УДК 004.62
DOI 10.21685/2072-3059-2017-2-2

А. С. Бождай, А. Ю. Тимонин

ИССЛЕДОВАНИЕ ПРОЦЕССА АНАЛИЗА ТЕКСТОВЫХ И МУЛЬТИМЕДИА ДАННЫХ СОЦИАЛЬНОГО ПРОФИЛЯ ИЗ ОТКРЫТЫХ ИСТОЧНИКОВ ИНФОРМАЦИИ

Аннотация.

Актуальность и цели. Наибольший научный интерес для аналитиков представляют открытые социальные данные из сети Интернет как имеющие непосредственную связь со всеми видами человеческой деятельности. Однако в своем исходном виде эти данные плохо подходят для автоматизированной прикладной обработки и должны быть представлены в структурированном, удобном для человеческого восприятия виде – социальном профиле. Построение социального профиля осуществляется через анализ отфильтрованных исходных данных из открытых источников сети Интернет. Динамические неструктурированные данные, включающие в себя как текстовую, так и мультимедиа информацию, не могут быть обработаны классическими средствами аналитики. Поэтому необходимо определить новые методы и подходы анализа в зависимости от типа рассматриваемой информации для наиболее эффективного и полного использования исходных данных.

Материалы и методы. Задача анализа данных социального профиля человека достигается за счет использования математического аппарата теории множеств, программных комплексов Big Data и NoSQL хранилищ данных, средств аналитики социальных медиа, а также современных методов анализа мультимедиа.

Результаты. Предложено разделить исходные данные социального профиля на статическую и динамическую части. В статье рассмотрены методы анализа неструктурированной текстовой информации социального профиля. Описывается технология поиска неявных зависимостей в текстах с использованием средств визуального анализа и обработки естественного языка. Также предлагается обзор методик для анализа мультимедиа контента (графика, звук).

Выводы. Этап анализа текстовых и мультимедийных данных социального профиля является наиболее важным с точки зрения получаемых результатов и достаточно сложным в реализации. Существует возможность частично автоматизировать процесс анализа информации за счет использования средств визуального анализа, обработки естественного языка (NLP), нейронных сетей и специализированных алгоритмов. Полученные данные обеспечивают подробный детальный обзор связей и сущностей социального профиля и могут использоваться в дальнейших более глубоких социальных исследованиях.

Ключевые слова: анализ данных, неструктурированные данные, мультимедиа, открытые источники информации, социальный профиль человека, Big Data.

A. S. Bozhday, A. Yu. Timonin

RESEARCHING THE PROCESS OF SOCIAL MEDIA AND TEXTUAL DATA ANALYZING FROM OPEN-ACCESS DATA SOURCES

Abstract.

Background. The greatest scientific analytical interest is drawn by open access social online data, as they are directly linked to all kinds of human activity. However, the initial form of such data is rather unsuitable for automated applied processing and should be presented in a structured, convenient, human-readable form – a social profile. Social profile building is carried out through analyzing filtered initial online data from open sources. Dynamically unstructured data, including textual and multimedia information, cannot be handled by classical analytic means. It is necessary to define new analytical methods and approaches depending on types of information for the most effective and full use of initial data.

Materials and methods. The task of personal social profile data analysis is achieved through the use of mathematical tools of the set theory, Big Data software and NoSQL data storages, analytic tools for social media, as well as modern methods for analyzing multimedia data.

Results. It is suggested to divide initial social profile data into static and dynamic parts. The article considers methods of unstructured textual social profile data analysis, describes a technology of searching implicit dependences in texts using visual analysis and natural language processing means, as well as offers a review of techniques for analyzing multimedia content (graphics, sound).

Conlusions. The stageof textual and multimedia social profile data analysis is the most important in terms of results and quite complicated to implement. There is a possibility to partially automate the process of information analyzing through the use of visual analysis, natural language processing (NLP), neural networks and specialized algorithms. Thedata obtained provide a detailed in-depth review of social profile entities and their relations. It can be used for further deeper social researches.

Key words: Big Data, Data Mining, data analysis, multimedia, personal social profile, public data sources, unstructured data.

