УДК 004.89

Э.Д. Павлыгин, А.Г. Подлобошников, Р.А. Савинов, Н.Г. Ярушкина,
А.М. Наместников, А.А. Филиппов, А.А. Романов, В.С. Мошкин,
Г.Ю. Гуськов, М.С. Григоричева

РАЗРАБОТКА ПРОГРАММНОГО КОМПЛЕКСА ДЛЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА СОЦИАЛЬНЫХ МЕДИА[1]

Павлыгин Эдуард Дмитриевич, *кандидат технических наук, окончил радиотехнический факультет Ульяновского политехнического института. Первый заместитель генерального директора по науке и инновационному развитию ФНПЦ АО «НПО «Марс». Имеет статьи в области статистических методов обработки сигналов. [e-mail: mars@mv.ru].*

Подлобошников Анатолий Геннадьевич, *окончил факультет информационных систем и технологий Ульяновского государственного технического университета, начальник научно-исследовательской лаборатории – главный конструктор ФНПЦ АО «НПО «Марс». Область научных интересов – специализированные информационные системы. [e-mail: mars@mv.ru].*

Савинов Руслан Анатольевич, *окончил машиностроительный факультет УлГТУ, инженер-программист I категории ФНПЦ АО «НПО «Марс». Область научных интересов – специализированные информационные системы. [e-mail: mars@mv.ru].*

Ярушкина Надежда Глебовна, *доктор технических наук, профессор, окончила Ульяновский политехнический институт. Первый проректор – проректор по научной работе УлГТУ, заведующая кафедрой «Информационные системы» УлГТУ. Имеет более 300 работ в области мягких вычислений, нечеткой логики, гибридных систем. [e-mail: jng@ulstu.ru].*

Наместников Алексей Михайлович, *доктор технических наук, доцент, окончил радиотехнический факультет УлГТУ. Профессор кафедры «Информационные системы» УлГТУ. Имеет более 80 работ в области автоматизированного проектирования и интеллектуальных систем. [e-mail: nam@ulstu.ru].*

Филиппов Алексей Александрович, *кандидат технических наук, окончил факультет информационных систем и технологий УлГТУ, доцент кафедры «Информационные системы» УлГТУ. Имеет статьи в области онтологического моделирования и построения автоматизированных систем обработки знаний. [e-mail: al.filippov@ulstu.ru].*

Романов Антон Алексеевич, *кандидат технических наук, окончил факультет информационных систем и технологий УлГТУ, доцент кафедры «Информационные системы» УлГТУ. Имеет статьи в области систем хранения и обработки информации и интеллектуального анализа временных рядов. [e-mail: romanov73@gmail.com].*

1 Работа выполнена при частичной финансовой поддержке РФФИ (проекты № 18-47-730035 р_а, 18-47-732007 р_мк).

Мошкин Вадим Сергеевич, *кандидат технических наук, окончил факультет информационных систем и технологий УлГТУ, доцент кафедры «Информационные системы» УлГТУ. Имеет более 70 статей в области интеллектуальных систем анализа данных. [e-mail: postforvadim@ya.ru].*

Гуськов Глеб Юрьевич, *окончил факультет информационных систем и технологий УлГТУ. Старший преподаватель кафедры «Информационные системы» УлГТУ. Имеет работы в области онтологического инжиниринга и интеллектуального анализа данных. [e-mail: g.guskov@ulstu.ru].*

Григоричева Мария Сергеевна, *аспирант кафедры «Информационные системы» УлГТУ, окончила факультет информационных систем и технологий УлГТУ. Ассистент кафедры «Информационные системы» УлГТУ. Имеет работы в области интеллектуального анализа данных. [e-mail: gms4295@mail.ru].*

Аннотация

В статье представлены результаты работы над программным комплексом для интеллектуального анализа социальных медиа. Описана архитектура программного комплекса, перечислены основные подсистемы программного комплекса, а также сторонние программные системы, используемые при разработке программного комплекса.

Рассмотрена организация подсистемы хранения данных программного комплекса, описана модель данных данной подсистемы.

Описан подход к организации хранилища онтологий программного комплекса на основе графовой базы данных, представлен метод трансляции онтологии в формате OWL/XML во фрагмент графовой базы данных.

Рассмотрена организация подсистемы поиска данных, описан метод расширения поискового запроса с применением онтологии для учета в процессе поиска особенностей предметной области. Также представлен метод форматирования поискового запроса для выделения смысловых элементов запроса для повышения качества поиска.

Описана организация подсистемы формирования социального портрета пользователя, рассмотрен метод определения категорий интересов пользователя социальной сети «ВКонтакте».

Ключевые слова: социальные медиа, онтологический анализ, анализ текстов на естественном языке, интеллектуальный анализ данных, программный комплекс.

doi: 10.35752/1991-2927-2019-2-56-23-36

DEVELOPMENT OF A SOFTWARE PACKAGE FOR DATA MINING OF SOCIAL MEDIA

Eduard Dmitrievich Pavlygin, *Candidate of Science in Engineering; graduated from the Radioengineering Faculty of Ulyanovsk Polytechnic Institute; First Deputy of Director General in Scientific Affairs and Innovations of Federal Research-and-Production Center Joint Stock Company 'Research-and-Production Association 'Mars'; an author of articles in the field of statistical methods of signal processing. e-mail: mars@mv.ru.*

Anatolii Gennadievich Podloboshnikov, *graduated from the Faculty of Information Systems and Technologies of Ulyanovsk State Technical University; Head of the Research Laboratory and Chief Designer at FRPC JSC 'RPA 'Mars'; research interests are in the field of special-purpose information system. e-mail: mars@mv.ru.*

Ruslan Anatolevich Savinov, *graduated from the Machine-building Faculty of UlSTU; Software Engineer at FRPC JSC 'RPA 'Mars'. research interests are in the field of special-purpose information system. e-mail: mars@mv.ru.*

Nadezhda Glebovna Yarushkina, *Doctor of Science in Engineering, Professor; graduated from Ulyanovsk Polytechnic Institute; First Vice-Rector, Vice-Rector in Scientific Affairs in UlSTU, Head of the Department of Information Systems at UlSTU; an author of more than 300 publications in the field of soft computing, fuzzy logic, and hybrid systems. e-mail: jng@ulstu.ru.*

Aleksei Mikhailovich Namestnikov, *Doctor of Science in Engineering, Associate Professor; graduated from the Radioengineering Faculty of UlSTU; an author of more than 80 publications in the field of computer-aided design and intelligent systems. e-mail: nam@ulstu.ru.*

Aleksei Aleksandrovich Filippov, *Candidate in Science in Engineering; graduated from the Faculty of Information Systems and Technologies of UlSTU; Associate Professor at the Department of Information Systems of UlSTU; an author of articles in the field of ontological modelling and building of computer-aided systems for knowledge processing. e-mail: al.filippov@ulstu.ru.*

Anton Alekseevich Romanov, *Candidate of Science in Engineering; graduated from the Faculty of Information Systems and Technologies of UlSTU; Associate Professor at the Department of Information Systems of UlSTU;*

an author of articles in the field of systems for data storage and processing, and time series mining. e-mail: romanov73@gmail.com.

Vadim Sergeevich Moshkin, *Candidate of Science in Engineering; graduated from the Faculty of Information Systems and Technologies at UlSTU; Associate Professor of Information Systems at UlSTU; an author of more than 70 articles in the field of data mining systems. e-mail: postforvadim@ya.ru.*

Gleb Iurevich Guskov, *graduated from the Faculty of Information Systems and Technologies of UlSTU; Senior Lecturer at the Department of Information Systems of UlSTU; an author of publications in the field of ontological engineering and data mining systems. e-mail: g.guskov@ulstu.ru.*

Maria Sergeevna Grigoricheva, *Postgraduate Student at the Department of Information Systems of UlSTU; graduated from the Faculty of Information Systems and Technologies of UlSTU; Assistant Lecturer at the Department of Information Systems of UlSTU; an author of publications in the field of data mining. e-mail: gms4295@mail.ru.*

Abstract

The article presents the results of work on a software package for data mining of social media. The architecture of the software package is described. Listed the main subsystems of the software package and third-party software systems used in the development of a software package.

