ISSN 1995-0802, Lobachevskii Journal of Mathematics, 2019, Vol. 40, No. 11, pp. 1831–1836. © Pleiades Publishing, Ltd., 2019.

The Algorithm for Decision-Making Supporting on the Selection of Processing Means for Big Arrays of Natural Language Data

K. Polshchykov[1*], S. Lazarev[1], O. Polshchykova[1***], and E. Igityan[1****]**

(Submitted by Vl. V. Voevodin)

[1] *Belgorod State University, Belgorod, 308015 Russia*

Received June 13, 2019; revised June 30, 2019; accepted July 14, 2019

Abstract—In this paper decision support algorithm for choosing the processing means of natural language big data arrays is proposed. In the process the algorithm uses the program for evaluating the effectiveness of text analyzers. This program is based on the operation of a fuzzy choice system, which serves to calculate the integral indicator of the text analyzer effectiveness. The quality and efficiency of getting answers to test questions are taken into account when evaluating the effectiveness of text analyzers.

DOI: 10.1134/S1995080219110222

Keywords and phrases: *natural language big data arrays, text analyzer effectiveness, decision support algorithm, fuzzy choice system, integral indicator, approximations of membership functions.*

1. INTRODUCTION

In large organizations, institutions and departments huge amounts of diverse information in the form of texts in natural language are accumulated for many years. Such information resources can be classified as big data containing information relating to a particular company, its goals, objectives, structure, personnel, activities, implemented projects, financial turnovers, reports, future plans, partners, etc. These large data files are stored electronically in various formats (MS Word, MS Excel, txt, pdf, djvu, HTML, etc.) in numerous corporate systems, archives, portals, databases of various departments, electronic document management systems, electronic mail, file directories, etc. In the process of making certain management decisions, it is very important for the manager to consider the information contained in all sources. However, due to the fact that these data volume is extremely large, heterogeneous, not systematized and distributed to different repositories, the manager does not have the ability to use the full information necessary for this to make decisions. According to IBM research, executives have no more than 7 percents of the information required to select the best solution [1]. As a result, the quality of management suffers, efficiency and competitiveness of the company decrease.

The above problem can be solved if the large and poorly structured natural-language arrays highlight the necessary information, giving the opportunity to the head in a timely manner to get reliable answers to specific questions. For this purpose, special software is used to allow the semantic analysis of texts receiving a request and giving response information in natural language. The basis of such tools is computer technology linguistic processor, aimed at extracting meaning from large arrays of natural language data. To solve this problem, various software solutions, called text analyzers, were developed: for example, Google Desktop, Yandex.Server. For now, AskNet's semantic question-answer search engines [2], the Russian Context Optimizer software package for the analysis of Russian texts [3], Ontos [4], information and analytical system ARION [5], etc. can be used.

*E-mail: polshchikov@bsu.edu.ru
**E-mail: lazarev_s@bsu.edu.ru
***E-mail: polshchikova@bsu.edu.ru
****E-mail: medvedeva_e@bsu.edu.ru

The purpose of the study outlined in the article is to ensure the choice of text analyzers that allow processing large natural-language arrays for making effective control decisions. To achieve this goal is required to solve the following tasks:

1) to offer analytical expressions for evaluating the software text analyzers effectiveness;

2) to develop an algorithm for decision support on the software tools choice that provides the most effective analysis of texts in a given subject area.

2. INDICATORS OF A TEXT ANALYZER EFFECTIVENESS

The user, trying to understand the details of the problem, forms a question in natural language and sends it to the program text analyzer. The analyzer performs semantic processing of existing text arrays and gives the user response in natural language.

It is proposed to evaluate the effectiveness of a software text analyzer based on the calculation of the values of two indicators:

1) Q is quality of the answers given by the analyzer;

2) D is average response time.

The indicator should consider the veracity and completeness of the issued response. The indicator value can be calculated by the formula

$$Q = N_Q/N, \tag{1}$$

where N_Q is the number of sufficiently high-quality answers given by the analyzer to the user's questions; N is the total number of formed question-answer pairs.

The indicator Q can take values from 0 to 1. The larger the value Q, the higher the efficiency of the software text analyzer. The indicator D value characterizes the ability of the analyzer to promptly provide answers to user questions. The value of this indicator can be calculated by the formula

$$D = \frac{1}{N}\sum_{i=1}^{N} T_i, \tag{2}$$

where T_i is the time during which the analyzer issued an answer to the question number.