Введение

Социальный профиль [1] – это множество информации, характеризующее социальные свойства человека и наглядно структурированное для удобства автоматизированной обработки и человеческого восприятия. Социальные профили могут найти свое применение в различных сферах деятельности, начиная с прикладных целей контекстной передачи информации и заканчивая исследованиями искусственного интеллекта и социума, а также противодействия терроризму. Задача построения социального профиля первоначально сводится к созданию математической модели и выбору структуры данных для хранения персонализированной информации [1]. Социальный профиль человека основывается на данных из открытых источников сети Интернет. Идентификация человека в сети производится через определение его точек вхождения – учетных записей веб-ресурсов, которые предоставляют ему ряд возможностей по использованию ресурса и выделяют его из массы остальных пользователей сети. Собираемые данные фильтруются от посторонней информации и разделяются по степени структурированности на статическую и динамическую части. С учетом экспоненциального роста информации в сети перед аналитиками стоит задача автоматизировать процессы анализа как структурированной, так и неструктурированной информации. В данной работе поднимаются вопросы анализа текстовых и мультимедийных данных социального профиля человека с использованием различных средств (Big Data, OCR, визуальный анализ, статистика и т.д.).

После идентификации человека в сети Интернет, сбора и фильтрации первичных данных социального профиля [1, 2] наступает черед для анализа полученной информации с целью формирования целостной структурированной социальной картины личности. Собранная ранее информация разделяется на две логические части [1]: информационную карту, содержащую уникальные идентифицирующие сведения о рассматриваемой персоне (статический контент), и динамический контент, состоящий из гетерогенных неструктурированных данных. Эти данные хранятся в различных типах хранилищ, зависящих от характера самой информации. Для хранения данных, не предусматривающих редактирования, предлагается использовать нереляционную распределенную базу данных с открытым исходным кодом HBase, которая запускается над распределенной файловой системой HDFS (Hadoop Distributed File System) и обеспечивает надежный способ хранения очень больших объемов разнородных данных. Соответствующая геоинформация помещается в графовое хранилище Neo4J [3], основными характеристиками которого являются: поддержка принципов ACID (атомарность, согласованность, изолированность, долговечность данных; англ. Atomicity, Consistency, Isolation, Durability), хорошая масштабируемость, поддержка популярных языков программирования (Java, Python, Ruby), мощный механизм обхода графа, подробная документация. Кроме того, данные динамической части могут быть как текстовыми, так и мультимедийными. Поэтому для анализа каждого типа данных требуется отдельный подход. Отдельной задачей при этом является анализ неструктурированных данных на предмет наличия неявных семантических зависимостей, искажений, информационного мусора, которая может быть решена при помощи сочетания встроенных механизмов поискового робота и прикладных программных модулей на основе фреймворка для разработки и выполнения распределенных программ Hadoop [2].

1. Анализ текстовых данных социального профиля

В настоящее время для обработки неструктурированных текстовых данных аналитиками применяются методы интеллектуального анализа текстов и обработки естественного языка. Также могут использоваться методы анализа тональности для определения эмоциональной окраски текстов и возможного выявления в них неявного или скрытого смысла [4].

Неструктурированный текст, являющийся исходным при построении социального профиля, может включать: электронные письма, публикации (посты, комментарии на социальных ресурсах), электронные документы, списки, расшифровки изображений и аудио записей, геоданные и т.д. В подсистеме анализа социальных связей и зависимостей осуществляется поиск и разделение по группам выражений, которые станут основой для узлов и связей социального графа. Анализ текстов состоит из четырех этапов, представленных на рис. 1.

Первым шагом анализа неструктурированных текстовых данных социального профиля является определение исходного языка для каждого элемента. Затем в текстах осуществляется поиск записей, входящих в информационную карту. Это обеспечит основу для последующего выстраивания связей социального профиля.

Дальнейший этап анализа заключается в распознавании именованных сущностей и извлечении признаков. Полученные результаты являются кан-

дидатами на роль объектов социального профиля. Необходимо учитывать наличие синонимов, производных слов, транскрипций для каждого объекта.