The organization of the data storage subsystem of the software package is considered, the data model of this subsystem is described.

The approach to organizing the ontology repository of a software package based on a graph database is described, and the ontology translation method in the OWL / XML format into a graph database fragment is presented.

The organization of the data search subsystem is considered, the method of extending the search query using ontology for accounting for the features of the subject area in the search process is described. Also presented is a method of formatting a search query to highlight the important elements of the query to improve the quality of the search.

The organization of the social portrait building subsystem of the user of the social network VKontakte is described, the method of determining the categories of interests of the user is considered.

Key words: social media, ontological analysis, natural language processing, data mining, software package.

Введение

В настоящее время анализ социальных медиа позволяет решать широкий круг задач [1–11], связанных с определением отношения группы людей к некоторой ситуации, с возможностью получения значений динамики и темпа роста интереса, а также принадлежности каждого члена такой группы к определенному социальному классу.

По данным статистических исследований, ежемесячно в русскоязычном секторе социальных медиа около 30 миллионов уникальных авторов публикуют почти 580 миллиардов сообщений, а количество электронных средств массовой информации (ЭСМИ, новостные сайты) в России превышает 7000 наименований. Сформировалась потребность в разработке программного комплекса для автоматизированного интеллектуального анализа социальных медиа, позволяющего за приемлемое время найти и получить необходимые сведения [6–11].

Автоматизация процесса информационного поиска обусловлена огромным количеством слабоструктурированных данных, содержащихся в социальных медиа. Первые автоматизированные информационно-поисковые системы работали преимущественно с информацией фактического характера, например, характеристиками объектов и их связей. Со временем появилась возможность обрабатывать текстовые документы на естественном языке и другие форматы представления данных [12].

Критерии качества поиска зависят не только от характеристик самой информационно-поисковой системы, но и от того, как сформулирован запрос. Идеальный запрос может быть составлен пользователем, в полном объеме знакомым с интересующей его предметной областью (ПрО), а также с применяемой системой, иначе пользователь вынужден довольствоваться или низкой точностью поиска, или низкой полнотой [13].

Разрабатываемый в рамках данного проекта программный комплекс для интеллектуального анализа социальных медиа (ПКИАД) предполагает использование методов онтологического анализа, инженерии знаний и анализа текстов на естественном языке. Данные подходы позволяют повысить эффективность анализа содержимого социальных медиа с учетом слабоструктурированного представления данных и нечеткости конструкций естественного языка [14–18].

В настоящее время в качестве языков представления онтологий используются языки RDF и OWL [19, 20], при этом язык OWL является более выразительным, так как содержит различные виды функциональных отношений.

Для работы с онтологиями в процессе функционирования интеллектуальных систем часто применяется библиотека OWL API [21], которая предполагает написание программного кода для получения необходимых фрагментов онтологии. На данный момент OWL API обладает наибольшими функциональными возможностями [22], но может быть использована только в программах, написанных для платформы Java Virtual Machine.

Таким образом, не существует универсального метода работы с онтологиями и формирования запросов к их содержимому. В качестве решения данной проблемы можно предложить использование специализированных хранилищ, например, StarDog [23], Virtuoso [24], RDF4j [25] и т. д. Однако, существующие хранилища онтологий обладают следующими недостатками:

• необходимость покупки лицензии для использования,

• требуют от разработчика знаний в области онтологического проектирования и инженерии знаний,

• полная поддержка только формата RDF.

Специфика работы со слабоструктурированными данными социальных медиа требует учета следующих особенностей:

• скорость доступа к информации,

• объем данных,

• неструктурированное представление информации,

• отсутствие единого понятийного аппарата.

В статье предложен метод лингвистического анализа и форматирования текстового запроса пользователя, целью которого является улучшение значений критериев качества информационного поиска.

Также рассмотрена работа хранилища онтологий, позволяющего:

- производить импорт онтологий в формате OWL;

- формировать запросы к содержимому хранилища;

- не требовать от разработчика знаний в области онтологического проектирования и инженерии знаний;

- организовать взаимодействие с хранилищем онтологий с помощью протокола HTTP, сделав хранилище максимально независимым от используемого языка программирования и операционной системы.

1 Архитектура программного комплекса

На рисунке 1 представлена архитектурная схема ПКИАД:

1. Подсистема импорта данных из социальных медиа позволяет загружать текстовые данные из социальной сети «Вконтакте» и ЭСМИ. Загрузчик онтологий позволяет импортировать описание особенностей ПрО в виде онтологий в формате OWL/XML.

2. Подсистема хранения данных обеспечивает представление информации, извлеченной из социальных медиа, в унифицированном виде, удобном для дальнейшей обработки и анализа. Данные хранятся в разрезе пользователей, коллекций, источников данных, версий и т. д. В качестве систем управления базами данных (СУБД) используются:

• Elasticsearch для организации информационного поиска [26],

• MongoDB для хранения данных [27],

• Neo4j для хранения графов социального взаимодействия и онтологий [28].

Конвертер данных обеспечивает преобразование импортированных из социальных медиа данных во внутреннее представление ПКИАД. Построитель социального графа формирует граф социального взаимодействия на основе отношений пользователей и сообществ,

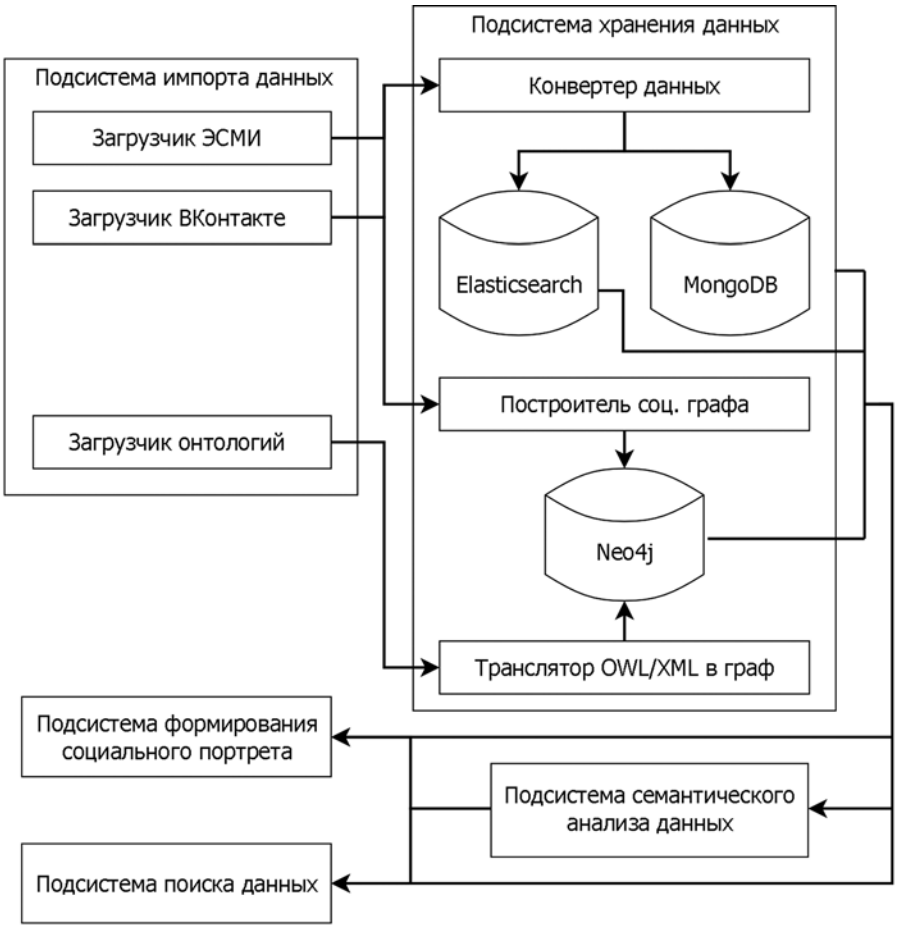

Рис. 1. Архитектурная схема ПКИАД

извлеченных из социальной сети «ВКонтакте». Транслятор OWL/XML в граф позволяет преобразовывать онтологию в содержимое хранилища онтологий.

