

The Social Portrait Building of a Social Network User Based on Semi-Structured Data Analysis

N.Yarushkina, A.Filippov, V.Moshkin, A.Namestnikov, G.Guskov

Ulyanovsk state technical university, Ulyanovsk, Russia
e-mail: {jng, al.filippov, v.moshkin, nam, g.guskov}@ulstu.ru

Abstract. The article presents the method of constructing a social portrait of a social network user. This technique is implemented in the module of system for opinion mining. The method of building a social portrait is based on collecting statistical information about the user, the dynamics of his activity and on the semantic analysis of the subject of his posts and comments using linguistic ontology.

1 Introduction

Active growth of social media audience on the Internet (social networks, forums, blogs and online media) made them a new source of data and knowledge. The specifics of working with social media has several advantages and disadvantages.

Advantages include:

- high speed of access to information;
- a broad audience;
- a wide range of data topics;
- large amount of data.

The disadvantages are:

- large amount of data;
- unstructured presentation of information;
- absence of a single conceptual framework.

A large amount of social media data is both an advantage and a disadvantage at the same time. Monthly in Russian social networks about 30 million unique authors publish 580 billion messages according to statistics for 2018.

However, a large amount of data makes it possible to obtain a large training sets, for machine learning methods and a large statistical sample for social studies.

The monthly billions of unstructured text messages and publications that users leave monthly cannot be processed manually.

There is a need for methods of automated intelligent and sentimental analysis of text data. These methods handle large amounts of data and understand their meaning

(Text Mining), determine the sentiment (Opinion Mining) of user messages and publications in a short time [1-5].

Understanding the meaning and sentiment of publications in social media is the most important and complex element of automated text processing [6-11].

Our scientific group has created an intelligent tool for Opinion Mining of social media. This tool includes new approaches to the hybridization of ontological analysis and methods of knowledge engineering with methods of nature language processing (NLP) for extracting the semantic and emotional component of semi-structured and unstructured text resources [12-16].

These approaches will improve the efficiency of the analysis of social media content-specific data and fuzziness of natural language.

The implementation of the model and methodology for obtaining a social portrait of a social network user will be considered in this paper. This model was developed on the example of the most popular social network in Russia VKontakte [17]. The number of users of this social network exceeded 528 million people at the time of publication of the article.

It is necessary to use intelligent data analysis systems to automate the process of analyzing the target audience. Large amounts of data, a variety of forms of their presentation and their unstructured presentation do not allow for quickly building a social portrait of a potential user of a company's product.

This problem increases the time frame for analyzing requirements, ideas, competitors, and target audiences. The ability to form a social portrait of a social network user is useful for:

- the fight against terrorism and extremism in the social network;
- the construction of a person-oriented education and health care system through the correct presentation of information about a healthy lifestyle, cultural and social values;
- sociological research;
- workforce planning, etc.

2 The architecture of System for Opinion Mining

Service-oriented approach is the basis of the architecture of the software system for Opinion Mining Social Media (SOM). This approach allows:

- Increase the overall fault tolerance of the SOM by performing services in different address spaces.
- Increase the scalability of the SOM by running several instances of services and balancing the load between them.
- Provide the ability to use different operating systems, programming languages, storage technologies, etc.
- Reduce the downtime of SOM when making changes, correcting errors, etc.
- Provide an opportunity to completely replace services while maintaining the interface of interaction with other parts of the SOM.

REST in conjunction with the HTTP protocol [18] is the basis for the organization of the interface for the interaction of SOM services. REST allows a distributed system of any type to have the following properties: performance, extensibility, simplicity, updatability, intelligibility, portability and reliability.

The architecture of SOM is shown in Figure 1.

