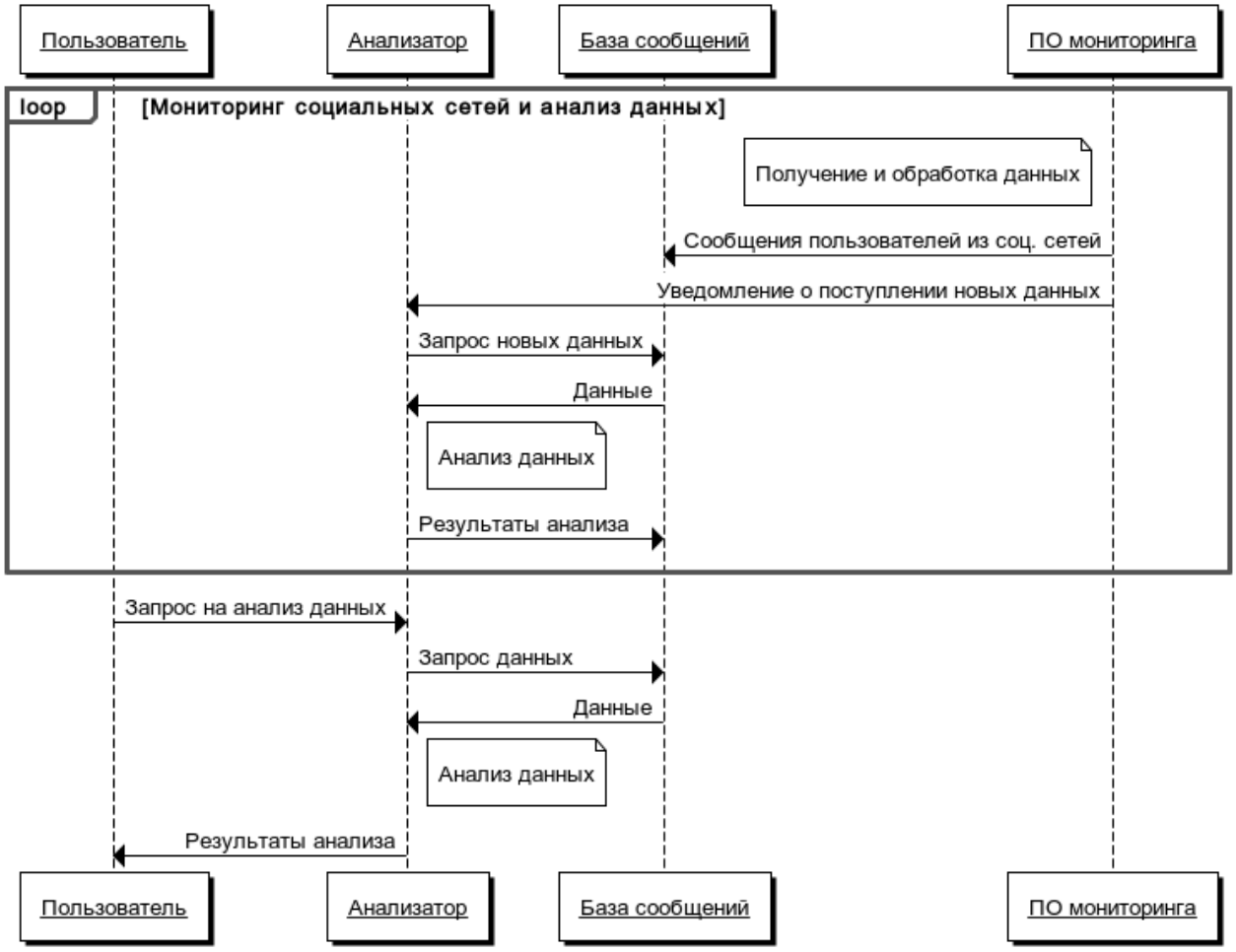

Рис. 2. Схема процесса мониторинга и анализа сообщений в социальных сетях

В общем случае под анализом тональности текста понимается класс методов, предназначенных для выявления эмоций в тексте. Он позволяет классифицировать текст по его эмоциональной окраске, например охарактеризовать текст как положительный, отрицательный либо нейтральный. Кроме того, возможно определение силы тональности, субъекта/объекта тональности и ряда других характеристик текста [9].