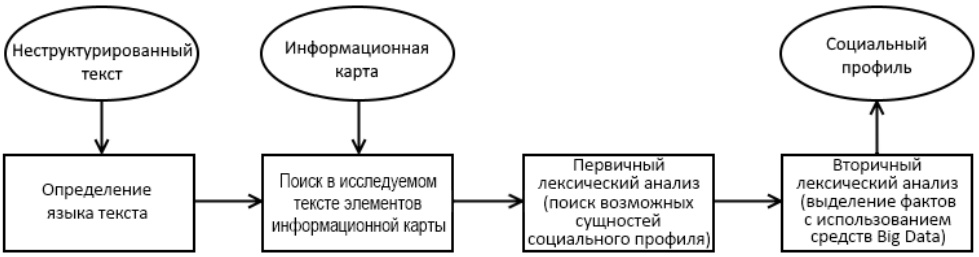

Рис. 1. Последовательность шагов анализа неструктурированного текста

После получения коллекции объектов социального профиля начинается извлечение информации и связей из неструктурированных данных для формирования целостной картины социального профиля. На данном этапе необходимо наличие средств обработки естественного языка, в частности, заполненных тезаурусов правил синтаксического и лексического анализа текста. Желательно наличие отдельного словаря для оценки настроений.

Ввиду очень большого количества обрабатываемых данных предлагается использовать решения на основе технологии Big Data. В качестве примера в данной работе рассмотрим применение программного средства IBM Content Analytics. Оно осуществляет поиск фактов на основе анализа контента, просмотра и импорта содержимого, синтаксический разбор и анализ содержимого, моделирование и прогнозирование, разработку интеллектуальных фильтров и создание пригодного для поиска индекса [5]. Рассмотрим алгоритм работы с Content Analytics (рис. 2).

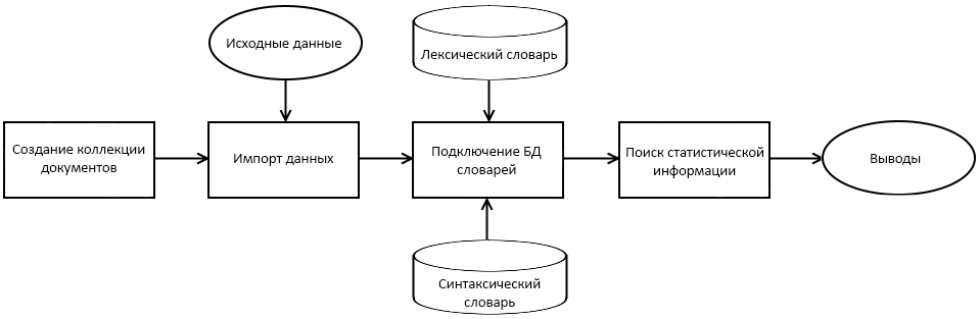

Рис. 2. Схема алгоритма анализа текстов в IBM Content Analytics

Первоначально создается коллекция документов – специальная индексируемая структура для работы со срезами данных или ограниченными совокупностями однородных значений по некоторому классификационному признаку. В нее импортируются исходные данные из различных источников. Далее подключаются базы данных (БД) словаря и синтаксических правил.

Затем поиск статистической информации осуществляется либо с помощью баз данных словарей обработки естественного языка, либо посредством запросов из ключевых слов. Тезаурусы состоят из локальной БД, xml-файла и

словаря в формате dict. Эти словари заполняются наиболее характерными выражениями из текстов исходных данных, после чего из них лексическим анализатором получают другие формы слов. Это повышает качество анализа и способствует нахождению информации, на основе которой делаются статистические выводы.

Результаты выделяются в тексте во вкладке Documents. Вкладка Facets показывает статистические данные результатов (например, количество повторений в тексте) в виде диаграмм. Пример диаграммы среза по ключевому слову «Возраст» показан на рис. 3.

