The analysis method of primary data for monitoring social processes using Big Data and fuzzy criteria

Alexey Y. Timonin
Postgraduate
Penza State University
40, Krasnaya street, Penza, Russia 440026
c013s017b301f018@mail.ru

Alexander M. Bershadsky
Doctor of technical sciences, professor
Penza State University
40, Krasnaya street, Penza, Russia 440026
bam@pnzgu.ru

Alexander S. Bozhday
Doctor of technical sciences, professor
Penza State University
40, Krasnaya street, Penza, Russia 440026
bozhday@yandex.ru

Abstract: The presented work deals with mathematical models' development for determining the significance degree of the social information. It contained in unstructured raw data from public web-sources. The significance criteria of social environment descriptions based on fuzzy logic and probability theory methods. Also, the set theory apparatus used to represent the extracted social objects. Results implementation allows to improve the data relevance and reliability determining in the social profiling system based on Big Data.

Keywords: Big Data, data analysis, evaluation, fuzzy logic, membership function, probability theory, set theory, social management, social profiling, unstructured data.

1 Introduction

Modeling and analysis of social processes is currently based both on the classical techniques of sociology and various theories of computer science [14]. The apparatus of graph theory, optimization methods, processing methods for scale-invariant [2] and "small-world" [18] networks are used to describe the society structure. In different work, analysts are increasingly resorting to collecting initial information from electronic sources of social media. This data is high available and complete. Big Data, Data Mining, Machine Learning and visual analytics techniques are used for processing such content. The society analysis is fraught with the difficulties of related activities. Some of them are the introducing information technologies and developing new associated scientific approaches, ensuring the confidentiality of the processed information, as well as obtaining timely and reliable results. The task of social profiling is no exception.

The social profile (SP) [3, 4] is a heterogeneous semantic network consisting of socialized data with varying structure degrees. That data is defining a particular person or social groups. The static part of the social profile is a structured set that uniquely identifies the selected subject from the social environment. The SP static part named the information card [4]. Its elements are verified definitions presented in verbal form. The SP dynamic part consists of heterogeneous information (such as text, multimedia, etc.) about personalities, social groups and phenomena. Dynamic content [4] is used to determine

implicit dependencies and expand the information card. Social profiles can be applied [3] in the tasks of assessing the social tension level and response to socially significant events. Other applied spheres [3] are designing a customer reliability rating for insurance and banking companies, compiling an employee's portfolio for human resources departments, issuing medical recommendations based both on lifestyle data and the environment states, etc.

Therefore, it is critical to identify discrepancies in the collected information and determine the data relevance degree. Results should be divided by importance in a particular research task. The initial SP information is taken from various sources and has a significant size (up to several terabytes). There is a problem of determining its reliability in a finite time. It is also impossible to exclude the human factor influence from the manual process of data`s social values. It is necessary to use automated tools to solve these problems. The instruments should be based on formalized mathematical criteria for the importance of social data.

2 Background

Studies of textual data have been conducted in order to automatically determine their relevance and consistency by the scientific community for a long time. Currently, Text Mining methods are used for this purpose. Over the past five years, a sufficient number of scientific papers have been published, which served as sources of ideas for developing a criterion for the significance of the initial social data. We give a description of quite interesting works.

Abdul Ghaffar Shoro and Tariq Rahim Soomro note the possibilities of Big Data analysis and recognize meaningful information from popular social networks [13], when explore perspectives Apache Spark framework. Authors of paper named "Learning to Match using Local and Distributed Representations of Text for Web Search" has develop [10] new document ranking model outperforms traditional baselines based on neural networks. Research results of the knowledge extraction from large clinical narratives in Bulgarian language are proposed in "Text Mining and Big Data Analytics for Retrospective Analysis of Clinical Texts from Outpatient Care" [5].

A number of articles are devoted to the use fuzzy logic methods in the semantic analysis of unstructured text and natural language. Farman Ali provides [1] fuzzy decision-making solution to monitor transportation activities and to make a city-feature polarity map for travelers in "Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling". "Public Sentiments Analysis Based on Fuzzy Logic for Text" devoted to the technique [17] defined as public sentiments discriminator and considering both fuzzy logic and sentiment complexity. Amir Karami and others solved the problem of increasing the productivity [8] for processing large-scale medical documents by using fuzzy clustering in topic modeling.