The value D indicates how much time the analyzer takes on average to generate and issue a response. The larger the value, the higher the efficiency of the software text analyzer.

Values Q and D are particular indicators of a software text analyzer effectiveness. On their basis, we can propose an integral index for making decisions on the choice of means for processing large arrays of natural language data.

The quality levels of the answers given by the analyzer and the promptness of their issuance are difficult to determine by strict numerical criteria. In this case, the fuzzy sets apparatus can be used. Then the quality of the analyzer's answers can be assessed using fuzzy sets "High quality answers" and "Low quality answers", and the promptness of giving answers to user questions may correspond to fuzzy sets "High efficiency answers" and "Low responsiveness answers".

It is possible to find the resulting evaluation of a software text analyzer effectiveness by calculating a certain integral index with the help of a fuzzy inference [6–11].

The value of the integral index is proposed to be calculated using the model corresponding to the zero-order Sugeno algorithm of fuzzy inference [12]. In this case, is used the following base of fuzzy rules:

$$If(Q = V_Q) \quad \text{and} \quad (D = V_D), \quad \text{then} \quad (E = Y_1),$$

$$If(Q = V_Q) \quad \text{and} \quad (D = W_D), \quad \text{then} \quad (E = Y_2),$$

$$If(Q = W_Q) \quad \text{and} \quad (D = V_D), \quad \text{then} \quad (E = Y_3),$$

$$If(Q = W_Q) \quad \text{and} \quad (D = W_D), \quad \text{then} \quad (E = Y_4),$$

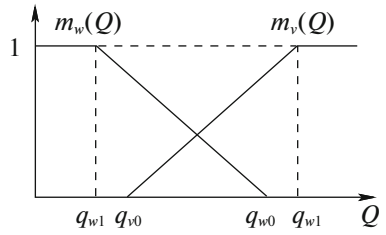

Fig. 1. Membership function $m_v(Q)$ and $m_w(Q)$.

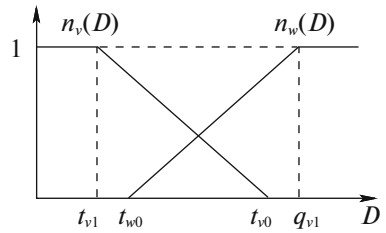

Fig. 2. Membership function $n_v(D)$ and $n_w(D)$.

where V_Q is fuzzy set "High quality answers"; W_Q is fuzzy set "Low quality answers"; V_D is fuzzy set "High responsiveness answers"; W_D is fuzzy set "Low responsiveness answers"; $Y_1 = 4$, $Y_2 = 3$, $Y_3 = 2$ and $Y_4 = 1$ are values of individual conclusions of fuzzy rules.

The belonging degree of quantities Q and D to fuzzy sets is determined by the values of the corresponding membership function $m_v(Q)$, $m_w(Q)$, $n_v(D)$ and $n_w(D)$ with the parameters q_{v0}, q_{v1}, q_{w0}, q_{w1}, t_{v0}, t_{v1}, t_{w0} and t_{w1} (Fig. 1 and Fig. 2).

The Sugeno zero-order fuzzy inference algorithm includes three main stages. At the first stage, fuzzification is performed:

$$m_v(Q) = \begin{cases} 0, & Q < q_{v0}; \\ \frac{Q-q_{v0}}{q_{v1}-q_{v0}}, & q_{v0} \le Q < q_{v1}; \\ 1, & Q \ge q_{v1}; \end{cases} \qquad m_w(Q) = \begin{cases} 1, & Q < q_{w1}; \\ \frac{q_{w0}-Q}{q_{w0}-q_{w1}}, & q_{w1} \le Q < q_{w0}; \\ 0, & Q \ge q_{w0}; \end{cases}$$

$$n_v(D) = \begin{cases} 1, & D < t_{v1}; \\ \frac{t_{v0}-D}{t_{v0}-t_{v1}}, & t_{v1} \le D < t_{v0}; \\ 0, & D \ge t_{v0}; \end{cases} \qquad n_w(D) = \begin{cases} 0, & D < t_{w0}; \\ \frac{D-t_{w0}}{t_{w1}-t_{w0}}, & t_{w0} \le D < t_{w1}; \\ 1, & D \ge t_{w1}. \end{cases}$$

At the second stage, aggregation is performed:

$$G_1 = m_v(Q) \wedge n_v(D); \quad G_2 = m_v(Q) \wedge n_w(D); \quad G_3 = m_w(Q) \wedge n_v(D); \quad G_4 = m_w(Q) \wedge n_w(D).$$

At the third stage, defuzzification is performed: $E = \sum\limits_{r=1}^{4} G_r Y_r / \sum\limits_{r=1}^{4} G_r$.