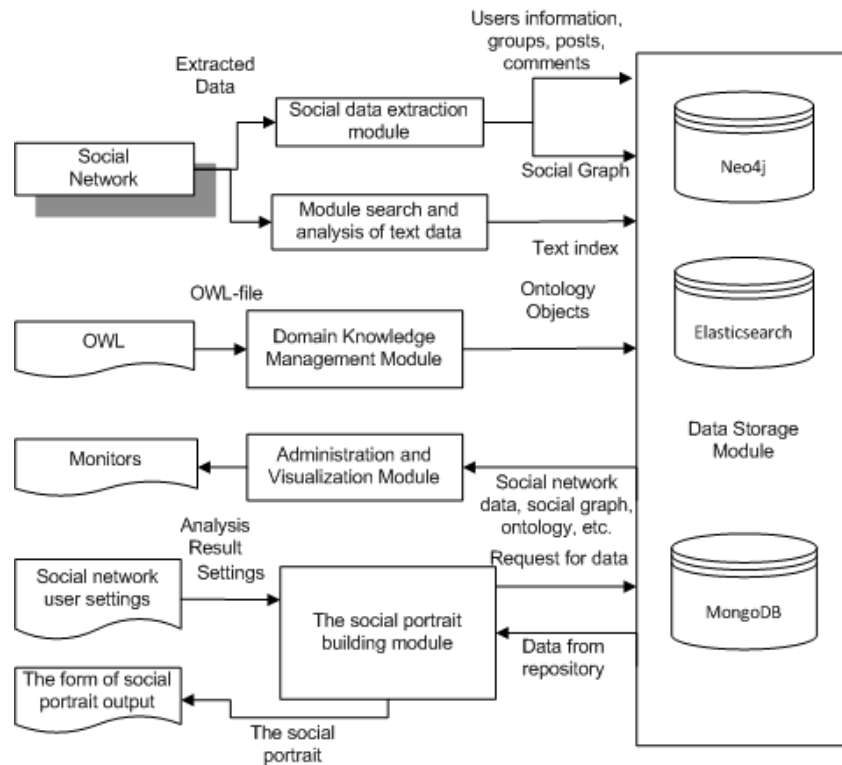


Figure 1. The architecture of SOM

1. Social data extraction module is a subsystem for importing data from social media. This subsystem works with social networks through the public application programming interface (Public API).

2. The data storage module provides the representation of information extracted from social networks in a unified structure that is convenient for further processing. The data is stored in the context of users, collections, data sources, versions, etc. As database management systems (DBMS) are used:

- Elasticsearch for indexing and retrieving data [19];
- MongoDB for storing data in JSON format [20];
- Neo4j for storing graphs of social interaction (social graph) and ontology [21].

3. Domain knowledge Management module translates OWL / RDF-ontology into the graph knowledge base [22].

4. Administration and visualization module manages user rights, tasks, provides a friendly system interface, and displays the necessary reports on the analysis of social network data.

5. Module search and analysis of text data performs preprocessing of text resources using statistical and linguistic methods. Also this module searches for objects related to a specific task. The task is presented in the form of a set of keywords. In this case, the user's query can be extended semantically using an ontology. Ontology contains descriptions of features of the PrA.

6. The social portrait building module performs semantic analysis of posts and user comments and collects statistics of its activity. The module work is described in more detail in the following chapters of the article.

3 Formal model of the social network user

A. Statistical social portrait of a social network user

Formally the statistical social portrait of the social network user is:

$$P = \{MI, S, G\},$$

where **MI** is user meta information; **S** – statistical information about the user; **G** – is the user's social graph.

The following expression is typical for any user of a social network:

$$\forall i \mathbf{p}_i \in \mathbf{P}: \mathbf{p}_i = \{< \mathbf{M}_i, \mathbf{S}_i, \mathbf{G}_i >\},$$

where **i** is a specific user ID.

Let us consider in more detail the components \mathbf{p}_i . Formally the attributes that determine the meta information \mathbf{M}_i of the *i*-th user are

$$\mathbf{M}_i = \left\{ \begin{array}{l} \mathbf{p}_i^{vk}, \mathbf{sa}_i^{vk}, \mathbf{wp}_i^{vk}, \mathbf{st}_i^{vk}, \mathbf{vf}_i^{vk}, \mathbf{na}_i^{vk}, \\ \mathbf{se}_i^{vk}, \mathbf{kon}_i^{vk}, \mathbf{int}_i^{vk}, \mathbf{ed}_i^{vk}, \mathbf{car}_i^{vk}, \mathbf{mil}_i^{vk} \end{array} \right\}$$

where

\mathbf{p}_i^{vk} is the social network Vkontakte page of the *i*-th user;