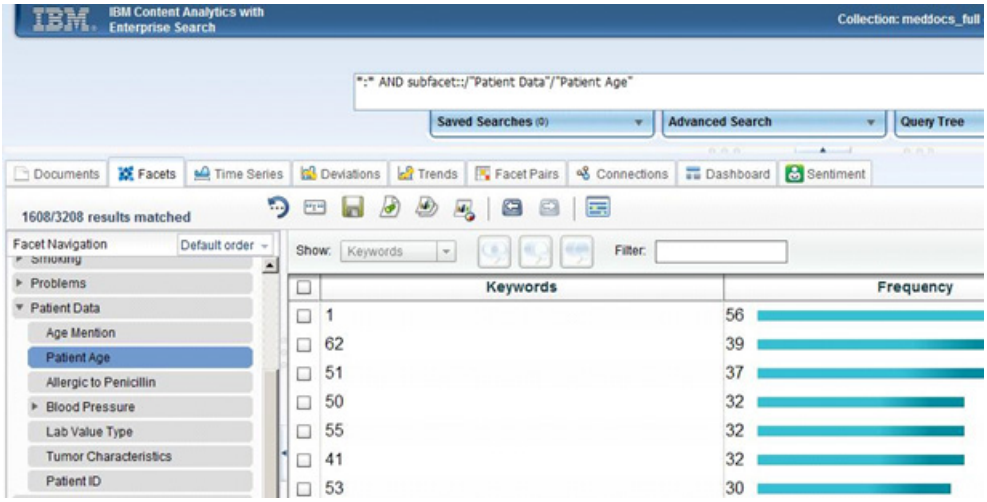

Рис. 3. Пример запроса для среза данных в Content Analytics [5]

Также возможно применение системы IBM BigInsights, использующей фреймворк Hadoop и модель распределенных вычислений MapReduce для выполнения задачи текстового анализа [6]. Встроенные функции системы включают нормализацию, разметку, идентификацию языка, классификацию текстов для фильтрации спама, распознавание и интеграцию сущностей, анализ настроений, извлечение связей. Результаты работы BigInsights можно использовать как исходные для других внешних обработчиков, в частности, для определения неявных связей внутри данных социального профиля.

2. Выявление неявных зависимостей в данных социального профиля

Результаты анализа неструктурированных данных даже с использованием специализированных средств NLP (обработка естественного языка, англ.: *natural language processing*), могут быть несовершенными: используемые тезаурусы могут быть неполными, исходные данные содержать лингвистические ошибки, иметь двоякий или скрытый смысл. Для разрешения подобных вопросов наиболее подходят средства визуального анализа, в качестве которых предлагается использовать программный инструментарий IBM i2. Он включает в себя следующие компоненты [7]:

– Text Chart – модуль для визуализации неструктурированных текстов. Он позволяет оперативно преобразовывать текстовую информацию в структурированный и легко анализируемый графический формат.

– Analyst's Notebook – предоставляет возможности быстрого сопоставления, анализа и наглядного представления данных из различных источников, обнаружения ключевой информации среди данных.

– iBase – позволяет совместно работающим коллективам аналитиков собирать, контролировать и анализировать данные из нескольких источников в единой защищенной рабочей среде.

Исходя из результатов анализа текстовых данных с помощью Big-Insights и Content Analytics создается математическая модель для постройки iBase базы данных социального профиля. На ее основе строится граф, задающий возможные взаимосвязи между *рассматриваемой персоной* и сущностями ее социального профиля (*упоминаемые персоны* – информация о лицах, связанных с рассматриваемой персоной в каком-либо контексте; *организации* – информация о различных учреждениях, связанных с персонами из социального профиля; *мероприятия* – информация о событиях, объединяющих группу людей по некоторым общим признакам; *контактные данные* персоны; *деятельность*, *специализация*, *достижения* и *хобби* персоны). Схематичный пример такого графа представлен на рис. 4.