В работе [10] предложена следующая классификация методов анализа тональности: методы, основанные на правилах; методы, основанные на словарях; машинное обучение с учителем; машинное обучение без учителя.

Методы, основанные на правилах, заключаются в применении тех или иных заранее заданных человеком правил, на основе которых делается общий вывод о тональности текстового фрагмента. Такие методы считаются наиболее точными при определении тональности и поэтому применяются в большинстве коммерческих систем. Пример сравнения методов, основанных на правилах, с другими приведен в работе [3]. Автор делает вывод о том, что другие виды методов (не основанные на правилах) используют теоретико-множественную модель текстов, в которой не учитывается контекст употребления слов.

К недостаткам данной группы методов можно отнести высокие трудозатраты, поскольку для получения хороших результатов экспертам необходимо составить значительный объем правил для обработки конструкций естественным языком.

Подходы, основанные на словарях, используют для анализа текста так называемые тональные словари (affective lexicons). В простейшем случае тональный словарь представляет собой список слов со значением тональности для каждого слова. Процесс анализа текста можно описать следующим алгоритмом: каждому слову в тексте присвоить его

значение тональности из словаря, а затем вычислить общую тональность текста (например, как среднее арифметическое значений тональности всех содержащихся в нем слов).

В чистом виде такие методы не применяются, но словари присутствуют в составе систем, использующих методы, основанные на правилах, или методы машинного обучения [12; 18].

Машинное обучение без учителя представляет собой наименее точный метод анализа тональности. Примером данного подхода может служить автоматическая кластеризация документов. В современных работах подобные методы не встречаются.

Последняя группа методов анализа тональности – машинное обучение с учителем. Фактически она представляет собой совокупность методов классификации, в которых заранее задан набор классов, к которым необходимо отнести документ, например «положительная тональность», «отрицательная тональность», «нейтральная тональность».

Требования к программному комплексу мониторинга социальных сетей и интеллектуальной обработки пользовательских сообщений. Основными требованиями, предъявляемыми к программному комплексу мониторинга социальных сетей с поддержкой анализа тональности текстовых сообщений, являются:

- периодический опрос социальных сетей для выборки новых текстовых сообщений, оставленных пользователями;
- сохранение полученной информации в базе данных;
- обработка информации – определение тональности текстовых сообщений;
- поддержка пользовательских запросов к данным.

Общая схема такого программного комплекса представлена на рис. 3.