\mathbf{sa}_i^{vk} is the short address of the *i*-th user page;

\mathbf{wp}_i^{vk} is the address of the *i*-th user's wall page;

$\mathbf{st}_i^{vk} \in \{0, 1\}$ is account status (0 - account is inactive, 1 - account is active);

$\mathbf{vf}_i^{vk} \in \{0, 1\}$ is a page verification (0 – no, 1 – yes);

\mathbf{na}_i^{vk} are *i*-th user's first name, last name and nickname;

$\mathbf{se}_i^{vk} \in \{f, m\}$ – is the user's gender (f is female, m is male);

\mathbf{kon}_i^{vk} is the contact list of the *i*-th user;

\mathbf{int}_i^{vk} is a set of interests of the *i*-th user (may be categorical);

\mathbf{ed}_i^{vk} is the information about the education received by the *i*-th user;

\mathbf{car}_i^{vk} is the information about the *i*-th user career;

\mathbf{mil}_i^{vk} is the information about the military service of the *i*-th user.

Statistical information about the social network user S_i is formed in various time sections with an indication of the time period:

$$S_i = \langle \{ \text{per}_j^{\text{vk}}, \text{sel}_k^{\text{vk}} \}, \{ \langle \text{par}_s^{\text{vk}}, \text{val}_s^{\text{vk}} \rangle \}_{s=1, |\text{par}^{\text{vk}}|} \rangle_{j=1, n; k=1, l}.$$

$\{ \text{per}_j^{\text{vk}}, \text{sel}_k^{\text{vk}} \}$ is a time section and a related period of time; n is the number of time sections; l is the number of time periods for analyzing the activity of a social network user;

The analysis is performed on a set of indicators. These indicators are presented as a set:

$$\text{par}^{\text{vk}} = \{ \text{ar}, \text{fr}, \text{subp}, \text{sub}, \text{pst}, \text{kom} \},$$

where

- ar** – the number of communities the user belongs to;
- fr** – the number of friends of a social network user;
- subp** – is the number of subscribers for the user;
- sub** – is the number of subscriptions a user has;
- pst** – is the number of user posts;
- kom** – is the number of user comments.

Each indicator corresponds to the value val_s^{vk} which characterizes its sum:

$$\text{val}_s^{\text{vk}} = \sum \text{item}_s.$$

The structure of the social relations of the analyzed user is represented as a graph:

$$\mathbf{G} = (\mathbf{V}^{\text{vk}}, \mathbf{E}^{\text{vk}}),$$

where \mathbf{V}^{vk} – is a finite set of social graph vertices; \mathbf{E}^{vk} is a finite set of edges defining pairs of adjacent identifiers of social network users.

The set of vertices of the social graph is represented as the union of the singleton set. This set contains the identifier of the analyzed social network user with set of his friends:

$$\mathbf{V}^{\text{vk}} = \{ \mathbf{v}_{\text{item}}^{\text{vk}} \} \cup \{ \mathbf{v}_1^{\text{vk}}, \mathbf{v}_2^{\text{vk}}, \dots, \mathbf{v}_p^{\text{vk}} \},$$

where p is the number of friends of the network user.

Formally set of edges is:

$$\mathbf{E}^{\text{vk}} = \{ \langle \mathbf{v}_{\text{item}}^{\text{vk}}, \mathbf{v}_1^{\text{vk}} \rangle, \langle \mathbf{v}_{\text{item}}^{\text{vk}}, \mathbf{v}_2^{\text{vk}} \rangle, \dots, \langle \mathbf{v}_{\text{item}}^{\text{vk}}, \mathbf{v}_p^{\text{vk}} \rangle \}.$$

B. Semantic representation of the social portrait of a social network user

The task of building a social portrait of a social network user is the task of classifying a set of users by classifying text fragments (social network posts, comments) of a specific user or his friends. Classes are categories of social network user interests. These categories may include topics related to subject areas: sports, IT-technologies, music, business and others.