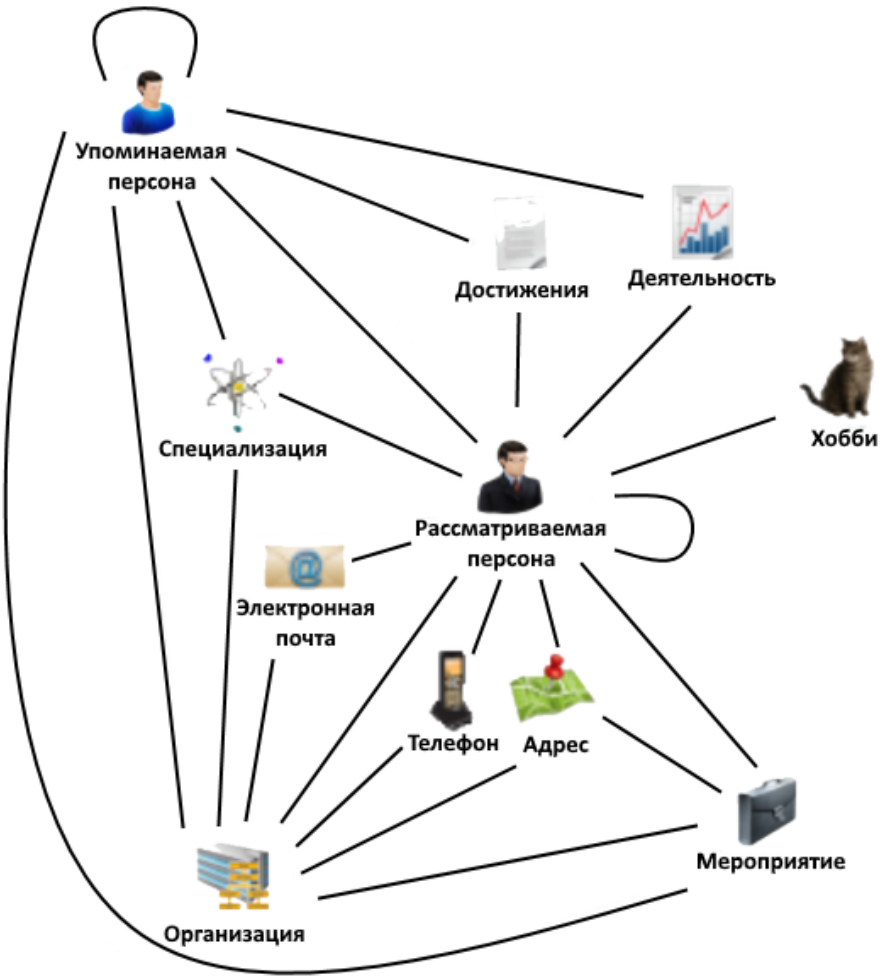

Рис. 4. Образец графа типов сущностей социального профиля

Основная работа по выявлению связей и зависимостей может производиться в программе IBM i2 TextChart. В этом случае анализ данных выполняется по следующему алгоритму:

1. Данные социального профиля заносятся в проект Text Chart через импорт CSV-файла с исходной информацией.

2. В тексте выделяется первая попавшаяся важная информация, после чего проводится поиск повторений и синонимов по всему тексту.

3. Результаты добавляются в проект как сущности социального профиля (рассматриваемая персона, деятельность и др.).

4. В тексте выделяются слова, выражающие отношения и связи между созданными объектами.

5. Посредством навигации между результатами поиска сущностей, аналогичным образом добавляются новые объекты и атрибуты уже существующих.

6. При нахождении противоречивой информации подсчитываются записи для каждого из вариантов, после чего делается вывод об их истинности: ложная информация удаляется из социального профиля или проводится уточняющий поиск.

Результатами визуального анализа являются социальный граф и iBase БД социального профиля. Помимо скрытой информации, выявленной непосредственно во время анализа, возможно нахождение неявных связей на полученном графе социального профиля.

3. Методы анализа мультимедиа информации социального профиля

В отличие от текстовых данных, мультимедиа контент трудно поддается анализу с использованием классических средств, поэтому приходится прибегать к использованию решений на основе больших данных, нейронных сетей и машинного обучения. Примерами таких систем можно считать Google Analytics, MS Azure, Multimedia Mining Marvel, Quaero.

В рамках построения социального профиля мультимедиа контент может рассматриваться в двух различных вариантах:

1. Данные мультимедиа, просматриваемые и создаваемые исследуемым лицом – информация, которая может говорить о деятельности и предпочтениях персоны. Также сюда относится и авторский контент. Задача анализа состоит в сравнении мультимедийных объектов с существующими образцами в сети Интернет и внутри социального профиля (например, определение музыкальных предпочтений по нескольким аудиозаписям).

2. Контент, содержащий в себе сведения непосредственно о рассматриваемом лице. Целью анализа является выделение существенной информации из самого мультимедийного объекта (например, распознавание эмоций на фото, выделение смысловых выражений из аудиозаписей).

Для обработки мультимедийной информации первой категории целесообразно использовать контентный метод анализа, суть которого состоит в разбиении данных на составные части и их прямом сравнении [8]. Для упомянутого выше примера определения музыкальных предпочтений будет выполнятся следующий алгоритм. Сначала у существующих в базе социального профиля аудиозаписей проверяются ID3 теги на наличие Исполнителя и Жанра музыки. Если таковые теги найдены, то они записываются в таблицу предпочтений. В ином случае производится сравнительный анализ аудиоза-

писи с образцами из Интернета (средствами утилит AudioTag, Shazam, Google Sound Search), после обнаружения совпадения искомые теги добавляются в таблицу. По завершении обработки всех имеющихся аудиозаписей производится подсчет тегов и делается вывод о преобладании определенного жанра или исполнителя в выборке.