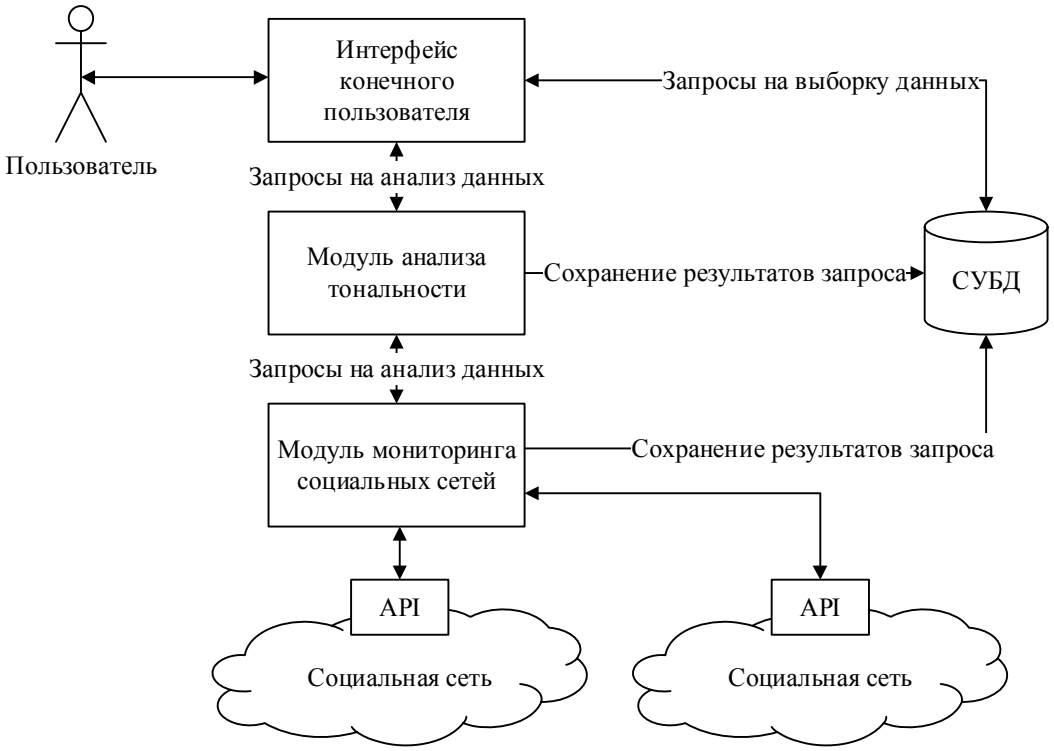

Рис. 3. Общая схема программного комплекса мониторинга социальных сетей с поддержкой анализа тональности текстовых сообщений

Рассмотрим более подробно назначение составляющих программного комплекса.

- СУБД (система управления базами данных) – зависимый модуль, основной задачей которого является хранение данных и доступ к ним. Зависимость модуля объясняется тем, что данные поступают и извлекаются в ходе работы других модулей системы.

- Модуль анализа тональности – зависимый модуль, основной задачей которого является проведение анализа тональности над текстовыми данными и сохранение результатов в СУБД (при необходимости). Зависимость модуля объясняется тем, что его функционирование определяется другими модулями.
- Модуль мониторинга социальных сетей – независимый модуль, задачей которого является периодическое опрашивание социальных сетей при помощи запросов на извлечение текстовых сообщений и последующее сохранение полученных данных в СУБД (при необходимости). Независимость модуля означает, что его работа автоматизирована и не определяется работой других модулей.
- Интерфейс конечного пользователя – независимый модуль, предоставляющий пользователю доступ к анализатору тональности и данным в СУБД в удобном виде. Независимость модуля также связана с тем, что работа данного модуля определяется только пользователем, а не другими модулями системы.

Связи между модулями отображены на схеме одно- и двунаправленными стрелками, обозначающими направление передачи данных. Например, модуль мониторинга передает данные в СУБД и может отправлять команды анализатору, не получая от других модулей каких-либо данных или команд; интерфейс пользователя отправляет запросы к СУБД и анализатору и получает в ответ наборы данных [13].

Результаты практической реализации и тестирования программного комплекса. В ходе проведения исследований был реализован программный комплекс, осуществляющий мониторинг сообщений в социальной сети Twitter. Анализ тональности выполнялся с использованием нейронной сети, представляющей собой двухслойный персептрон. В пределах заданной тематики спортивных новостей удалось достичь точности классификации сообщений 70 %, что является приемлемым результатом.

Применение программного комплекса позволило сделать ряд выводов и определить направления дальнейших исследований.

Прежде всего, для корректной работы алгоритма машинного обучения с учителем необходимо иметь заранее подготовленный набор данных для обучения. В ходе исследования была сделана выборка сообщений из социальной сети. Эти сообщения были оценены экспертами, и каждому сообщению была присвоена оценка тональности – «негативная», «нейтральная» или «позитивная». При этом разные эксперты могли по-разному оценивать одно и то же сообщение. Более того, один и тот же эксперт мог в процессе мониторинга изменять свое суждение по поводу тональности определенного сообщения. Неопределенность возникала при установлении степени позитивности или негативности, поэтому сообщениям присваивался либо тональный, либо нейтральный класс. Таким образом, возникает задача, связанная с обработкой нечеткой информации [1] и введением нечетких оценок тональности сообщений. Технически это означает, что нейронную сеть, в настоящий момент представляющую собой тернарный классификатор, необходимо модифицировать одним из двух следующих способов:

- расширить количество классов, обеспечив возможность использования нечетких оценок тональности, таких как «сильно негативная», «умеренно негативная», «слабо негативная» и т.д.;
- обеспечить регрессионные значения на выходе нейронной сети, которые можно интерпретировать как степень принадлежности сообщения к позитивным или негативным.

Развитие архитектур нейронных сетей и методов глубокого обучения (deep learning) [16] обеспечивает потенциальную возможность значительного улучшения качества распознавания тональности текста по сравнению с классической персептронной архитектурой. Следовательно, необходимо провести исследование и сравнить результаты работы классификатора при различных конфигурациях архитектуры нейронной сети.