The calculation of the indicator E is the result of fuzzy inference system operation designed to evaluate the effectiveness of a software text analyzer. The higher the value of the integral index, the more efficient the software text analyzer is.

3. SETTING THE PARAMETERS OF THE SYSTEM FOR FUZZY INFERENCE

The proposed fuzzy selection system needs to be configured, i.e. determine the correct parameters' values of the membership functions.

It is proposed to conduct a series of K experiments with the participation of M_0 experts. In each experiment, the number k (where $k = 1, 2, ..., K$) of a specially created program for setting up a fuzzy inference system asks questions and gives answers to them from a subject area in which experts have

in-depth knowledge and experience. In this case, using the specified program, the quality of the response and the time of its issuance are recorded. As a result of each experiment k is determined quantities N, N_Q and T_i. Then the indicator values Q_k and D_k are calculated using formulas (1) and (2). At the end of the experiment, in the interface of the program for setting up a fuzzy inference system, each expert notes which (high or low), in his opinion, were the quality and efficiency of responding with a test system.

Next, are calculated estimates of the membership functions for each value Q_k and D_k:

$$m^*_{vk} = \frac{M^+_Q(Q_k)}{M_0}; \quad m^*_{wk} = \frac{M^-_Q(Q_k)}{M_0}; \quad n^*_{vk} = \frac{M^+_D(D_k)}{M_0}; \quad n^*_{wk} = \frac{M^-_D(D_k)}{M_0},$$

where $M^+_Q(Q_k)$ is the number of experts who noted that with Q_k the quality of the answers given by the test system was high; $M^-_Q(Q_k)$ is the number of experts who noted that with Q_k the quality of the answers given by the test system was low; $M^+_D(D_k)$ is the number of experts who noted that with D_k the system answer response time was high; $M^-_D(D_k)$ is the number of experts who noted that with D_k the system answer response time was low.

Then from the set $m^*_v = \{m^*_{v1}, m^*_{v2}, ..., m^*_{vk}, ..., m^*_{vK}\}$ by selecting those estimates whose values satisfy the condition $0.1 \le m^*_{vk} \le 0.9$, a set $\widetilde{m}_v = \{\widetilde{m}_{v1}, \widetilde{m}_{v2}, ..., \widetilde{m}_{vs}, ..., \widetilde{m}_{vS_1}\}$ is formed, where s is the number of the current element of the set formed; S_1 is the number of elements in the formed set.

Similarly, from sets $m^*_v = \{m^*_{w1}, m^*_{w2}, ..., m^*_{wk}, ..., m^*_{wK}\}$, $n^*_v = \{n^*_{v1}, n^*_{v2}, ..., n^*_{vk}, ..., n^*_{vK}\}$ and $n^*_w = \{n^*_{w1}, n^*_{w2}, ..., n^*_{wk}, ..., n^*_{wK}\}$ sets $\widetilde{m}_w = \{\widetilde{m}_{w1}, \widetilde{m}_{w2}, ..., \widetilde{m}_{ws}, ..., \widetilde{m}_{wS_2}\}$, $\widetilde{n}_v = \{\widetilde{n}_{v1}, \widetilde{n}_{v2}, ..., \widetilde{n}_{vs}, ..., \widetilde{n}_{vS_3}\}$ and $\widetilde{n}_w = \{\widetilde{n}_{w1}, \widetilde{n}_{w2}, ..., \widetilde{n}_{ws}, ..., \widetilde{n}_{wS_4}\}$ are formed accordingly, whose elements have values in the range from 0.1 to 0.9.

Next, need to obtain the equations of the lines $y_1(Q) = a_1Q + b_1$, $y_2(Q) = a_2Q + b_2$, $y_3(D) = a_3D + b_3$ and $y_4(D) = a_4D + b_4$ smoothing the values of the sets' elements $\widetilde{m}_v$, $\widetilde{m}_w$, $\widetilde{n}_v$ and $\widetilde{n}_w$ respectively.