There are works on determining the records tonality from the Internet social media. Han Liu and Mihaela Cocea review the concepts and techniques of granular computing in general, and focus on [9] the characteristics of fuzzy information granulation in their investigation. The paper [6] of James M. Heilman, et. al. quantifies the production and consumption of Wikipedia's medical content in order to determine the demographic characteristics of authors and readers, assess the completeness of the contained medical information.

Also, there are researches on measuring the quantitative and bibliometric characteristics of information sources and contained content. Mehdi Jafari and others simulate human-like methods by integrating fuzzy logic with traditional statistical approaches to improve [7] textual summarization accuracy. Dejian Yu has make a retrospective analysis of scientific publications for 50 years with text mining and bibliometric [20]. He provides the complex list of bibliometric characteristics.

Works "Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media" of Kashyap Popat [11] and "Evaluating the credibility of English web sources as a foreign-language searcher" of Alyson L. Young [19] are devoted to the methods of determining the unstructured texts reliability.

We take as a basis the scientific principles and analytical technologies from the above works. The popularity of using fuzzy logic methods in analysis of natural language texts was discovered. A methods review was conducted for assessing the accuracy of textual data from the world wide web public sources. The most interpreted bibliometric parameters for digital sources of textual data are determined. However, we note that the developed significance criterion must be adaptable to use in the system of a social profile building [3]. Also, it will be applied to each individual object, connection or property of the social environment within the proposed [15] social profile and social environment models.

3 The determining problem of the information importance in the social profiling process

The developed mathematical model [15] of the social environment provides information relating to the information cards of individuals, groups and social phenomena. They are divided into actants — the definitions of real-world objects & phenomena, and predicates — the relations between actants. Each individual element of the social environment has its own unique list of attributes, presented in tabular form. Let's consider them in more detail.

A personal SP contains many of the names and pseudonyms of the concerned person. Also, it consists key dates of life and related locations, personal characteristics and the most significant social phenomena, as well as links to profiles of other people and social groups, including a list of the person`s social roles and group relationships types. All this have links to sources in dynamic content. A conclusion is made about the mood of the person, both as a whole and in relation to the surrounding objects of the social environment. It based on the analysis of the source data context and tonality containing the elements presented above. Such content may include fairly subjective evaluation information, which is nonetheless important for drawing up a complete picture of the social profile.

In a group SP, its own unique parameters are added to the information from the profiles of its members. They can change drastically over time. Therefore, the included personal profiles have timestamps in the group relation descriptions to organize the analysis results. The complexity of defining the social groups boundaries also depends on the fact that they may be informal, conditional and implicit. In view of this, we restrict ourselves to the description of the personal social profile model only.

Social phenomena are information about the objects and events of the social environment. They are connected with a person or group in the task of a social profile building. Phenomena can be described directly or indirectly - when existence of one indicated by other phenomena. What is more, they have a number of properties [15]: time and place of occurrence, the sets of involved participants and accompanying real world objects, a list of sources confirming the existence of this phenomenon, relationships between phenomena and other social objects, the related concepts that give a descriptive characteristic of this phenomenon.

All the above information should have the following qualities to achieve the most truthful and complete results:

- Reliability - the analyst must know whether the collected initial information is true or false, and to what extent. This directly affects the veracity of the social profiling results.
- Relevance - social profile data from public sources can be updated at different intervals, ranging from a few changes per hour to just one change per week or even longer. The less time passes from the origin of a certain tangible event to its recording into the sources of the initial SP data and from its gathering to the completion of the analytical processing, the more accurately the results of the analysis will reflect reality.
- Detail - also affects the final results of social profiling. This quality allows for a more in-depth analysis of the social environment. The more properties an object has, the easier it is to compare one with other social environment elements.

The initial data for a social profile construction is taken from a variety of public Internet sources. It is impossible to state unequivocally that the sample will contain information that fully satisfies the indicated qualities. There may be a contradictory or completely false information in the samples. Moreover, even if the SP consist of only true information, it is not always fully used in further research due to the insignificant value of its individual components. Semantic value is determined primarily by the number of links of the selected element with the surrounding social environment objects. Therefore, it is required to introduce additional criteria related to information sources. The reliability and moderability show resources that have a high reputation. The attendance provides information how much the resource is known, readability — how popular the published content is. The appearance speed and the number of new unique publications give an idea how active the authors of the content are.