3. Подсистема семантического анализа данных обеспечивает выполнение предобработки текстовых данных, проведение статистического и лингвистического анализов текстовых ресурсов.

4. Подсистема формирования социального портрета позволяет определить классы пользователей социальной сети «ВКонтакте» посредством классификации текстовых фрагментов (посты социальной сети, комментарии) и данных со страницы. Классами являются категории интересов пользователя – темы, связанные с предметными областями: спорт, IT-технологии, музыка, бизнес и прочие.

5. Подсистема поиска данных обеспечивает возможности информационного поиска объектов хранилища данных, имеющих отношение к возникшей ситуативной задаче, формируемой в виде множества ключевых слов. При этом существует возможность семантического расширения запроса пользователя на основе описания особенностей ПрО, представленного в виде онтологии.

2 Модель данных подсистемы хранения данных программного комплекса

Для хранения данных, извлеченных из социальных медиа, используется документоориентированная СУБД MongoDB. К основным преимуществам MongoDB можно отнести:

- высокую производительность,
- документоориентированный язык запросов,
- отказоустойчивость,
- масштабирование.

Для унификации данных, загруженных из различных социальных медиа, необходимо выделить основные сущности модели данных ПКИАД.

Сущность MassMedia – коллекция, элементы которой соответствуют определенным социальным медиа. Загрузку данных социальных медиа поддерживает подсистема импорта данных ПКИАД.

Сущность Person – коллекция, элементы которой содержат список пользователей, извлеченных из социальных медиа. Сущность Person имеет набор признаков, соответствующий атрибутам, часто используемым в социальных сетях: фамилия, имя, отчество, дата рождения, увлечения, сведения об образовании и т. д.

Сущность Group – коллекция, элементы которой содержат информацию о сообществах, извлеченных из социальных медиа. Сущность Group имеет набор признаков, соответствующих атрибутам, часто используемым в социальных сетях: название группы, описание группы, возрастные ограничения, дата создания и т. д.

Сущность Post – коллекция, элементы которой содержат информацию о записях в социальных медиа. Сущность Post имеет следующий набор атрибутов: автор, заголовок, содержимое, дата создания, вложения и т. д.

Сущность Comment – коллекция, элементы которой содержат информацию о комментариях в социальных медиа. Сущность Comment имеет следующий набор атрибутов: автор, заголовок, содержимое, дата создания, вложения и т. д.

Сущность Attachment – коллекция, элементы которой содержат информацию о вложениях записей и комментариев в социальных медиа. Сущность Attachment имеет несколько типов и позволяет хранить следующие виды вложений: фотографии, фотоальбомы, аудиозаписи, видеозаписи, гиперссылки, документы (файлы), опросы и т. д.

В таблице 1 представлено соответствие сущностей социальных медиа, загрузку которых поддерживает подсистема импорта данных ПКИАД, и сущностей ПКИАД.

Как видно из таблицы 1, выделенные ранее основные сущности модели данных ПКИАД позволяют хранить данные, загруженные из большинства существующих социальных медиа. Унифицированное представление данных в ПКИАД позволяет эффективно производить процесс их обработки, анализа и поиска. Для трансформации данных, загруженных из социальных медиа, во внутреннее представление ПКИАД используется конвертер данных. Для каждого нового Интернет-ресурса необходима разработка своего модуля

Таблица 1

Соответствие сущностей популярных социальных медиа сущностям подсистемы хранения данных ПКИАД

ПКИАД	ВКонтакте, Facebook, Одноклассники	Twitter	Instagram	Youtube	ЭСМИ
MassMedia	URL, например, vk.com	URL	URL	URL	URL
Person	Пользователь	Пользователь	Пользователь	Пользователь	-
Group	Сообщество	-	-	-	-
Post	Запись	Твит	Фотография	Видео	Новость, Статья
Comment	Комментарий	Комментарий	Комментарий	Комментарий	Комментарий
Attachment	Вложения	Вложения	Тэги, ссылки	Ссылки	Вложения

в рамках конвертера данных. Загрузчик ЭСМИ формирует одинаковое представление данных для всех сайтов, следовательно, нет необходимости адаптировать конвертер для каждого сайта в отдельности.

3 Модель данных хранилища онтологий программного комплекса

Для хранения описания ПрО в форме графовой базы знаний используется графовая СУБД Neo4j. К основным преимуществам Neo4j можно отнести:

1. Естественный (native) формат хранения графов;
2. Один экземпляр СУБД может обслуживать графы с миллиардами узлов и связей;
3. Может обрабатывать графы, которые полностью не помещаются в оперативную память;
4. Графо-ориентированный язык запросов – Cypher.

Для успешной трансляции онтологии, представленной в формате OWL/XML, в содержимое хранилища онтологий ПКИАД необходимо выделить структурные элементы, которые будут отнесены к TBox (структура, схема) и ABox (наполнение, содержимое) модели данных хранилища онтологий.

Формально модель данных хранилища онтологий ПКИАД (модель онтологии) можно представить в виде следующего выражения:

$$O = \langle C, I, R \rangle, \quad (1)$$

где $C = \{C_1, C_2, ..., C_n\}$ – множество классов модели онтологии;

$I = \{I_1, I_2, ..., I_n\}$ – множество объектов модели онтологии;

R – множество отношений модели онтологии следующего вида:

$$R = \{R_C, R_I\}, \quad (2)$$

где R_C – множество отношений, формирующих иерархию классов модели онтологии;

R_I – множество отношений, определяющих связь вида «класс-объект» модели онтологии.

При этом отношения R_C и R_I могут представлять функциональные отношения, характерные для языка OWL.

Функции трансляции OWL-онтологии в содержимое хранилища онтологии ПКИАД (1) можно представить следующим выражением:

$$f_O^{OWL} : OWL \to O,$$

где $OWL = \langle C^{OWL}, I^{OWL}, R^{OWL} \rangle$ – множество сущностей онтологии в формате OWL (соответствует сущностям выражения 1),

O – множество сущностей модели данных хранилища онтологий ПКИАД (1).

В таблице 2 представлено соответствие сущностей OWL-онтологии сущностям модели данных хранилища онтологий ПКИАД.

Таким образом, модель данных хранилища онтологий ПКИАД позволяет разработчикам формировать запросы к содержимому хранилища онтологий на языке запросов Cypher. Данный метод извлечения знаний из хранилища онтологий является более привычным для разработчика, чем работа с методами OWL API.

4 Организация подсистемы информационного поиска программного комплекса

Для организации поиска данных используется программная поисковая система Elasticsearch. К основным преимущества Elasticsearch можно отнести:

1. Может обрабатывать петабайт структурированных и неструктурированных данных.
2. Использование денормализации для увеличения эффективности поиска.
3. Одна из популярных поисковых систем, которая в настоящее время используется многими крупными организациями, такими как Wikipedia, The Guardian, StackOverflow, GitHub и т. д.

Однако программная поисковая система Elasticsearch обладает рядом недостатков:

• запросы пользователя обрабатываются системой без учета особенностей ПрО, что снижает полноту поиска;

• используется обычный поиск по словам, объединенным оператором ИЛИ, что снижает точность поиска.

Для решения проблемы учета особенностей ПрО используется метод расширения запроса терминами из онтологии ПрО [29, 30].