Formally the task of classifying text fragments is described by a set of text fragments sets:

$$\mathbf{D}^{\text{vk}} = \{ \mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n \}.$$

User interest categories are set:

$$\mathbf{C}^{\text{vk}} = \{ \mathbf{c}_r \}, \text{ where } r = 1, \dots, m.$$

A hierarchy of categories will represent this set of pairs. This set of pairs determines the relationship of nesting between rubrics:

$$\mathbf{H}^{\text{vk}} = \{ \langle \mathbf{c}_j, \mathbf{c}_p \rangle, \mathbf{c}_j, \mathbf{c}_p \in \mathbf{C}^{\text{vk}} \}$$

(the category \mathbf{c}_p is nested in the category \mathbf{c}_j).

The hierarchy of categories is formed as domain ontology and each category is represented as a class (concept). In the classification problem it is necessary to build a procedure based on this data. The procedure should find the most likely category from the set \mathbf{C}^{vk} for the text fragment \mathbf{d}_i .

Our method for classifying text fragments is based on the assumption that texts belonging to the same category contain the same attributes (words or phrases). The presence or absence of such attributes of the text fragment indicates its belonging or non-belonging to one or another topic.

For each category a set of attributes is:

$$\mathbf{F}^{vk}(\mathbf{C}^{vk}) = \cup (\mathbf{c}_r),$$

where $\mathbf{F}^{vk}(\mathbf{c}_r) = \langle \mathbf{f}_1, \dots, \mathbf{f}_1, \dots, \mathbf{f}_z \rangle$.

The specified set of attributes defines the dictionary. This dictionary consists of tokens, including words and phrases characterizing the category. This dictionary is considered as the linguistic basis of the ontological resource of the developed system.

Each text fragment also has attributes similar to topics or categories. A fragment of these attributes can be attributed to one or more categories with some degree of probability:

$$\mathbf{F}(\mathbf{d}_i) = \langle \mathbf{f}_1^i, \dots, \mathbf{f}_1^i, \dots, \mathbf{f}_y^i \rangle.$$

The set of all text fragments attributes should be equal to the set of attributes of interests categories of social network users, ie:

$$\mathbf{F}^{vk}(\mathbf{C}^{vk}) = \mathbf{F}(\mathbf{D}) = \cup \mathbf{F}(\mathbf{d}_i).$$

The decision to classify the text fragment \mathbf{d}_i as \mathbf{c}_r is made on the basis of the intersection:

$$\mathbf{F}(\mathbf{d}_i) \cap \mathbf{F}^{vk}(\mathbf{c}_r).$$

The category of a specific social network user is determined on the basis of a numerical indicator that aggregates the values of text fragments of posts and user comments.

Formally the metric for calculating the degree of conformity of the text (post, comment) to the description of the area of interest from the ontology is:

$$\mathbf{Val}_{ir} = \frac{\text{count}(\mathbf{F}(\mathbf{d}_i) \cap \mathbf{F}^{vk}(\mathbf{c}_r))}{\text{count}(\mathbf{F}^{vk}(\mathbf{c}_r))}, \mathbf{Val}_{ir} = [0..1],$$

where $\text{count}(\mathbf{F}(\mathbf{d}_i) \cap \mathbf{F}^{vk}(\mathbf{c}_r))$ – is the number of matched attributes of the dictionaries $\mathbf{F}(\mathbf{d}_i)$ and $\mathbf{F}^{vk}(\mathbf{c}_r)$ respectively; $\text{count}(\mathbf{F}^{vk}(\mathbf{c}_r))$ – is the number of attributes in the dictionary $\mathbf{F}^{vk}(\mathbf{c}_r)$.

A set of degrees of correspondence of a text fragment to a set of categories of interests \mathbf{C}^{VK} formed:

$$\boldsymbol{\delta}(\mathbf{d}_i) = \langle \mathbf{Val}_{i1}, \mathbf{Val}_{i2}, \dots, \mathbf{Val}_{im} \rangle$$

The following expression is used to calculate the severity of the user's interest category in the process of forming a social portrait:

$$\boldsymbol{\mu}_r = \frac{\sum_i^n (1, \max(\langle \mathbf{Val}_{i1}, \mathbf{Val}_{i2}, \dots, \mathbf{Val}_{im} \rangle) = \mathbf{Val}_{ir})}{\sum_i^n (0, \max(\langle \mathbf{Val}_{i1}, \mathbf{Val}_{i2}, \dots, \mathbf{Val}_{im} \rangle) \neq \mathbf{Val}_{ir})},$$

where n is the number of text fragments; m - the number of categories.