Accounting for these qualities leads to conduct activities to study the significance of the collected data variety used in the social profiling process. As a result, each element from the social environment description must be assigned a certain weighting factor. Weights show the conformity of the element to these qualities. Assessment of the quality indicators carried out not only by computer means, but also handled by experts. They can determine arbitrary values from Absolute Truth to Absolute Falsehood. Therefore, the apparatus of the fuzzy logic theory is the most suitable for modeling the significance criterion of the initial social data.

Then the mathematical model of personal social profile will take the form [15]:

$$PSP = \left\{ P\left(X, v\right), R\left(Y, u\right), Q\left(m+n\right) \right\}, \tag{1}$$

where $P = \left\{ P_1, \ldots, P_m \right\}$ – the social objects array of the person in question (themes, events, persons, etc.);

$R = \left\{ R_1, \ldots, R_n \right\}$ – the social connections between social objects array of the person in question;

m, n - the number of social objects and connections in the personal SP, respectively, and $m - 1 \le n$;

$Q(m+n) = \{Q_1, \ldots, Q_{m+n}\}$ – the fuzzy set of "intonations" of social characteristics, fundamental for a person`s complex mood assessment and depending on the number of social characteristics; it may take values in the interval: $Q \in \left[-1;\ 1\right]$;

X, Y – property sets of social objects and relationships, respectively;
v, u – weights of social objects and connections, respectively, they are fuzzy values.

4 The model development for the significance criterion of SP objects

Approaches to the processing [16] of text and multimedia data differ. We give their descriptions. The algorithm for analyzing an unstructured social profile source text consists of the following steps:

1. Determining the language of the text;
2. Detection of the information card elements in the initial textual data;
3. The primary lexical analysis is the compilation of a linguistic network, which is the basis for representing a social profile as a graph:
   3.1. The division of text into logical parts;
   3.2. Search for grammar foundations;
   3.3. Identify the text tonality;
   3.4. Determination of the predicates;
   3.5. Highlighting definitions;

4. Secondary lexical analysis - grouping the elements of the resulting linguistic network:

4.1. Search for sentences with verbal constructions (keywords) from the information card;

4.2. Selection from the found sentences of words - candidates for the role of social objects and social connections;

4.3. Recursive processing of the source text to the manual depth, performing steps 4.1 - 4.2;

4.4. Determining the mood of social objects by summarizing of the total mood of their attributes. Search for emotionally colored phrases in the information card descriptions, adding conclusions about the expert assessment of the associated multimedia content tonality. In conclusion, the adjustment of the object`s mood is carried out considering the emotional background of the social environment.

A SP multimedia content is analyzed as follows:

1. Content is divided by nature (audio, images);

2. Content is divided into author's content and information about the selected person or group;

3. For author's content, multimedia objects are compared with existing samples from the Internet and the social profiles; also, the service information is extracted to find out information about the studied person`s activities and preferences;

4. For information about the studied person, a detailed content-analysis is carried out using the available tools of multimedia analytics (such as Machine Learning) in order to isolate essential information about the social environment from the multimedia object itself;

5. For all types of data, tonality analysis is carried out on the basis of the content expert assessment itself and the phrases extracted from it;

6. The results of the analysis are recorded to the information card with links to the processed content.

In this way, textualized information from multimedia content can also be evaluated. The most efficient way is to check the data significance at the stage of secondary lexical analysis. The SP information card is filled at this step. It also shows the semantic properties of the future social profile elements.

We pay attention to the belief functions and plausible reasoning definitions from Dempster-Shafer evidence theory [12] in drawing up the data significance criterion. Schafer's approach allows us to interpret trust and credibility as interval limits of the possible hypothesis truth. This is useful in assessing indicators whose probabilistic values are unknown. For example, when the source of information previously not considered is included in the raw data sample, when the data is fragmentary and incomplete for a comprehensive assessment, or when an expert has questions because of a lack of knowledge in a certain area of interest.

Initial data characteristics were selected empirically. The weight coefficients *u, v* are calculated on its basis. There are verifiable difficulties, subjectivity of expert assessments and the prevalence of the human factor in data interpreting. It is not possible to give an accurate importance assessment of an object or phenomenon. Formalized criteria can only be considered ratings V_{deg}, presented in numerical or percentage form. The fuzzy logic methods are proposed to solve this problem. We defined the fuzzy variables that affect the value of the significance coefficient for SP entities: relevance (V_{vol}), meaning (V_{val}), originality (V_{unq}), validity (V_{arg}), credibility (V_{ver}) of data and source dependability (V_{auth}). Their characteristic functions μ are:

– Relevance - the freshness of initial information. It actually represents the information appearance time into the raw sample. We will consider the interval of updating SP information as one day. This value provides the convenience of carrying out relevance calculations. The choice of a shorter period is not advisable, since the changes are mostly insignificant during this time. A longer period may affect that some gathered information will be irrelevant at the collection time.