Таблица 2

Соответствие сущностей онтологии в формате OWL/XML сущностям модели данных хранилища онтологий ПКИАД

OWL	Хранилище онтологии ПКИАД
TBox	
owl:Thing	$C = \{C_1, C_2, ..., C_n\}$
owl:Class	$C = \{C_1, C_2, ..., C_n\}$
owl:SubclasOf	R_C
owl:ObjectProperty owl:DataProperty	R_C
owl:ObjectPropertyDomain owl:DataPropertyDomain	R_C
owl:ObjectPropertyRange owl:DataPropertyRange	R_C
ABox	
owl:NamedIndividual	$I = \{I_1, I_2, ..., I_n\}$
owl:ClassAssertion	R_I
owl:ObjectPropertyAssertion owl:DataPropertyAssertion	R_I

Поисковый запрос $Q = \{\hat{w}_1, \hat{w}_2, ..., \hat{w}_n\}$ есть множество ключевых слов запроса к данным, агрегированным из ЭСМИ и социальных сетей, проходит процесс предобработки:

- удаление стоп-слов,
- стемминг (выделение основы слова, получение термов).

Для расширения поискового запроса используется функция F^{Ext}, сопоставляющая набору терминов запроса Q подмножество объектов онтологии ПрО $\hat{I}$ (1), фактически получая множество терминов, которыми будет расширен запрос:

$$Q^{Ext} = Q \cup F^{Ext}(Q),$$

где Q^{Ext} – расширенный поисковый запрос;

$F^{Ext}(Q)$ – функция расширения поискового запроса Q, которая имеет вид:

$$F^{Ext} : Q \to G^{Ext} \to w^{Ext},$$

где Q – исходный поисковый запрос;

w^{Ext} – множество терминов, расширяющих исходный поисковый запрос Q, такое, что $w^{Ext} = \hat{I}^M \cup \hat{I}^{Ext}$;

$G^{Ext} = \left\langle \hat{I}^M, \hat{R}_I^M, \hat{C}^M, \hat{R}_C, \hat{C}^{Ext}, \hat{R}_I^{Ext}, \hat{I}^{Ext} \right\rangle$ – фрагмент онтологии ПрО, состоящий из следующих множеств:

$\hat{I}^M$ – множество объектов онтологии ПрО таких, что $\hat{I}^M \in Q$,

$\hat{R}_I^M$ – множество отношений онтологии ПрО вида «класс-объект», соединяющих множество объектов онтологии ПрО $\hat{I}^M$ с множеством классов онтологии ПрО $\hat{C}^M$;

$\hat{R}_C$ – множество отношений онтологии ПрО вида «класс-класс», соединяющих множество классов онтологии ПрО $\hat{C}^M$ с множеством классов онтологии ПрО $\hat{C}^{Ext}$;

$\hat{R}_I^{Ext}$ – множество отношений онтологии ПрО вида «класс-объект», соединяющих множество классов онтологии ПрО $\hat{C}^{Ext}$ с множеством объектов онтологии ПрО $\hat{I}^{Ext}$.

Таким образом, расширение запроса сводится к заданию обратной и прямой функций интерпретации.

Фрагмент онтологии, используемой для учета особенностей ПрО, представлен на рисунке 2.

Таким образом, если на вход поисковой системы поступит поисковый запрос «*улгту*»:

1. Будет выполнен поиск объектов, текст которых содержит слово из поискового запроса – объект «улгту».

2. Далее происходит переход к классу, соединенному отношением «класс-объект» с найденным объектом «улгту» – «УлГТУ».

3. Затем будет произведен переход к классам, имеющим связь с классом «УлГТУ» – классы «Ульяновск» и «Государственный».

4. Конечное множество терминов формируется из объектов, имеющих отношение «класс-объект» с помеченными ранее классами, включая исходный: «ульяновск», «ульск», «государственный», «улгту», «политех» и «технический университет».

В результате выполнения данных шагов результирующий поисковый запрос будет иметь вид: «*улгту ульяновск ульск государственный "технический университет" политех*».

Для оценки качества предложенного метода используется показатель полноты информационного поиска. Полнота поиска определяется как отношение числа релевантных документов, найденных подсистемой поиска данных ПКИАД, к общему числу релевантных документов [13]:

$$R = \frac{a}{r},$$

где a – количество полученных в результате поиска релевантных документов;

r – общее количество релевантных документов, содержащихся в подсистеме хранения данных ПКИАД.

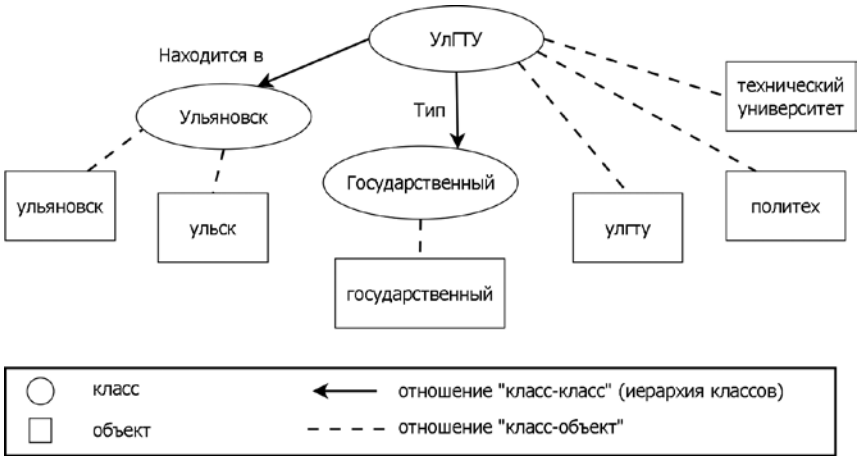

Рис. 2. Фрагмент онтологии для учета особенностей ПрО

Для исходного запроса Q = *улгту* количество полученных в результате поиска релевантных документов равно 336, общее количество релевантных документов равняется 866. В данном случае значение полноты будет равно $R(Q) = \frac{336}{866} = 0,388$.

Для результирующего запроса Q^{Ext} = *улгту ульяновск ульск государственный "технический университет" политех*, полученного из исходного запроса Q = *улгту*, количество релевантных документов равно 534, а общее количество релевантных документов – 866. В данном случае полнота информационного поиска будет

равна $R\left(Q^{Ext}\right)=\frac{534}{866}=0,617$.

Для решения проблемы отсутствия учета связей между словами поискового запроса используется подход с применением синтаксического анализа поискового запроса, к которому применяется множество правил для выделения из запроса смысловых групп – устойчивых словосочетаний, состоящих из связанных между собой слов.

Сначала необходимо произвести синтаксический анализ запроса, вводимого пользователем, с целью получения дерева зависимостей. Полученное дерево хранит данные о структуре запроса, зависимостях между словами и типах этих зависимостей и в дальнейшем будет использовано для обработки текста.

Дерево синтаксического разбора можно представить в виде следующего множества:

$$T=\{t_1,t_2,...,t_k\}, \quad (3)$$

где каждый элемент t_i – узел дерева, который описывается набором своих характеристик:

$$t_i=(i,\ w_i,\ m_i,\ c_i),$$

где k – количество элементов в дереве синтаксического разбора;

i – порядковый номер слова в запросе, причем $i=\{1, 2, ..., k\}$, $w_i \in W$, $W=\{w_1, w_2, ..., w_k\}$;

W – множество слов, из которых состоит исходный запрос;

$m_i \in M$, M = {*Существительное, Глагол, Прилагательное, Числительное, Наречие, Местоимение, Предлог, Союз, Частица, Междометие*} – множество частей речи, используемых в русском языке;

$p_i \in P$ – это порядковый номер слова в предложении запроса, которому подчинено данное слово, причем $P=\{0,1,...,k\}$.

Таким образом, каждому слову запроса ставится в соответствие его часть речи, порядковый номер слова в запросе, а также отражаются связи с другими словами запроса, если такие имеются.

Далее происходит выделение из полученного синтаксического дерева смысловых групп с использованием набора правил. В рамках данного процесса в соответствии с правилом элемент помечается определенным символом (символами) информационно-поискового языка (ИПЯ) Elasticsearch [31] в начале и при необходимости в конце:

- \+ – обязательное наличие элемента запроса в искомом документе;
- "слово1 слово2" – поиск по словосочетанию, используется логический оператор И.