C. Building a social portrait of a social network user

The social portrait model of a social network user is a SOM module. Experiments on the construction of a social portrait were carried out on the open data of users of the social network VKontakte.


The constructed social portrait consists of four sections:

- User information.
- Statistical data.
- The interests of the user and user friends.
- Social graph.

The first block contains the main public data from the user's page (Fig.2).

User information

Version from: 23.01.2019 17:37 [Update](#)


User Name: Pavel Novikov ✓

<p>Page: https://vk.com/id12827954</p> <p>Short Link: https://vk.com/mellvin</p> <p>User wall: https://vk.com/wall12827954</p> <p>Account Status: active</p> <p>Page Verification: <input type="checkbox"/></p> <p>Name Lastname Nickname: Pavel Noviko</p> <p>Sex: Man</p> <p>Contacts:</p> <p>Web-site: http://scaleofuniverse.com/</p> <p>Skype: nokerakuta</p>	<p>Interests:</p> <p>Activity: 1C programmer, accountant</p> <p>Interests: Bicycles, historical fencing, computers, board games, anime, books, archery, pneumatics, music</p> <p>The favorite music: Electronic, Experimental, Dubstep, Glitch, 8bit, Neo-Classical, Classical, New-Age, Jazz, Post-Rock, Ambient, Trance, Alternative Rock, Rock, Heavy Metal, Gothic Metal, Progressive Metal</p> <p>Favorite books: science fiction, fantasy</p> <p>Favorite games: board games, Minecraft, Transport Tycoon</p> <p>Education:</p> <p>School: Russia, Ulyanovsk, School №28, Not specified</p> <hr/> <p>School: Russia, Ulyanovsk, School №76, Not specified</p> <hr/> <p>University: Russia, Moscow, Ulyanovsk State Technical university</p> <p>Faculty: Faculty of Information Systems and Technology</p> <p>Department: Information Systems</p> <hr/> <p>Career: missing</p> <p>Military Service: missing</p>
---	--

Figure 2. User information

This block contains a list of interests of the user, his education, career, etc.

The second block is a graph of the dynamics of user activity:

- Communities
- Friends
- Subscribers
- Subscriptions
- Posts
- Comments (Fig.3)

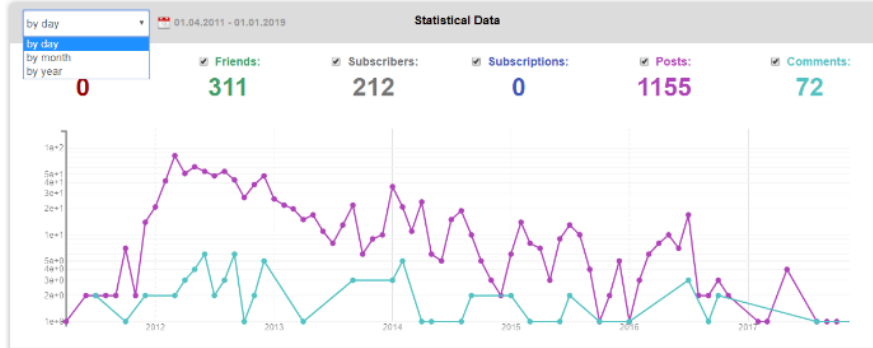


Figure 3. Statistical Data

Data can be presented by day, month and year.

The third block contains the results of semantic analysis of social network user posts and comments (Fig.4). The calculation model is presented in paragraph 3.