Рассмотрим подробнее вторую категорию. Для ее анализа используется контентно-интерпретационный метод, согласно которому составным частям мультимедийных данных присваиваются понятия на формальном языке, после чего выстраиваются связи между ними. Подходы к анализу аудио- и графической информации различаются между собой.

Основным направлением анализа аудиозаписей при построении социального профиля является распознавание речи и ее интонации. Результатами аудиоанализа являются: звуковые характеристики голоса (спектрально-временные, кепстральные, амплитудно-частотные и др.) рассматриваемой персоны, прикрепляемые к социальному профилю; распознанные тексты, связывающиеся с исходной записью. В настоящее время существует достаточное количество свободно распространяемых систем распознавания речи с открытым исходным кодом: CMU Sphinx, Julius, HTK, Praat, SHoUt, VoxForge и др. Многие из них основаны на использовании скрытых Марковских моделей и нейронных сетей.

Распознавание интонации речи человека может проходить в три этапа: запись речи человека и разделение ее на законченные интонационные конструкции, выделение тона голоса в каждой из частей, построение классификатора интонационных конструкций [9]. Точность работы подобного алгоритма зависит от качества и продолжительности записей, особенностей речи и т.д.

Анализ графической информации включает в себя: распознавание образов, текста и сопоставление результатов с датой создания рассматриваемого файла. Подходы к распознаванию графической информации делятся на три категории: методы перебора, искусственные нейронные сети и поиск контуров объекта с дальнейшим исследованием их свойств. Технология OCR (оптическое распознавание символов, англ.: *optical character recognition*) позволяет находить печатный и в меньшей степени рукописный текст. Как и в случае с анализом аудиотекста, распознанный на изображении текст должен привязываться к исходному объекту и в дальнейшем рассматриваться с помощью средств текстовой аналитики. После завершения обработки изображения полученные данные сводятся в результирующую таблицу, содержащую следующие ключевые параметры: распознанные лица, их эмоции, список надписей, данные об окружении (распознанные объекты), сервисная информация об изображении (размер, дата создания, название и т.п.).

Распознавание образов в рамках построения социального профиля подразделяется на поиск лиц на изображениях, определение их выражений, а также выделение элементов окружения. Услуги распознавания лиц предоставляют такие сервисы, как ASID, FaceID, FindFace, Vissage Gallery. Распознавание эмоций является более сложной процедурой, основными проблемами которой являются определение положения и цвета лица, степень освещения, наличие посторонних объектов на переднем плане изображения. Однако, несмотря на это, существуют готовые решения, такие как FaceReader, FaceSecurity и др. Разработка систем определения элементов окружения не-

достаточно развита на сегодня, поэтому для решения этой задачи целесообразно использовать специально обученные нейронные сети.

Заключение

Степень проработанности аналитической подсистемы построения социального профиля влияет на информативность и корректность конечного социального профиля. Показано, что автоматизированная обработка неструктурированных текстовых данных с использованием средств Big Data, визуального анализа и обработки естественного языка на примере использования программных продуктов IBM BigInsights, Content Analytics, i2 позволяет построить подробный граф, учитывающий явные и скрытые взаимосвязи между объектами социального профиля. Предложен обзор существующих подходов к анализу мультимедийного контента, а также рассмотрена возможность их применимости в задаче построения социального профиля. Выявлено, что для обработки аудиоданных возможно использование уже существующих алгоритмов, а анализ графической информации требует совершенствования технологии распознавания.