Определим функцию:

$$F^{Sem}\left(T,R\right)=Q^{Sem}. \quad (4)$$

Функция F^{Sem} получает на входе дерево синтаксического разбора исходного запроса и набор правил, а возвращает результирующий запрос с выделенными и отформатированными согласно ИПЯ Elasticsearch смысловыми группами.

В выражении 4 параметр T – дерево синтаксического разбора (3), параметр R представляет собой множество правил поиска значимых элементов и их форматирования в исходном запросе:

$$R=\{R_1,R_2,...,R_n\},$$

где n – количество правил.

Каждое правило можно описать в виде следующего выражения:

$$R_i\left(p,t_1,t_2,...,t_m\right)\rightarrow Q^{Sem},$$

где t_m – элемент правила, который соответствует узлу синтаксического дерева;

m – количество узлов дерева, задействованных в правиле;

p – приоритет правила;

$Q_i^{Sem} \in Q^{Sem}$ – элемент результирующего запроса Q^{Sem}. При этом каждый элемент форматированного запроса Q_i^{Sem} содержит слово/слова исходного запроса, экранированного символом из множества допустимых операторов ИПЯ Elasticsearch.

Например, формальное описание правила поиска именных групп (существительное, связанное с прилагательными) в исходном запросе, можно представить в следующем виде:

$$\begin{gathered}R_{\text{именная\_группа}}\left(4,\left\langle i,w_i,Существительное,c_i\right\rangle,\right.\\ \left\langle i+1,w_{i+1},Прилагательное,i\right\rangle,\\ \left\langle i+2,w_{i+2},Прилагательное,i\right\rangle,...,\\ \left.\left\langle i+d,w_{i+d},Прилагательное,i\right\rangle\right)\rightarrow\\ \rightarrow +"w_iw_{i+1}w_{i+2}...w_{i+d}",\end{gathered}$$

где d – количество прилагательных в именной группе.

Поиск именных групп в запросе сводится к обходу дерева и поиску одного или нескольких прилагательных, подчиненных одному и тому же узлу дерева, являющегося существительным.

Поиск групп связанных между собой существительных в запросе сводится к обходу дерева и поиску одного или нескольких существительных, подчиненных одному и тому же узлу дерева, являющегося существительным.

Поиск имени собственного в запросе сводится к обходу дерева и поиску имени собственного, подчиненного узлу дерева, являющегося именем собственным, если такой узел есть.

Поиск существительного и связанного с ним имени собственного в запросе сводится к обходу дерева и поиску имени собственного, подчиненного узлу дерева, являющегося существительным.

Правило поиска обособленного существительного обладает низшим приоритетом и осуществляется только после того, как не выполнились правила

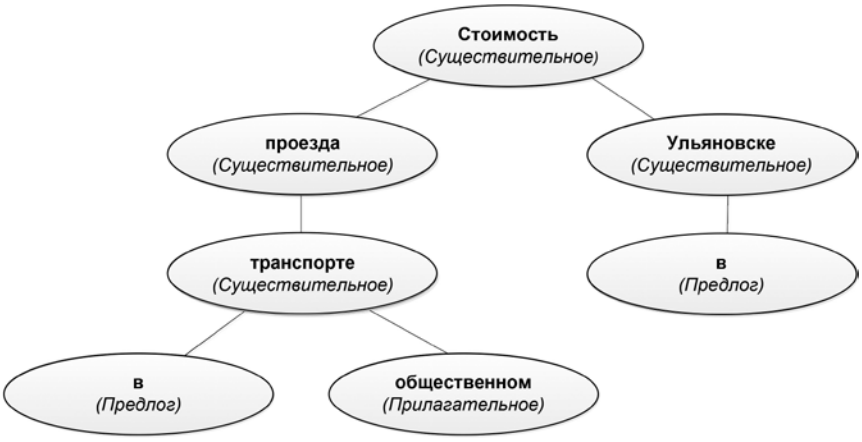

Рис. 3. Дерево синтаксического разбора запроса:
Стоимость проезда в общественном транспорте в Ульяновске

с приоритетом выше. Правило предполагает поиск узла с существительным, который не связан ни с другим существительным, ни с прилагательным.

Правило поиска глагола обладает высшим приоритетом и не пересекается ни с одним другим правилом. Оно сводится к поиску узла с данной частью речи.

Дерево синтаксического разбора, полученное для текста исходного запроса: *Стоимость проезда в общественном транспорте в Ульяновске*, представлено на рисунке 3. Узлами дерева являются слова предложения исходного запроса, каждому из которых в соответствие поставлена часть речи.

В результате работы предложенного метода в исходном запросе были найдены смысловые группы, представленные на рисунке 4 в виде дерева семантического разбора. Текст результирующего запроса выглядит следующим образом: *+"Стоимость проезда" + "общественном транспорте" +Ульяновске*.

Для оценки качества предложенного метода используется показатель точности информационного поиска. Точность поиска определяется как отношение числа релевантных документов, найденных подсистемой поиска данных ПКИАД, к общему числу найденных документов [13]:

$$P = \frac{a}{b},$$

где a – количество полученных в результате поиска релевантных документов;

b – общее количество документов, выданных подсистемой поиска данных ПКИАД.

Для исходного запроса Q = *Стоимость проезда в общественном транспорте в Ульяновске* количество полученных в результате поиска релевантных документов равно 8, количество выданных документов равняется 44 857. В данном случае значение точности будет равно

$$P(Q) = \frac{8}{44857} = 0,00018.$$

Для результирующего запроса Q^{Sem} = *+"Стоимость проезда" +"общественном транспорте" +Ульяновске*, полученного из исходного запроса Q = *Стоимость проезда в общественном транспорте в Ульяновске*, количество релевантных документов равно 8, а количество выданных документов – 8. В данном случае точность информационного поиска будет равна $P\left(Q^{Sem}\right) = \frac{8}{8} = 1.$

Таким образом, применение методов онтологического анализа и анализа текстов на естественном языке позволяют выделить в поисковых запросах значимые смысловые элементы запроса, а также расширить запрос терминами, характерными для некоторой ПрО, зафиксированной в виде онтологии в хранилище онтологий ПКИАД.

5 Организация подсистемы формирования социального портрета программного комплекса

Возможность формирования социального портрета пользователя социальной сети может оказаться полезной для следующих направлений:

- осуществление борьбы с проявлениями терроризма и экстремизма в сети как с целью выявления очагов зарождения подобных течений, так и в воспитательных и профилактических целях;
- формирование ориентированной на человека системы образования и здравоохранения за счет правильной подачи информации о здоровом образе жизни, культурных и социальных ценностях;
- проведение социологических исследований;
- кадровое планирование и др.

Подсистема формирования социального портрета пользователя включает в себя метод для определения предпочтений пользователя социальной сети «ВКонтакте».

Формально задача определения предпочтений сводится к классификации множества текстовых фрагментов:

$$D^{vk} = \{d_1, d_2, ..., d_n\}.$$

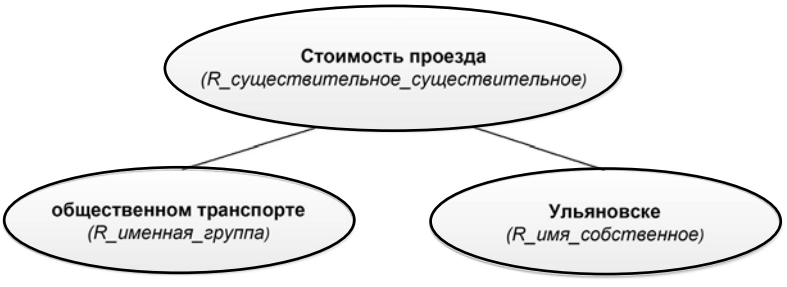

Рис. 4. Дерево значимых смысловых элементов запроса:
Стоимость проезда в общественном транспорте в Ульяновске

В качестве классификатора выступает множество категорий интересов пользователей C^{vk}, зафиксированное в виде фрагмента онтологии ПрО в хранилище онтологий ПКИАД.