The fewer days that have passed since the publication until the moment of data collection or processing the more relevant it is. We suppose the published social information will lose half its relevance a month later. Further reduction of the relevance follows a power law. The 31 days interval was chosen because the people's activities associated to calendar cycles. For example, payouts and vacations are proportional to the months. The choice of the relevance weakness degree = 2 is justified by the fact that a person is usually socially active in the long term, but shortly he can have frequent intervals of 'silence'. We also determine that the information is most relevant during day of publication. Author may make changes to the published message during this period, and readers may not be able to immediately get acquainted with it (e.g. when they are in different time zones). This number will decrease in the future because the pace of life acceleration and the spread of mobile telecommunications grow.

However, sometimes information is not relevant yet at the collection time. This is usually represented by forecasts and conditional assumptions. Past events information always relevant at the time of publication, as well as facts are not tied to time.

V_{vol} = {not actual yet, actual, not actual anymore} – possible values of relevance criteria;

vol = {no, yes} – shows the availability of more recent information;

time = {past, present, future} – shows the time of statements, which contain the social essence, to determine their relevance;

$x_{vol} = [0; \infty)$ – elapsed time since publication;

$$\mu_{V_{vol}} = \begin{cases} 0, vol = 1 \wedge time = \text{"future"}; \\ 1, x_{vol} \le 1 \wedge time = \text{"past"}; \\ \dfrac{1}{1 + (\dfrac{x_{vol} - 1}{30})^2}, x_{vol} > 1, \end{cases} \quad (2)$$

– The meaning of information - shows the power of connectedness with other entities SP. This indicator depends on their weights. We modify the class S membership function so that information is considered secondary if objects with the connections no more than half of the all considered entity links, otherwise - primary.

V_{val} = {meaningless, minor, major, critical} – information may not relate to the actual problem, have little or great importance in solving it, and also be central by providing a task solution or opening previously unknown areas for research;

x_{val} = [0; *nlink*] – the integer number of connections between social entities whose weight is greater than a certain threshold value u_{min};

nlink – total number of unique connections with the entity under investigation;

$$\mu_{V_{val}} = \begin{cases} 0, x_{val} = 0; \\ \dfrac{2x_{val}^{\ 2}}{nlink^2}, 0 < x_{val} \le \dfrac{nlink}{2}; \\ 1 - \dfrac{2(x_{val} - nlink)^2}{nlink^2}, \dfrac{nlink}{2} < x_{val} < nlink; \\ 1, x_{val} = nlink, \end{cases} \quad (3)$$

– The source dependability is determined empirically by an expert. There are no exact criteria. A zero value of the indicator means that the source authority cannot be determined.

V_{auth} = {not defined, low, high} – values showing how dependable the source is;

x_{auth} = [0; 1] – average expert scale of authority;

$$\mu_{V_{auth}} = \begin{cases} 0, x_{auth} = 0; \\ 2x_{auth}^{\ 2}, 0 < x_{auth} \le 0.5; \\ 1 - 2(x_{auth} - 1)^2, 0.5 < x_{auth} \le 1, \end{cases} \quad (4)$$

– Data validity – defined by the existence of arguments, evidence, references of considered social element. The resulting assessment is influenced by the fact that the sample contains the most reliable sources of information and the average level of sources dependability. This is done in order to identify information noise: some social media resources copy facts from the outside hastily. The assessment of this indicator more than 0.5 says that the considered fact has weighty evidence of truth.

V_{arg} = {not enough, enough} – states of validity criteria;

x_{arg} = [0; ∞) – number of possible arguments;

$$\mu_{V_{arg}} = \begin{cases} 0, x_{arg} = 0; \\ \dfrac{1}{1 + e^{2(2 - \max_{x_{arg}} \mu_{V_{auth}} - \frac{\sum^{x_{arg}} \mu_{V_{auth}}}{x_{arg}})}}, x_{arg} > 0, \end{cases} \quad (5)$$

– Data originality - the number of information mentions. This index used to determine the less common facts. This value shows primary sources of information. Like validity, this indicator depends on the source dependability. It allows you to separate the facts from the unsubstantiated rumors and assumptions. This simplifies the process of identifying verified primary sources.