Первоначально тестирование программного комплекса проводилось на наборе сообщений спортивной тематики, что позволяло заведомо исключить двусмысленность некоторых терминов. В связи с этим важной задачей является исследование алгоритмов глубокого обучения при работе с сообщениями произвольной тематики. В частности, интерес представляет способность данных алгоритмов автоматически решать проблему омонимии для текстовых сообщений на русском языке.

СПИСОК ЛИТЕРАТУРЫ

1. Аверченков, В.И. Представление и обработка нечеткой информации в многокритериальных моделях принятия решений для задач управления социально-экономическими системами / В.И. Аверченков, А.В. Лагерев, А.Г. Подвесовский // Вестн. Брян. гос. техн. ун-та. – 2012. – № 2 (34). – С. 97-104.
2. Базенков, Н.И. Обзор информационных систем анализа социальных сетей / Н.И. Базенков, Д.А. Губанов // Управление большими системами: сб. тр. – М.: ИПУ РАН, 2013. – № 41. – С. 357-394.
3. Васильев, В.Г. Классификация отзывов пользователей с использованием фрагментных правил / В.Г. Васильев, М.В. Худякова, С.А. Давыдов // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог»: сб. ст. – М.: Изд-во РГТУ, 2011. – Т. 2. – С. 66-76.
4. Губанов, Д.А. Концептуальный подход к анализу онлайновых социальных сетей / Д.А. Губанов, А.Г. Чхартишвили // Управление большими системами: сб. тр. – М.: ИПУ РАН, 2013. – № 45. – С. 226-236.
5. Губанов, Д.А. Модели информационного влияния и информационного управления в социальных сетях / Д.А. Губанов, Д.А. Новиков, А.Г. Чхартишвили // Проблемы управления. – М.: СенСиДат-Контрол, 2009. – № 5. – С. 28-35.
6. Губанов, Д.А. Социальные сети: модели информационного влияния, управления и противоборства / Д.А. Губанов, Д.А. Новиков, А.Г. Чхартишвили. – М.: Изд-во физ.-мат. лит., 2010. – 228 с.
7. Для чего люди используют интернет? – URL: http://fom.ru/SMI-i-internet/11088.
8. Интернет в России: динамика проникновения. Лето 2014. – URL: http://fom.ru/SMI-i-internet/11740.
9. Меньшиков, И.Л. Анализ тональности текста на русском языке при помощи графовых моделей / И.Л. Меньшиков // Доклады Всероссийской научной конференции АИСТ’2013: сб. ст. – Екатеринбург, 2013. – С. 151-155.
10. Пак, А. Обучаем компьютер чувствам (sentiment analysis по-русски). – URL: http://habrahabr.ru/post/149605.
11. Паклин, Н.Б. Бизнес-аналитика: от данных к знаниям: учеб. пособие / Н.Б. Паклин, В.И. Орешков. – 2-е изд. – СПб.: Питер, 2013. – 704 с.
12. Паничева, П.В. Система сентиментного анализа ATEX, основанная на правилах, при обработке текстов различных тематик / П.В. Паничева // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог»: сб. ст. – М.: Изд-во РГТУ, 2013. – Т. 2. – С. 101-112.
13. Подвесовский, А.Г. Особенности реализации программного комплекса мониторинга социальных сетей с поддержкой анализа тональности текстовых сообщений / А.Г. Подвесовский, Д.В. Будыльский // Вопросы информационных технологий: междунар. сб. науч. ст. – Липецк: Гравис, 2014. – Вып. I. – С. 23-33.
14. Brand Analytics – система мониторинга и анализа социальных медиа. – URL: http://br-analytics.ru.
15. BrandSpotter – система мониторинга социальных медиа. – URL: http://brandspotter.ru.
16. Deng, L. Deep Learning: Methods and Applications / L. Deng, Y. Dong. – Now Publishers, 2014. – 134 p.
17. IQBuzz – профессиональный сервис мониторинга для маркетологов, PR и SMM. – URL: http://iqbuzz.ru.
18. Socher, R. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank / R. Socher [et al.] // EMNLP. – 2013. – P. 1-12.
19. YouScan.ru – система мониторинга социальных медиа и социальных сетей. – URL: http://youscan.ru.

Материал поступил в редколлегию 10.11.14.