Фрагмент онтологии ПрО для определения предпочтений пользователя имеет следующий вид:

$$G^{vk} = \left\langle C^{vk}, R_C^{vk}, I^{vk}, R_I^{vk} \right\rangle,$$

где C^{vk} – множество категорий интересов пользователя;

R_C^{vk} – множество отношений, определяющее иерархию категорий интересов;

$I^{vk} = \left\langle f_1, ..., f_l, ..., f_z \right\rangle$ – множество признаков категории интересов. Указанное множество признаков определяет словарь, который состоит из лексем, включающих слова и словосочетания, характеризующие категорию;

R_I^{vk} – множество отношений, определяющее связи между категориями и их признаками.

Задачей классификации является нахождение наиболее вероятной категории из множества C^{vk} для текстового фрагмента d_i. Предложенный метод классификации текстовых фрагментов основан на предположении, что тексты, относящиеся к одной категории, содержат одинаковые признаки (слова или словосочетания). Наличие или отсутствие таких признаков в текстовом фрагменте сигнализирует о его принадлежности или непринадлежности к той или иной теме.

Аналогично категориям каждый текстовый фрагмент также имеет признаки, по которым его можно отнести с некоторой степенью вероятности к одной или нескольким категориям:

$$I_{d_i} = \left\langle f_1, ..., f_l, ..., f_y^i \right\rangle.$$

Решение об отнесении текстового фрагмента d_i к категории C_r^{vk} принимается на основе пересечения:

$$I_{d_i} \cap I_{C_i^{vk}}^{vk}.$$

Метрика для расчета степени соответствия текстового входа (пост, комментарий) категории имеет вид:

$$Val_{ir} = \frac{\text{count}\left(I_{d_i} \cap I_{C_i^{vk}}^{vk}\right)}{\text{count}\left(I_{C_i^{vk}}^{vk}\right)}, \; Val_{ir} = [0..1] \qquad (5)$$

где $\text{count}\left(I_{d_i} \cap I_{C_i^{vk}}^{vk}\right)$ – количество совпавших атрибутов словарей I_{d_i} и $I_{C_i^{vk}}^{vk}$ соответственно;

$\text{count}\left(I_{C_i^{vk}}^{vk}\right)$ – количество атрибутов в словаре $I_{C_i^{vk}}^{vk}$.

В результате для каждого текстового входа формируется множество степеней его соответствия множеству категорий интересов C^{vk} следующего вида:

$$\delta(d_i) = \left\langle Val_{i1}, \; Val_{i2}, \; \ldots, \; Val_{im} \right\rangle$$

Для вычисления итогового значения степени принадлежности текстового фрагмента d_i категории интересов пользователя в процессе формирования социального портрета используется следующее выражение:

$$\mu_r = \frac{\sum_i^n \begin{cases} 1, & \max\left(\delta(d_i)\right) = Val_{ir} \\ 0, & \max\left(\delta(d_i)\right) \neq Val_{ir} \end{cases}}{n}, \qquad (6)$$

где n – количество текстовых фрагментов.

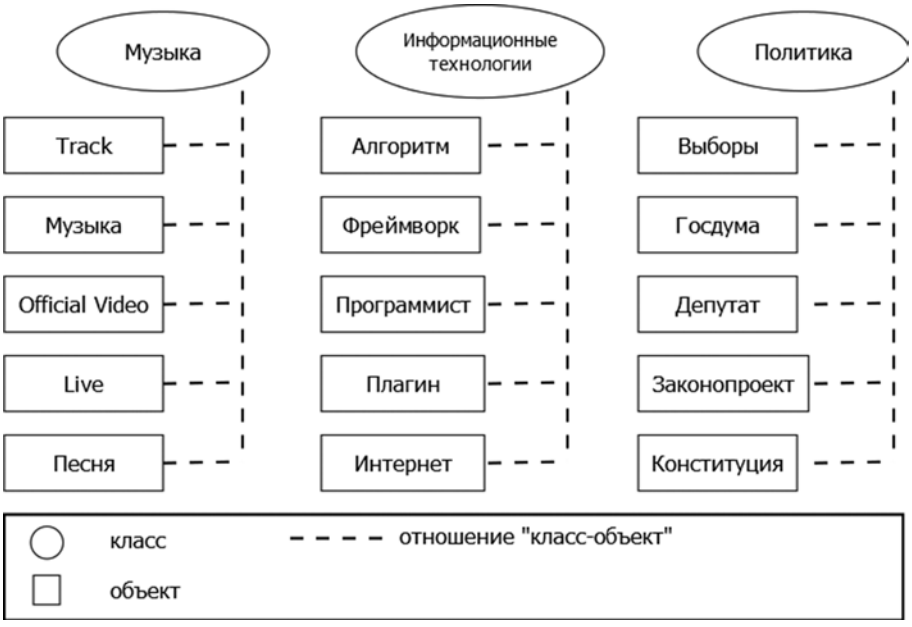

Рис. 5. Фрагмент онтологии, используемой для определения предпочтений пользователя социальной сети «ВКонтакте»

На рисунке 5 представлен фрагмент онтологии, используемой для определения предпочтений пользователя социальной сети «ВКонтакте».

Для определения степени принадлежности текстового фрагмента «Депутаты Госдумы приняли в первом чтении законопроект об изоляции российского сегмента интернета» к некоторой категории необходимо:

1. Выполнить разделение текста на слова с последующей лемматизацией каждого слова. Лемматизация – процесс приведения слова к его словарной форме, например, депутаты – депутат, интернета – интернет, изоляции – изоляция и т. д.

2. Используя выражение 5, определить степень соответствия

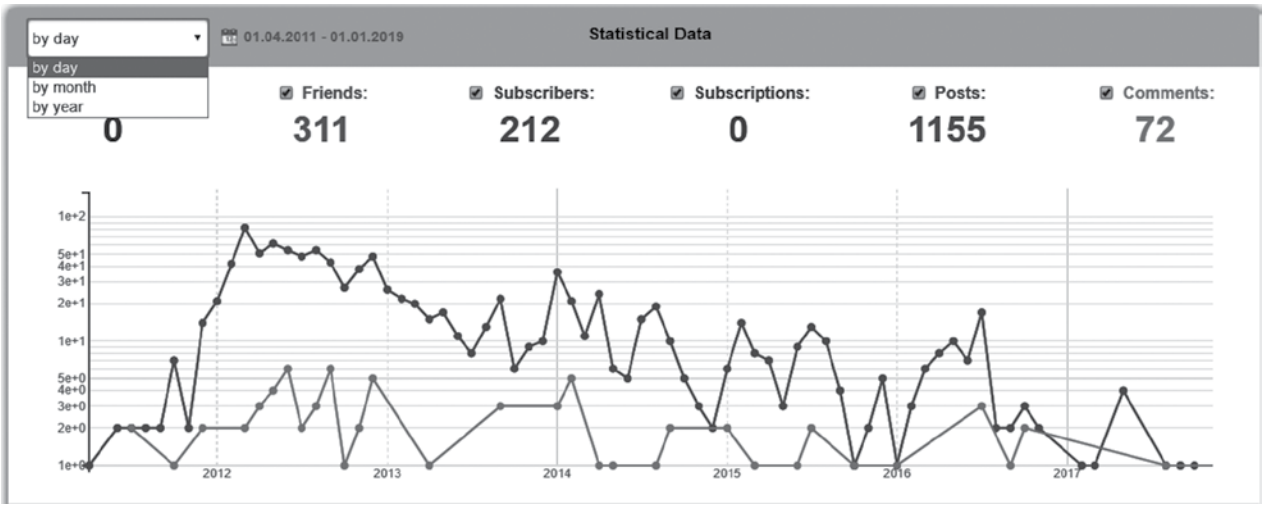

Рис. 6. График динамики активности пользователя

текстового входа каждой категории:

- $Val_{музыка} = \frac{0}{5} = 0,$
- $Val_{ИТ} = \frac{1}{5} = 0,2$ (интернет),
- $Val_{Политика} = \frac{3}{5} = 0,6$ (депутат, Госдума, законопроект).