V_{unq} = {no mention, few mention, widely known} – list of uniqueness values;

x_{unq} = [0; *ns*] – the number of sources containing the analyzed statement, with a publication date less than in the original;

ns = [1; ∞) – the total number of sources containing the subject statement from the SP;

$$\mu_{V_{unq}} = \begin{cases} 1, x_{unq} = 0; \\ 1 - \dfrac{2(\sum_{x_{unq}} \mu_{V_{auth}})^2}{ns^2}, 0 < x_{unq} \le \dfrac{ns}{2}; \\ \dfrac{(ns - 1 - \sum_{x_{unq}} \mu_{V_{auth}})^2}{ns^2}, \dfrac{ns}{2} < x_{unq} < ns, \end{cases} \quad (6)$$

– Data credibility - shows the different points of view (PoV) prevalence. It depends on the validity and originality. If there is a single displayed opinion, we will consider it true until alternatives appear. Otherwise, the credibility of this PoV will be depend on confidence indicator of others sights.

$V_{ver} = \{\text{false, undefined, true}\}$ – possible states of credibility criteria;

$x_{ver} = [0; \infty)$ – number of different points of view in initial dataset;

$$\mu_{V_{ver}} = \begin{cases} 1 - \max_{i=1}^{x_{ver}} (ver_i), x_{ver} \ge 1; \\ 1, x_{ver} < 1, \end{cases} \quad (7)$$

where ver – the confidence indicator of proceeded data. It defined as the most authoritative opinion, which must either be supported by many facts or be novel.

$$ver = \max\left[\mu_{V_{arg}} (\mu_{V_{auth}}), \mu_{V_{unq}} (\mu_{V_{auth}}, ns_{min}) \right], \quad (8)$$

Thus, the summary mathematical representation of the weighting coefficient for data significance definition will be as follows:

$$u = \mu_{V_{vol}} * \mu_{V_{val}} (u_{min}) * \mu_{V_{ver}} * ver\ , \quad (9)$$

The coefficient *v* is depicted similarly. The difference is that it focuses directly on the definition of actants.

5 Discussion

The introduction of the developed significance criterion for the initial text data allowed to detect and exclude from consideration extraneous and fake information sources. Testing was conducted on a sample of unstructured social texts from 30 public sources. These results were issued by Google search engine for the query "famous blockchain developers biography". Some selected resources in the sample contained a meaningless compilation of sentences and phrases borrowed from trusted sources. Thereby they mimicking popular information resources. Some data sources were randomly sampled and did not contains the required information.

Also, the usage of a fuzzy significance criterion (relevance indicator) has simplified the sorting of links between social objects in chronological order. It is a part of the ongoing research devoted to methods creation for social profiling task.

In the future, it is planned to introduce the developed criteria into text-analytical module of the social profile building system based on Big Data technologies. Detailed performance testing of the weight criterion will be carried out on a sample of 1500 unique public web-pages with socialized data. If necessary, the characteristic functions of existing quantitative indicators will be corrected to process bibliometric attributes.

6 Conclusion

The result of this work is a fuzzy criterion model. It determines the significance coefficient of social profile elements in a comprehensive analysis. The criterion based on the such qualitative information characteristics as: connectivity, reasoning, relevance, reliability, uniqueness of data and expert assessment of the source verification level. These parameters are primary and necessary to identify outdated and unreliable information in the SP, as well as finding unreliable sources of social media. The resulted fuzzy model can be implemented in the developed concept of the social environment states description. It allows you to increase the consistency degree of the aggregated social profile elements. Also developed significance criterion does simple the tonality analysis procedure of social media data by identifying the semantic content of the particular textual elements. Next stage of our research work devoted to the development of modular implementation for the software and instrumental complex of social profiling.

Acknowledgments

This work is carried out with the support of the Russian Foundation for Basic Research grant №18-07-00408 in a research project named "Fundamental theoretical bases development for self-adaptation of applied software systems".

References

[1] Ali F., et al. Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling. Transportation Research Part C: Emerging Technologies 77: pp. 33-48, 2017.

[2] Barabási A. L., Albert R. Emergence of scaling in random networks // science, 1999. – V. 286, №. 5439. pp. 509-512.