Библиографический список

1. **Бождай, А. С.** Исследование процесса идентификации человека в сетях открытого доступа и построения его социального профиля на основе технологий Big Data / А. С. Бождай, А. Ю. Тимонин // Модели, системы, сети в экономике, технике, природе и обществе. – 2016. – № 2 (18). – С. 112–119.
2. **Бождай, А. С.** Исследование проблемы фильтрации исходных данных социального профиля / А. С. Бождай, А. Ю. Тимонин // Математическое и компьютерное моделирование естественно-научных и социальных проблем : материалы X Междунар. науч.-техн. конф. молодых специалистов, аспирантов и студентов / под ред. И. В. Бойкова (Пенза, 23–27 мая 2016 г.). – Пенза : Изд-во ПГУ, 2016. – С. 130–135.
3. Официальный сайт Neo4j: The World's Leading Graph Database. – 2017. – URL: https://neo4j.com (дата обращения: 02.02.2017).
4. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Е. И. Большакова, Э. С. Клышинский, Д. В. Ландэ, А. А. Носков, О. В. Пескова, Е. В. Ягунова. – М. : МИЭМ, 2011. – 272 с.
5. Анализ структурированных и неструктурированных данных с помощью Content Analytics // Центр компетенции по технологии IBM Big Data. – М., 2014. – 66 с.
6. Официальный сайт проекта Apache Hadoop. – 2017. – URL: http://hadoop.apache.org (дата обращения: 02.02.2017).
7. Выявление скрытых связей на основе анализа текстов с помощью i2 // Центр компетенции по технологии IBM Big Data. – М., 2014. – 47 с.
8. **Яковлев, В. Е.** Макромедиа: анализ мультимедиа информации. M-Lang / В. Е. Яковлев // Молодой ученый. – 2011. – Т. 1, № 4. – С. 105–108.
9. **Бойков, И. В.** Алгоритм построения статистического дискретно-континуального описания длительности звуков потока осмысленной речи диктора / И. В. Бойков, А. И. Иванов, Д. М. Калашников // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2015. – № 4 (36). – С. 64–78.

References

1. Bozhday A. S., Timonin A. Yu. *Modeli, sistemy, seti v ekonomike, tekhnike, prirode i obshchestve* [Models, systems, networks in economy, engineering, environment and society]. 2016, no. 2 (18), pp. 112–119.

2. Bozhday A. S., Timonin A. Yu. *Matematicheskoe i komp'yuter-noe modelirovanie estestvenno-nauchnykh i sotsial'nykh problem: materialy X Mezhdunar. nauch.-tekhn. konf. molodykh spetsialistov, aspirantov i studentov (Penza, 23–27 maya 2016 g.)* [Mathematical and computer modeling of natural scientific and social problems: proceedings of X International scientific and technical conference of young scientists, undergraduate and postgraduate students (Penza, 23rd–27th May 2016)]. Penza: Izd-vo PGU, 2016, pp. 130–135.
3. *Official site Neo4j: The World's Leading Graph Database*. 2017. Available at: https://neo4j.com (accessed February 02, 2017).
4. Bol'shakova E. I., Klyshinskiy E. S., Lande D. V., Noskov A. A., Peskova O. V., Yagunova E. V. *Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i komp'yuternaya lingvistika: ucheb. posobie* [Automatic processing of natural language texts and computer linguistics: teaching aid]. Moscow: MIEM, 2011, 272 p.
5. *Tsentr kompetentsii po tekhnologii IBM Big Data* [IBM Big Data competence center]. Moscow, 2014, 66 p.
6. *Official site project Apache Hadoop*. 2017. Available at: http://hadoop.apache.org (accessed February 02, 2017).
7. *Tsentr kompetentsii po tekhnologii IBM Big Data* [IBM Big Data competence center]. Moscow, 2014, 47 p.
8. Yakovlev V. E. *Molodoy uchenyy* [Young scientist]. 2011, vol. 1, no. 4, pp. 105–108.
9. Boykov I. V., Ivanov A. I., Kalashnikov D. M. *Izvestiya vysshikh uchebnykh zavedeniy. Povolzhskiy region. Tekhnicheskie nauki* [University proceedings. Volga region. Engineering sciences]. 2015, no. 4 (36), pp. 64–78.

Бождай Александр Сергеевич
доктор технических наук, профессор, кафедра систем автоматизированного проектирования, Пензенский государственный университет (Россия, г. Пенза, ул. Красная, 40)

E-mail: bozhday@yandex.ru

Bozhday Aleksandr Sergeevich
Doctor of engineering sciences, professor, sub-department of CAD systems, Penza State University (40 Krasnaya street, Penza, Russia)

Тимонин Алексей Юрьевич
аспирант, Пензенский государственный университет (Россия, г. Пенза, ул. Красная, 40)

E-mail: c013s017b301f018@mail.ru

Timonin Aleksey Yur'evich
Postgraduate student, Penza State University (40 Krasnaya street, Penza, Russia)

УДК 004.62

Бождай, А. С.

Исследование процесса анализа текстовых и мультимедиа данных социального профиля из открытых источников информации / А. С. Бождай, А. Ю. Тимонин // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2017. – № 2 (42). – С. 19–28. DOI 10.21685/2072-3059-2017-2-2