Approximations of membership functions estimates are proposed to be performed using the least squares method, according to which the coefficients of straight lines are calculated using the formulas:

$$a_1 = \frac{S_1 \sum\limits_{s=1}^{S_1} Q_s\widetilde{m}_{vs} - \sum\limits_{s=1}^{S_1} Q_s \sum\limits_{s=1}^{S_1} \widetilde{m}_{vs}}{S_1 \sum\limits_{s=1}^{S_1} Q_s^2 - (\sum\limits_{s=1}^{S_1} Q_s)^2}; \quad b_1 = \frac{\sum\limits_{s=1}^{S_1} \widetilde{m}_{vs} - a_1 \sum\limits_{s=1}^{S_1} Q_s}{S_1};$$

$$a_2 = \frac{S_2 \sum\limits_{s=1}^{S_2} Q_s\widetilde{m}_{ws} - \sum\limits_{s=1}^{S_2} Q_s \sum\limits_{s=1}^{S_2} \widetilde{m}_{ws}}{S_2 \sum\limits_{s=1}^{S_2} Q_s^2 - (\sum\limits_{s=1}^{S_2} Q_s)^2}; \quad b_2 = \frac{\sum\limits_{s=1}^{S_2} \widetilde{m}_{ws} - a_2 \sum\limits_{s=1}^{S_2} Q_s}{S_2};$$

$$a_3 = \frac{S_3 \sum\limits_{s=1}^{S_3} D_s\widetilde{n}_{vs} - \sum\limits_{s=1}^{S_3} D_s \sum\limits_{s=1}^{S_3} \widetilde{n}_{vs}}{S_3 \sum\limits_{s=1}^{S_3} D_s^2 - (\sum\limits_{s=1}^{S_3} D_s)^2}; \quad b_3 = \frac{\sum\limits_{s=1}^{S_3} \widetilde{n}_{vs} - a_3 \sum\limits_{s=1}^{S_3} D_s}{S_3};$$

$$a_4 = \frac{S_4 \sum\limits_{s=1}^{S_4} D_s\widetilde{n}_{ws} - \sum\limits_{s=1}^{S_4} D_s \sum\limits_{s=1}^{S_4} \widetilde{n}_{ws}}{S_4 \sum\limits_{s=1}^{S_4} D_s^2 - (\sum\limits_{s=1}^{S_4} D_s)^2}; \quad b_4 = \frac{\sum\limits_{s=1}^{S_4} \widetilde{n}_{ws} - a_4 \sum\limits_{s=1}^{S_4} D_s}{S_4}.$$

Having obtained the smoothing line coefficients a_1 and b_1, we can determine the desired parameters q_{v0} and q_{v1} from the equations $a_1q_{v0} + b_1 = 0$ and $a_1q_{v1} + b_1 = 1$.

Having solved the equations, we get $q_{v0} = -b_1/a_1$; $q_{v1} = (1 - b_1)/a_1$. Similarly displayed the values of the parameters q_{w0}, q_{w1}, t_{v0}, t_{v1}, t_{w0} and t_{w1}. The resulting formulas for their calculation:

$$q_{w0} = -\frac{b_2}{a_2}; \quad q_{w1} = \frac{1 - b_2}{a_2}; \quad t_{v0} = -\frac{b_3}{a_3}; \quad t_{v1} = \frac{1 - b_3}{a_3}; \quad t_{w0} = -\frac{b_4}{a_4}; \quad t_{w1} = \frac{1 - b_4}{a_4}.$$

4. PROGRAM FOR EVALUATING THE EFFECTIVENESS OF TEXT ANALYZERS

After setting the parameters of the synthesized system for fuzzy inference, it can be used to assess the effectiveness of software text analyzers and then decide on the means choice for processing large arrays of natural language data. It is necessary to create a program for evaluating the effectiveness of text analyzers, including the above-mentioned system of fuzzy inference, a database of test questions and correct answers, as well as means for calculating quantities Q and D using formulas (1) and (2). With this program and participation of a human tester, it is possible to choose the most effective text analyzer from a set of alternative options with the help of the proposed algorithm:

1. Launch of a program for evaluating the effectiveness of a text analyzer.

2. The investigated software text analyzer is launched and connected to large arrays of natural language data containing information about the required subject area.