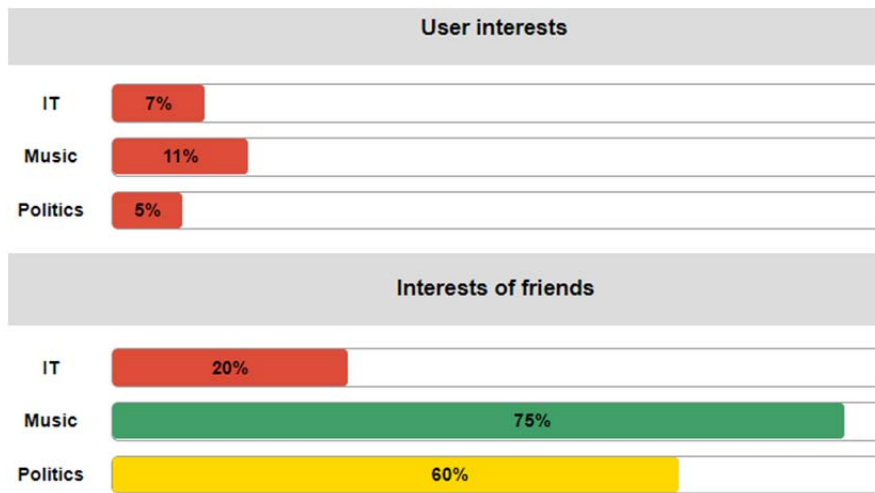


Figure 4. User interests and user friends interests

The fourth block of the social portrait is a social graph. The social graph contains data about the social network users associated with a specific user of various types of connections: friend, follower, etc. (Fig.5).

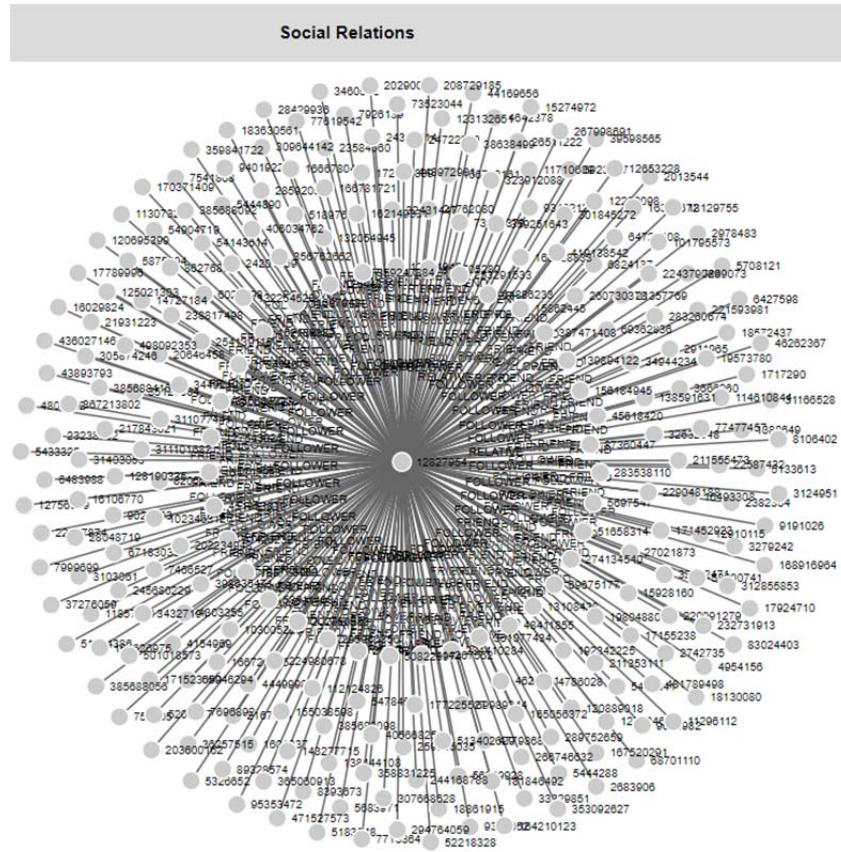


Figure 5. Social graph

Conclusion

Data analysis of social networks can be useful in the management of company personnel, since it is often possible to learn more from social networks about a person's professional and personal qualities than from his resume.

The analysis of a person's interests as well as his psycho-physiological characteristics is important from the point of view of ensuring the integrated safety of the organization's functioning.

The developed algorithm for the formation of a social portrait in the framework of SOM will allow HR specialists of any company, which first of all need this system, to quickly get an objective understanding of the personal, psycho-physiological and business qualities of a person. The use of this system will reduce the company's risks associated with the work of the specialists involved.

