УДК 004.777-047.36

УЧЕБНО-ИССЛЕДОВАТЕЛЬСКАЯ СИСТЕМА МОНИТОРИНГА И АНАЛИЗА ВЕБ-ПРОСТРАНСТВА

М.В.Стержанов
Доцент кафедры информатики БГУИР, кандидат технических наук, доцент

М.А. Медунецкий
Студент БГУИР

М.П. Хоронеко
Студент БГУИР

А.С. Дроздов
Студент БГУИР

Н.В. Харитонов
Студент БГУИР

Е.И. Нехведович
Студент БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: sterjanov@bsuir.by

Аннотация. В данной работе описывается учебно-исследовательская система веб-мониторинга и анализа данных. Описаны цели и задачи разработки. Представлены фазы исследований.

Ключевые слова: анализ данных, тематическое моделирование, учебно-исследовательская система.

Введение. Получение ценной информации через автоматизированный анализ больших объемов данных сегодня является крайне популярным направлением и тесно сопряжено с такими терминами, как Big Data («большие данные») и Data Mining (в русском языке нет устоявшегося перевода для данного термина, но принято понимать его как «интеллектуальный анализ данных»).

Целью учебно-исследовательской системы мониторинга и анализа веб-пространства является разработка программного обеспечения для выявления контента, нежелательного для потребления детьми и подростками и защиты несовершеннолетних от информации, причиняющей вред здоровью и (или) развитию, побуждающей к самоубийству.

Дети в возрасте до 18 лет являются активными пользователями сети интернет, проводя время обучаясь, играя, коммуницируя со сверстниками. Пользование социальными сетями, программами обмена сообщениями, блогами и форумами позволяет детям активно распространять и обмениваться информацией.

На данный момент в русскоязычном интернете существует сайты и сообщества, пропагандирующие субкультуру, связанную с дискредитацией общепринятых ценностей и норм поведения в обществе, а также пропагандой деструктивных форм поведения.

Навязанное поведение и негативное мышление зачастую являются губительными для подростка, поскольку он независимо от своего желания начинает вести асоциальный образ жизни, выражающийся в отказе от учебы, нормальных отношений со сверстниками и родными.

Информационная безопасность детей – состояние защищенности детей, при котором отсутствует риск, связанный с причинением информацией вреда их здоровью и (или) физическому, психическому, духовному, нравственному развитию.

В целях предотвращения вовлечения детей и подростков в просмотр контента вредоносного характера, а также оповещения родителей и педагогов о возможности депрессивных состояний детей в необходимо программно анализировать качество потребляемого подростком контента.

Данная задача является актуальной не только в Республике Беларусь, но и за ее пределами, так как на данный момент отсутствуют системы анализа и фильтрации нежелательного контента глобального уровня. Существующая база fortiguard.com направлена на анализ и

фильтрацию зарубежных ресурсов, она не способна классифицировать русскоязычный контент, содержащий явную или скрытую пропаганду алкоголя, наркотиков, асоциального образа жизни.

Тема разработки является современной и актуальной, она соответствует Перечню приоритетных направлений научных исследований Республики Беларусь на 2016–2020 гг., в частности, направлениям «00005 Информатика и космические исследования» и «00013 Безопасность человека, общества и государства».

В процессе исследований и разработок нами решаются следующие задачи:

1. Исследование методов и средств хранения, организации и анализа больших массивов данных.
2. Создание математической модели информационного мониторинга.
3. Разработка алгоритмов анализа тональности текстов на русском языке.
4. Разработка и усовершенствование алгоритмов тематического моделирования.

Ведется создание универсальной системы автоматизированного мониторинга, которая позволит организовать данные для их анализа на основе технологий машинного обучения и провести собственно анализ этих данных. Архитектура программного комплекса состоит из нескольких логически обособленных уровней:

1. Автоматизированный краулер веб-ресурсов;
2. Система обработки текстовых данных;
3. Система анализа полученных данных.

При обработке данных больших объемов разумным является применение распределенных вычислений, к примеру, использование парадигмы MapReduce.

Приведем основные этапы исследования:

1. Построение высокопроизводительных модулей сбора информации (краулеров) из различных источников (сайты, блоги, тематические порталы, открытый API).
2. Организация подсистемы хранения данных с использованием трех шагов обработки данных: консолидация, трансформация и очистка.
3. Исследование методов морфоанализа, синтаксического анализа.
4. Исследование существующих методов вероятностного тематического моделирования и методов оценки их качества. Реализация алгоритма построения тематической модели. Оценка качества полученной тематической модели.
5. Исследование и реализация основных алгоритмов машинного обучения, используемых для решения задачи автоматического определения тональности. Исследование качества работы алгоритмов.
6. Исследование методов семантического анализа. Выделение именованных объектов и их атрибутов.
7. Реализация модулей комплексной бизнес-аналитики в режиме реального времени с развитой системой BI-отчетности и наглядной визуализацией, объединяющей обработку и анализ структурированных и неструктурированных данных.
8. Тестирование созданной системы на больших массивах данных.

В результате проведения исследований предполагается:

- разработать модуль эффективного получения данных из веб-пространства.
- разработать алгоритм вероятностного тематического моделирования.
- разработать алгоритм анализа тональности русского языка.

В рамках предыдущих исследований нами разработан метод получения данных из открытых источников [1], получены результаты в области анализа тональности текста на русском языке, использованы методы машинного обучения для классификации текстов.