3. The program for evaluating the effectiveness of the text analyzer asks a test question.

4. The tester receives the question posed by the text analyzer performance evaluation program.

5. The tester enters the asked question in the interface of the analyzed text analyzer and starts the timer to wait for an answer.

6. The software text analyzer forms and gives the answer to the asked question.

7. The tester receives the answer given by the text analyzer and stops the response timer.

8. The tester in the interface of the program for assessing the effectiveness of the text analyzer chooses the option that best corresponds to the response received.

9. If all test questions were asked, the program for evaluating the effectiveness of the text analyzer calculates the value of the integral index E, otherwise returns to step 3 of the algorithm.

10. If all text analyzers are examined, then the analyzer with the highest value of the integral efficiency indicator is recommended for further use, otherwise returns to step 2 of the algorithm.

5. CONCLUSION

Thus, the article presents an overview of the program text analyzers and proposes an algorithm to support decision making on the choice of solutions for processing large arrays of natural language data. In the process the algorithm uses the program for evaluating the effectiveness of text analyzers. This program is based on the operation of a fuzzy choice system, which serves to calculate the integral indicator of the text analyzer effectiveness. To set up a fuzzy choice system, special software and the participation of experts in a given subject area are required. In the process of setting up this system, the membership functions are approximated and the values of their parameters are determined. When evaluating the effectiveness of text analyzers, the quality and efficiency of getting answers to test questions are taken into account.

Text analyzer with the highest integral efficiency indicator value is recommended for further use in the relevant subject area.

The application of the algorithm presented in the article allows taking the best version of a management decision based on information obtained as a result of processing big data arrays of natural language.

Further research on the article topic will be devoted to developing software based on the proposed algorithm and obtaining experimental results of its application.

FUNDING

This article contains the results of the project “Development of tools for implementation of natural-language systems for processing big data” carried out within the framework of the implementation of the Program of the Center for Competence of the National Technology Initiative “Big Data Storage and Analysis Center”, supported by the Ministry of Science and Higher Education of the Russian Federation under the Lomonosov MSU with the Fund for Support of Projects of the National Technology Initiative no. 13/1251/2018 dated December 11, 2018.

REFERENCES

1. SyTech. http://www.sytech.ru/about.php?id=60. Accessed 2019.
2. AskNet. http://asknet.ru/. Accessed 2019.
3. RCO. http://www.rco.ru/. Accessed 2019.
4. Business Robotics. http://ontos.com/. Accessed 2019.
5. Information and Analytical System ARION. http://sytech.ru/about.php?id=5. Accessed 2019.
6. N. Rvachova, G. Sokol, K. Polschykov, and J. Davies, "Selecting the intersegment interval for TCP in Telecomms networks using fuzzy inference system," in *Proceedings of the 6th International Conference on 2015 Internet Technologies and Applications (ITA), Glyndwr Univ., Wrexham, 2015,* pp. 256–260.
7. K. Polshchykov, Y. Zdorenko, and M. Masesov, "Neuro-fuzzy system for prediction of telecommunication channel load," in *Problems of Infocommunications Science and Technology, Proceedings of the 2nd International Scientific-Practical Conference, Kharkiv, 2015,* pp. 33–34.
8. O. Ivaschuk, K. Polschykov, and S. Lazarev, et al., "Integral estimate of terrestrial compartment condition in management of Biotechnosphere of Rural and Urban Areas," Int. J. Pharm. Technol. **8**, 27032–27038 (2016).
9. I. Konstantinov, K. Polshchykov, and S. Lazarev, "The algorithm for neuro-fuzzy controlling the intensity of retransmission in a mobile ad-hoc network," Int. J. Appl. Math. Stat. **56** (2), 85–90 (2017).
10. I. Konstantinov, K. Polshchykov, and S. Lazarev, "Model of neuro-fuzzy prediction of confirmation timeout in a mobile ad hoc network," CEUR Workshop Proc., Math. Inform. Technol. **1839**, 174–186 (2017).
11. K. Polshchykov, S. Lazarev, and A. Zdorovtsov, "Neuro-fuzzy control of data sending in a mobile ad hoc network," J. Fundam. Appl. Sci. **9**, 1494–1501 (2017).
12. T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," IEEE Trans. Syst., Man, Cybern. **15**, 116–132 (1985).