Acknowledgements. This study was supported by the Russian Foundation for Basic Research (Grants No. 18-47-732007 and 18-47-730035).

References

1. Leskovec J., Faloutsos C. Sampling from large graphs //Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. pp. 631-636 (2006).
2. Gjoka M. et al. Practical recommendations on crawling online social networks //Selected Areas in Communications, IEEE Journal on. Vol. 29. №. 9. pp. 1872-1892 (2011).
3. Boyd D., Ellison N. Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication. Vol. 13(1). article 11. (2007)
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 Volume 331, pp 213-234 (2011).
5. Key Trends to Watch in Gartner 2012 Emerging Technologies Hype Cycle. <http://www.forbes.com/sites/gartnergroup/2012/09/18/key-trends-to-watch-in-gartner2012-emerging-technologies-hype-cycle-2>, last accessed 2019/08/08.
6. Korshunov A. Tasks and methods for determining the attributes of users of social networks // Proceedings of the 15th All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" - RCDL'2013
7. Korshunov A., Beloborodov I., Gomzin A., Chuprina K., Astrakhantsev N., Nedumov J., Turdakov D. Determination of demographic attributes of users of microblogging // Proceedings of the Institute of System Programming of RAS. Vol. 25, 2013 DOI : 10.15514 / ISPRAS-2013-25-10.
8. Fleuret F. Fast Binary Feature Selection with Conditional Mutual Information // JMLR, 5:1531–1555 (2004).
9. Crammer K., Dekel O., Keshet J., Shalev-Shwartz S., Singer Y. Online Passive-Aggressive Algorithms // JMLR, 7(Mar):551–585 (2006).
10. Pang B., Lee L., Vaithyanathan S. Thumbs up? Sentiment Classification using Machine Learning Techniques. pp. 79–86 (2002).
11. Turney P. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews // Proceedings of the Association for Computational Linguistics. pp. 417–424. arXiv: LG/0212032 (2002)
12. Chetviorkin I., Loukachevitch N. Sentiment Analysis Track at ROMIP-2012. Computer linguistics and intellectual technologies. Computer linguistics and intellectual technologies: Dialogue-2013. Sat. scientific articles volume 2, p. 40-50.
13. Antonova A., Soloviev A., Using the method of conditional random fields for processing texts in Russian. Computer linguistics and intellectual technologies: Dialogue-2013. Sat. scientific articles / Issue. 12 (19)- Moscow: Publishing house of the RSUH. pp.27-44 (2013).
14. Pazelskaya A., Soloviev A. Method of definition of emotions in texts in Russian. Computer linguistics and intellectual technologies. Computer linguistics and intellectual technologies: Dialogue-2011. Sat. scientific articles / Issue. 11 (18). Moscow: Publishing House of the RSUH. pp. 510-523 (2011).
15. García-Moya, L., Anaya-Sánchez, H., Berlanga-Llavori, R.: Retrieving product features and opinions from customer reviews. IEEE Intelligent Systems 28(3), pp. 19–27 (2013)

16. Tarasov D. Deep Recurrent Neural Networks for Multiple Language Aspect-Based Sentiment Analysis // Computational Linguistics and Intellectual Technologies: Proceedings of Annual International Conference “Dialogue-2015”. Issue 14(21), Vol.2, pp. 65-74 (2015).
17. Representational state transfer, https://en.wikipedia.org/wiki/Representational_state_transfer, last accessed 2019/08/08.
18. Social Network VKontakte <https://vk.com> last accessed 2019/08/08
19. The Heart of the Elastic Stack, <https://www.elastic.co/products/elasticsearch>, last accessed 2019/08/08.
20. MongoDB. For Giant ideas, <https://www.mongodb.com>, last accessed 2019/08/08.
21. Introducing the Neo4j Graph Platform, <https://neo4j.com>, last accessed 2019/08/08.
22. Yarushkina N., Filippov A., Moshkin V. Development of the unified technological platform for constructing the domain knowledge base through the context analysis. Communications in Computer and Information Science. 2017. Vol. 754. pp. 62-72.