[3] Bershadsky A. M., Bozhday A. S., Koshevoy O. S., Timonin A. Y. Social profiles - Methods of solving actual socio-economic problems using digital technologies and Big Data // Third International Conference «Digital Transformation and Global Society (DTGS)», 2018, St. Petersburg, Russia, May 30 - June 2, 2018, Digital Transformation and Global Society. DTGS 2018. Communications in Computer and Information Science, vol 858, pp. 436-445, doi: 10.1007/978-3-030-02843-5_35

[4] Bershadsky A. M., Bozhday A. S., Timonin A. Y. The Process of Personal Identification and Data Gathering Based on Big Data Technologies for Social Profiles // First International Conference «Digital Transformation and Global Society (DTGS)», 2016, St. Petersburg, Russia, June 22-24, 2016, Volume 674 of the series Communications in Computer and Information Science, Springer International Publishing Switzerland, pp. 576-584, ISBN: 978-3-319-49699-3, doi: 10.1007/978-3-319-49700-6_57

[5] Boytcheva S., Angelova G., Angelov Z., Tcharaktchiev D. Text mining and big data analytics for retrospective analysis of clinical texts from outpatient care. Cybernetics and Information Technologies, 15(4), pp. 58-77. 2015.

[6] Heilman J.M., West A.G. Wikipedia and medicine: quantifying readership, editors, and the significance of natural language. Journal of medical Internet research, 17(3), e62. 2015.

[7] Jafari M., Wang J., Qin Y., Gheisari M., Shahabi A.S., Tao X. Automatic text summarization using fuzzy inference. In 2016 22nd International Conference on Automation and Computing (ICAC), pp. 256-260. IEEE. 2016, September.

[8] Karami A., Gangopadhyay A., Zhou B., Karrazi H. Flatm: A fuzzy logic approach topic model for medical documents. In 2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC) (pp. 1-6). IEEE. 2015, August.

[9] Liu H., Cocea M. Fuzzy information granulation towards interpretable sentiment analysis. Granular Computing 2.4: pp. 289-302, 2017.

[10] Mitra B., Diaz F., Craswell N. Learning to match using local and distributed representations of text for web search. In Proceedings of the 26th International Conference on World Wide Web (pp. 1291-1299). International World Wide Web Conferences Steering Committee. 2017, April.

[11] Popat K., Mukherjee S., Strötgen, J., Weikum G. (2017, April). Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 1003-1012). International World Wide Web Conferences Steering Committee.

[12] Shafer G. Dempster-Shafer theory. Encyclopedia of artificial intelligence, 1, pp. 330-331. 1992.

[13] Shoro A. G., Soomro T. R. Big data analysis: Apache spark perspective. Global Journal of Computer Science and Technology. 2015.

[14] Sowa J. F. Conceptual Structures: Information Processing in Mind and Machine // Reading, MA: Addison-Wesley, 1984. – 481 p.

[15] Timonin A. Y. Set-theoretic and mathematical modeling of social environment states // Izvestiya vysshikh uchebnykh zavedenii. Volga region. Engineering Sciences, no. 1 (49), - Penza: PSU Publishing House, 2019, pp. 18-33. doi: 10.21685/2072-3059-2019-1-2

[16] Timonin A. Y., Bozhday A. S. Investigation of the analyzing process textual and multimedia social profile data from open information sources. Izvestiya vysshikh uchebnykh zavedenii. Volga region. Engineering science, №2 (42), - Penza: PSU Publishing House, 2017, pp. 19-28. doi: 10.21685 / 2072-3059-2017-2-2

[17] Wang X., Zhang H., Xu Z. Public sentiments analysis based on fuzzy logic for text. International Journal of Software Engineering and Knowledge Engineering, 26(09n10), pp. 1341-1360. 2016.

[18] Watts D. J., Strogatz S. H. Collective dynamics of 'small-world' networks // Nature. 1998. V. 393 (6684). pp. 440-442.

[19] Young A. L., Komlodi A., Rózsa G., Chu P. Evaluating the credibility of english web sources as a foreign-language searcher. In Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives through Information & Technology (p. 42). American Society for Information Science. 2016, October.

[20] Yu D., Xu Z., Pedrycz W., Wang W. Information Sciences 1968–2016: A retrospective analysis with text mining and bibliometric. information sciences, 418, pp. 619-634. 2017.