Таким образом, данный текстовый фрагмент со степенью принадлежности 0,6 относится к категории «Политика» и степенью принадлежности 0,2 – к категории «Информационные технологии».

3. Используя выражение 6, определить значения степени принадлежности текстового фрагмента категории интересов пользователя:

- $\mu_{музыка} = 0,$
- $\mu_{ИТ} = 0,$
- $\mu_{политика} = 1.$

Следовательно, к предпочтениям пользователя скорее можно отнести категории «Политика».

Социальный портрет состоит из четырех разделов:

1. Информация о пользователе – содержит основные открытые данные со страницы пользователя.

2. Статистические данные – блок представляет собой график динамики активности пользователя: количество сообществ, друзей, подписчиков, подписок, постов, комментариев. Данные можно представить в разрезе дня, месяца и года (рис. 6).

3. Интересы пользователя – содержит результаты анализа постов и комментариев пользователя социальной сети с последующим отнесением к некоторой категории интересов (рис. 7).

4. Социальный граф пользователя – содержит данные о социальных связях пользователя: друг, подписчик, родственник, член сообщества и т. д.

Работа с социальными сетями может принести пользу при реализации функции системы управления персоналом и обеспечения комплексной безопасности компании, так как зачастую из социальных сетей о профессиональных и личностных качествах человека можно узнать больше, чем из его резюме.

Заключение

Представленный в данной работе ПКИАД позволяет загружать данные из социальной сети «ВКонтакте» и ЭСМИ.

В процессе загрузки данных из социальной сети «ВКонтакте» формируется граф социального взаимодействия, учитывающий следующие виды отношений: является другом, является подписчиком, является родственником, состоит в отношениях, состоит в сообществе. Также при загрузке данных средствами программной поисковой системы Elasticsearch формируется статистический индекс текстовых данных, а сами данные конвертируются в сущности модели данных подсистемы хранения ПКИАД и сохраняются в MongoDB.

С помощью подсистемы поиска данных существует возможность организации процесса поиска данных по ключевым словам в разрезе источников данных и видов сущностей: пользователи, сообщества, записи, комментарии и вложения. В процессе поиска исходный поисковый запрос пользователя может быть расширен и отформатирован с учетом семантических элементов на основе сведений, содержащихся в хранилище онтологий ПКИАД. Содержимое хранилища онтологий формируется в процессе трансляции онтологии в формате OWL/XML в узлы и отношения графовой базы данных Neo4j.

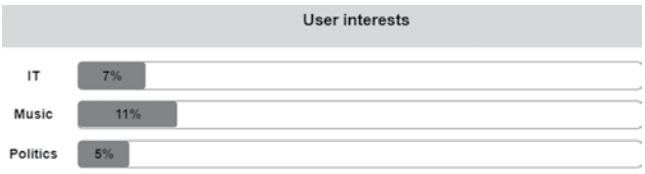

Рис. 7. Пример раздела с интересами пользователя

Возможность формирования социального портрета пользователя социальной сети может оказаться полезной для следующих направлений:

• осуществление борьбы с проявлениями терроризма и экстремизма в сети как с целью выявления очагов зарождения подобных течений, так и в воспитательных и профилактических целях;

• формирование ориентированной на человека системы образования и здравоохранения за счет правильной подачи информации о здоровом образе жизни, культурных и социальных ценностях;

• проведение социологических исследований;

• кадровое планирование и др.

В качестве дальнейшего развития ПКИАД планируются следующие действия:

1. Тестирование работы подсистемы хранения данных на больших объемах данных.

2. Разработка подсистемы сентимент-анализа данных.

3. Доработка пользовательского интерфейса.

Предполагается, что полученный в результате данного проекта программный комплекс должен повысить эффективность анализа содержимого социальных медиа с учетом специфики представления данных и нечеткости конструкций естественного языка.

СПИСОК ЛИТЕРАТУРЫ

1. Leskovec J., Faloutsos C. Sampling from large graphs // Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. 2006. pp. 631–636.

2. Practical recommendations on crawling online social networks / M. Gjoka et al. // Selected Areas in Communications, IEEE Journal. 2011. Vol. 29, Iss. 9. pp. 1872–1892.

3. Boyd D., Ellison N. Social network sites: Definition, history, and scholarship // Journal of Computer-Mediated Communication. 2007. Vol. 13, Iss. 1. pp. 210–230.

4. Pallis G., Zeinalipour-Yazti D., Dikaiakos M. Online Social Networks: Status and Trends // New Directions in Web Data Management 1, Studies in Computational Intelligence. 2011. Vol. 331. pp. 213–234.

5. Key Trends to Watch in Gartner 2012 Emerging Technologies Hype Cycle. – URL: http://www.forbes.com/sites/gartnergroup/2012/09/18/key-trends-to-watch-in-gartner2012-emerging-technologies-hype-cycle-2 (дата обращения: 01.05.2019).

6. Коршунов А. Задачи и методы определения атрибутов пользователей социальных сетей. – URL: http://ceur-ws.org/Vol-1108/paper23.pdf (дата обращения: 01.05.2019).

7. Определение демографических атрибутов пользователей микроблогов / А. Коршунов, И. Белобородов, А. Гомзин, К. Чуприна, Н. Астраханцев, Я. Недумов, Д. Турдаков. – URL: https://cyberleninka.ru/article/n/opredelenie-demograficheskih-atributov-polzovateley-mikroblogov (дата обращения: 01.05.2019).

8. Fleuret F. Fast Binary Feature Selection with Conditional Mutual Information. – URL: http://www.jmlr.org/papers/volume5/fleuret04a/fleuret04a.pdf (дата обращения: 01.05.2019).

9. Online Passive-Aggressive Algorithms / K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, Y. Singer // JMLR. 2006. pp. 551–585.

10. Pang B., Lee L., Vaithyanathan S. Thumbs up? Sentiment Classification using Machine Learning Techniques // In Proceedings of the ACL-02 conference on Empirical methods in natural language processing. 2002. Vol. 10. Association for Computational Linguistics, Stroudsburg, PA, USA. pp. 79–86.

11. Turney P. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews // In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Association for Computational Linguistics, Stroudsburg, PA, USA, 2002. pp. 417–424.

12. Цукерт А.Г. Проблемы и перспективы информационного поиска // Изв. Таганрог. гос. радиотехн. ун-та. – 2001. – Т. 21, № 3 (21). – С. 194–201.

13. Manning C., Schutze H. Foundations of Statistical Natural Language processing. The MIT Press Cambridge, MA, 1999.

14. Chetviorkin I., Loukachevitch N. Sentiment Analysis Track at ROMIP-2012. Computer linguistics and intellectual technologies // Computer linguistics and intellectual technologies: Dialogue-2013. 2013. Vol. 2. pp. 40–50.

15. Антонова А., Соловьев А. Использование метода условных случайных полей для обработки текстов на русском языке // Компьютерная лингвистика и интеллектуальные технологии: «Диалог-2013». – 2013. – Вып. 12, № 19. – С. 27–44.

16. Пазельская А., Соловьев А. Метод определения эмоций в текстах на русском языке. // Компьютерная лингвистика и интеллектуальные технологии: «Диалог-2011». – 2011. – Вып. 11, № 18. – С. 510–523.

17. García-Moya L., Anaya-Sanchez H., Berlanga-Llavori R. Retrieving product features and opinions from customer reviews // IEEE Intelligent Systems. 2013. Vol. 28, no 3. pp. 19–27.

18. Тарасов Д. Глубокие рекуррентные нейронные сети для аспектно-ориентированного анализа тональности отзывов пользователей на различных языках // По матер. ежегод. Междунар. конф. «Диалог». – 2015. – Вып. 14, № 21, Т. 2. – С. 53–64.

19. Resource Description Framework (RDF). – URL: https://www.w3.org/RDF (дата обращения: 01.05.2019).

20. OWL Web Ontology Language Overview. – URL: https://www.w3.org/TR/owl-features.

21. The OWL API. – URL: http://owlcs.github.io/owlapi (дата обращения: 01.05.2019).

22. Owlcpp: a C++ library for working with OWL ontologies. – URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4574266 (дата обращения: 01.05.2019).