Построение тематической модели по коллекции документов является некорректно поставленной оптимизационной задачей, которая может иметь бесконечное множество решений. Одним из недостатков тематических моделей является использование модели «мешка слов»,

в которой каждый документ представляется в виде множества несвязанных между собой слов. Данная модель не учитывает порядок слов и основывается на гипотезе независимости появлений слов друг от друга в текстах. Нами планируется решение задачи улучшения качества тематических моделей за счет добавления в них подходящих словосочетаний и многословных выражений (в частности, терминов) и учета связей между ними и образующими их словами.

Современные методы построения тематических моделей продуцируют довольно большой список слов и словосочетаний, являющихся лишь кандидатами в термины, которые затем должны быть проанализированы и подтверждены экспертами по предметной области текстов. Нами планируется выделение максимального числа признаков, описывающих кандидаты с разных сторон, и проведение в ходе извлечения ранжирования кандидатов, чтобы в начале итогового списка стояли слова и словосочетания, действительно являющиеся терминами. Планируется исследование характеристик тематической информации на качество извлечения терминов.

Тематическое моделирование — это одно из современных направлений статистического анализа текстов, один из основных способов извлечения знаний из текстовых коллекций. Вероятностное тематическое моделирование — это набор алгоритмов, позволяющих анализировать слова в больших наборах документов и извлекать из них темы, связи между темами и изменение их во времени [3]. Вероятностная тематическая модель (BTM) описывает каждую тему дискретным распределением на множестве терминов, а каждый документ — дискретным распределением на множестве тем.

Пусть $D = \{d1 \dots d\ |D|\}$ – множество документов, $C = \{c1 \dots c\ |C|\ \}$–множество тематик, Ф: $D \times C \rightarrow \{0,1\}$ – неизвестная целевая функция, которая для пары (di, cj) говорит, принадлежит ли документ di тематике cj. Задача состоит в построении функции Ф′, максимально близкой к Ф.

В качестве тематических моделей использованы PLSA и LDA модели. PLSA [4] позволяет строить тематические модели с любым числом матриц Ф (тематическое распределение слов). Поэтому она может решать задачи классификации и категоризации документов, строить мультиязычные модели. Мультимодальные модели позволяют эффективно использовать имеющиеся в документе мета-данные (авторы, дата издания, теги и т.п.). Скрытое размещение Дирихле (LDA) — генеративная графическая вероятностная модель, предложенная Дэвидом Блеем и соавторами в 2003 году [5]. Латентное размещение Дирихле является базовой вероятностной тематической моделью и из-за большого количества приложений и обобщений является самой распространенной вероятностной тематической моделью. Для идентификации параметров модели LDA по коллекции документов можно применить семплирование по Гиббсу — алгоритм для генерации выборки совместного распределения множества случайных величин. Он используется для оценки совместного распределения и для вычисления интегралов методом Монте-Карло.

Алгоритм (TF-IDF) основывается на учете частоты слов. При этом для того, чтобы не учитывать слова которые встречаются во всех документах используется обратная частота документа IDF, которая равна логарифму отношения количества документов к количеству документов в которых они встретились. Частота слова рассматривается относительно конкретного документа.

В результате эксперимента было получено порядка полумиллиона текстовых документов на русском и белорусском языках, которые были разбиты на 150 тем.

Планируется продолжать улучшать метод выделения тематического ядра, т.к. замечены существенные недостатки текущего метода на сайтах с политематичным контентом. Значительным улучшением качества построения тематической модели будет являться увеличение качества алгоритмов лемматизации текста на белорусском языке

Таким образом, реализация системы может послужить решением важной теоретико-прикладной задачи автоматизации первичного анализа и структуризации инфопространства..

Список литературы

[1]. Стержанов М.В. Модуль получения данных из внешних открытых источников / М.В. Стержанов и др. Третья Международная научно-практическая конференция «BIG DATA and Advanced Analytics. BIG DATA и анализ высокого уровня» (Минск, Республика Беларусь, 3-4 мая 2017 года) – Минск: БГУИР, 2017. – с. 174-177.

[2]. М.И. Селюк Использование вероятностного латентнтно- семантического анализа для построения вероятностных тематических моделей текстовых коллекций. М.И. Селюк, Н.Н. Шинкевич, М.В. Стержанов / Технические средства защиты информации: Тезисы докладов XV Белорусско-российской научно технической конференции, 6 июня 2017 г., Минск: БГУИР, 2017. - с.101.

[3]. Blei D. M. Probabilistic topic models //Communications of the ACM. – 2012. – Т. 55. – No. 4. – P. 77-84.

[4]. Hofmann T. Probabilistic Latent Semantic Indexing // Proceedings of the 22nd Annual Internationl SIGIR Conference on Research and Development in Information Retrieval. – 1999. – P. 50-57.

[5]. Blei D. M. Latent dirichlet allocation / D. M. Blei, A. Y. Ng, M. I. Jordan//the Journal of machine Learning research. – 2003. – Т. 3. – С. 993-1022.

EDUCATIONAL-RESEARCH SYSTEM OF MONITORING AND ANALYSIS OF WEB-SPACE

M.STERJANOV , PhD
Associate Professor, Department of Computer Science BSUIR

M.MEDUNETSKI
Student of BSUIR

M.HORONEKO
Student of BSUIR

A. DROZDOV
Student of BSUIR

N. KHARITONAU
Student of BSUIR

E. NEHVEDOVICH
Student of BSUIR

Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
E-mail: sterjanov@bsuir.by

Abstract. This work describes educational research system for web monitoring and data analysis. Goals and tasks of developing are described. Research phases are provided.

Key words: data analysis, topic modeling, educational research system.