23. The Knowledge Graph Platform for the Enterprise. – URL: https://www.stardog.com (дата обращения: 01.05.2019).

24. About OpenLink Virtuoso. – URL: https://virtuoso.openlinksw.com (дата обращения: 01.05.2019).

25. Eclipse RDF4J. – URL: http://rdf4j.org (дата обращения: 01.05.2019).

26. The Heart of the Elastic Stack. – URL: https://www.elastic.co/products/elasticsearch (дата обращения: 01.05.2019).

27. MongoDB. For Giant ideas. – URL: https://www.mongodb.com (дата обращения: 01.05.2019).

28. Introducing the Neo4j Graph Platform. – URL: https://neo4j.com (дата обращения: 01.05.2019).

29. Токмаков Г.П. Онтологии и их применение для интеграции информационных ресурсов // Автоматизация процессов управления. – 2010. – № 1 (19). – С. 37–49.

30. Субхангулов Р.А. Онтологический поиск технических документов на основе модели интеллектуального агента // Автоматизация процессов управления. – 2014. – № 4 (38). – С. 85–91.

31. Elasticsearch Query DSL. – URL: https://www.elastic.co/guide/en/elasticsearch/reference/2.3/query-dsl-query-string-query.html (дата обращения: 01.05.2019).

REFERENCES

1. Leskovec J., C. Faloutsos. Sampling from Large Graphs. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM.* 2006, pp. 631–636.

2. M. Gjoka et al. Practical Recommendations on Crawling Online Social Networks. *Selected Areas in Communications, IEEE Journal,* 2011, vol. 29, iss. 9, pp. 1872–1892.

3. Boyd D., Ellison N. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication,* 2007, vol. 13, iss. 1, pp. 210–230.

4. Pallis G., Zeinalipour-Yazti D., Dikaiakos M. Online Social Networks: Status and Trends. *New Directions in Web Data Management 1, Studies in Computational Intelligence,* 2011, vol. 331, pp.213–234.

5. Key Trends to Watch in Gartner 2012 Emerging Technologies Hype Cycle. Available at :http://www.forbes.com/sites/gartnergroup/2012/09/18/key-trends-to-watch-in-gartner2012-emerging-technologies-hype-cycle-2 (accessed: 01.05.2019).

6. Korshunov A. *Zadachi i metody opredeleniia atributov polzovatelei sotsialnykh setei* [Problems and Methods for detecting attributes of social network users]. Available at: http://ceur-ws.org/Vol-1108/paper23.pdf (accessed: 01.05.2019).

7. Korshunov A., Beloborodov I., Gomzin A., Chuprina K., Astrakhantsev N., Nedumov Ia., Turdakov D. *Opredelenie demograficheskikh atributov polzovatelei mikroblogov* [Definition of Demographic Attributes of Microblogs Users]. Available at: https://cyberleninka.ru/article/n/opredelenie-demograficheskih-atributov-polzovateley-mikroblogov (accessed: 01.05.2019).

8. Fleuret F. *Fast Binary Feature Selection with Conditional Mutual Information.* Available at: http://www.jmlr.org/papers/volume5/fleuret04a/fleuret04a.pdf (accessed: 01.05.2019).

9. Crammer K., O. Dekel, J. Keshet, S. Shalev-Shwartz, Y. Singer. Online Passive-Aggressive Algorithms. *JMLR,* 2006, pp. 551–585.

10. Pang B., Lee L., Vaithyanathan S. Thumbs up? Sentiment Classification Using Machine Learning Techniques. *In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, vol. 10, pp. 79–86.

11. Turney P. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02).* Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, pp. 417–424.

12. Tsukert A.G. Problemy i persektivy informatsionnogo poiska [Problems and Prospects of Information Search]. *Izv. Taganrog. gos. radiotekhn. un-ta* [Proc. of Taganrog State Radioeng. University], 2001, vol. 21, no. 3 (21), pp. 194–201.

13. Manning C., Schutze H. *Foundations of Statistical Natural Language Processing.* The MIT Press Cambridge, MA, 1999.

14. Chetviorkin I., Loukachevitch N. Sentiment Analysis Track at ROMIP-2012. Computer Linguistics and Intellectual Technologies. *Computer Linguistics and Intellectual Technologies: Dialogue-2013.* 2013, vol. 2. pp. 40–50.

15. Antonov A., Solovev A. Ispolzovanie metoda uslovnykh sluchainykh polei dlia obrabotki tekstov na russkom iazyke [Conditional Random Field Models for the Processing of Russian]. *Kompiuternaia lingvistika i intellektualnye tekhnologii: Dialog-2013* [Computer Linguistics and Intellectual Technologies: Dialogue-2013]. 2013, iss. 12, no. 19, pp. 27–44.

16. Pazelskaia A., Solovev A. Metod opredeleniia emotsii v tekstakh na russkom iazyke [A Method of Sentiment Analysis in Russian Texts]. *Komputernaia lingvistika i intellektualnye tekhnologii: Dialog-2011* [Computer Linguistics and Intellectual Technologies: Dialogue-2013]. 2011, iss. 11, no. 18, pp. 510–523.

17. García-Moya L., Anaya-Sanchez H., Berlanga-Llavori R. Retrieving Product Features and Opinions from Customer Reviews. *IEEE Intelligent Systems,* 2013, vol. 28, no. 3, pp. 19–27.

18. Tarasov D. Glubokie rekurrentnye neironnye seti dlia aspektno-orientirovannogo analiza tonalnosti otzyvov polzovatelei na razlichnykh iazykakh [Deep Recurrent Neural Networks for Multiple Language Aspect-based Sentiment Analysis of User Reviews]. *Po materialam ezhegodnoi Mezhdunarodnoi konferentsii 'Dialog'* [Materials of Annual Int. Sci. Conf. 'Dialogue']. 2015, iss. 14, no. 21, vol. 2, pp. 53–64.

19. Resource Description Framework (RDF). Available at: https://www.w3.org/RDF (accessed: 01.05.2019).

20. *OWL Web Ontology* Language Overview. Available at: https://www.w3.org/TR/owl-features.

21. The OWL API. Available at: http://owlcs.github.io/owlapi (accessed: 01.05.2019).

22. *Owlcpp: a C++ Library for Working with OWL Ontologies.* Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4574266 (accessed: 01.05.2019).

23. *The Knowledge Graph Platform for the Enterprise.* Available at: https://www.stardog.com (accessed: 01.05.2019).

24. *About OpenLink Virtuoso.* Available at: https://virtuoso.openlinksw.com (accessed: 01.05.2019).

25. *Eclipse RDF4J.* Available at: http://rdf4j.org (accessed: 01.05.2019).

26. *The Heart of the Elastic Stack.* Available at: https://www.elastic.co/products/elasticsearch (accessed: 01.05.2019).

27. *Mongo DB. For Giant Ideas.* Available at: https://www.mongodb.com (accessed: 01.05.2019).

28. *Introducing the Neo4j Graph Platform.* Available at: https://neo4j.com (accessed: 01.05.2019).

29. Tokmakov G.P. Ontologii i ikh primenenie dlia integratsii informatsionnykh resursov [Ontology and its Application for Information Resources Integration]. *Avtomatizatsiia protsessov upravleniia* [Automation of Control Processes], 2010, no. 1 (19), pp. 37−49.

30. Subkhangulov R.A. Ontologicheskii poisk tekhnicheskikh dokumentov na osnove modeli intellektualnogo agenta [Ontological Retrieval of Technical Documents Based on an Intelligent Agent Model]. *Avtomatizatsiia protsessov upravleniia* [Automation of Control Processes], 2014, no. 4 (38), pp. 85−91.

31. *Elasticsearch Query DSL.* Available at: https://www.elastic.co/guide/en/elasticsearch/reference/2.3/query-dsl-query-string-query.html (accessed: 01.05.2019).