UDC 004.67
DOI: 10.15587/2706-5448.2020.208400

Yaremenko V.,
Syrotiuk O.

DEVELOPMENT OF A MULTI-AGENT SYSTEM FOR SOLVING DOMAIN DICTIONARY CONSTRUCTION PROBLEM

The object of research is the use of multi-agent systems for text data mining. The need for this study arose with a tendency to increase the amount of textual information generated in the world. Accordingly, it is necessary to develop and research methods of its processing, as well as ways to use the results of this processing, because the methods can't exist in isolation from practice. At the same time, there is a development of multi-agent systems (MAS), where agents are endowed with some kind of intelligence, these systems can be easily scaled. The use of MAS for text analysis is a promising area.

The following methods of text data analysis were used in this study: TF-IDF and RAKE methods, Word2Vec neural network models, and TextRank. The algorithms were compared for their work and the results were compared. The corpus of documents (10–12 texts, 5732–12331 words) from the subject areas of physics and biology were used as a test set. According to the results of the study, one method was chosen, on the basis of which the MAS was built to solve the problem. Additionally, Schulze methods (with one and several winners) were used for voting. With the received system additional researches concerning accuracy and speed of work, and also – influence are carried out system parameters for its operation.

It has been found that TF-IDF-based analysis is useful for finding terms in documents with a weak context. The resulting system shows an accuracy of 75 % (3 of the 4 words proposed by the system are terms). The maximum operating time on test cases is 2–3 seconds, which is achieved through the use of parallel calculations and modification of the Schulze method. The results obtained in this paper are heuristic (ontology is a rather vague concept) and require additional elaboration by experts in the relevant fields. However, the results are positive within this experiment.

Keywords: *TF-IDF, RAKE, TextRank, Word2Vec, Schulze method, text data, frequency analysis, parallel computing, multi-agent system.*

Received date: 16.04.2020
Accepted date: 22.05.2020
Published date: 31.08.2020

1. Introduction

Information retrieval and analysis of textual data is an important area of science that is becoming increasingly important in today's realities and is an integral part of life. Each generates a large amount of information from which it is necessary to extract the main content for further use.

Increasing the amount of information leads to deterioration in the quality of its perception and this problem affects all areas of the human information space, including scientific activity. Therefore, the actual problems are the systematization of information and simplification of its perception in the volumes it is available today. In studies related to text data analysis it is noted that the quality of systems for working with textual data primarily depends on data and the chosen method [1, 2]. In works related to multi-agent systems the main goal is to study the impact of the chosen architecture on the quality of the system, as well as on the amount of data the system can process [3, 4]. Based on the analyzed works, the following *object of research* was identified: the use of multi-agent systems (MAS) for the text data analysis. This work is a generalization of attempts to solve the problem of finding the importance of the word for the subject area. Word ranking, morphological search and syntactic similarity methods were chosen to compare. *The aim of research* is to develop a multi-agent system for solving the problem of constructing a dictionary of the subject area from a minimal corpus of documents and a minimal initial dictionary on the basis of comparison of methods for assessing the similarity of words.

2. Methods of research

Theoretical research includes the study of approaches used in the analysis of textual data, namely – in the search for keywords, finding the semantic similarity of words, as well as in the tasks of creating a brief description of the document. Based on this study, the selection of the target group of methods that will be used in the future. Comparison of existing methods, namely: TF-IDF, RAKE, TextRank, and models of the neural network Word2Vec. The methods were divided into two main categories according to the number of documents they process: document (RAKE, TextRank) and the whole body (TF-IDF, Word2Vec) [5, 6].

ISSN 2664-9969

The comparison of TF-IDF and Word2Vec methods was performed on two sets of text data corpora and basic dictionaries of subject areas: physics and biology. The parameters of these sets are given in Tables 1, 2. TextRank and RAKE were compared on randomly selected documents from these sets.

Table 1

Parameters of the documents set (the physics domain)

Characteristics	Value
The number of documents	11 documents
The number of symbols	27701 symbols
The number of words	5732 words
The size in bytes	27100 bytes
The dictionary size in N-Grams	135 N-Grams

Table 2

Parameters of the documents set (the biology domain)

Characteristics	Value
The number of documents	6 documents
The number of symbols	43105 symbols
The number of words	12527 words
The size in bytes	62311 bytes
The dictionary size in N-Grams	44 N-Grams

The practical part includes the design and development of the system based on the method selected according to the results of the previous stage. This stage includes both theoretical system development and analysis of existing solutions and architectures for working with data and practical implementation. Kappa and Lambda architectures, as well as MAS-based systems, were considered [7–9].

The last stage of the study is the analysis of the results and an attempt to explain them in terms of algorithms used and architectural approaches to creating a system. A plan for further research that includes opportunities to improve the system was created.

3. Research results and discussion

At the methods comparison stage, the following results were obtained: keywords extracted using TF-IDF and words similar to the ones in the subject area dictionary derived from Word2Vec cover a fairly broad ontology, as shown in Tables 3, 4. However, N-grams that were obtained with the TF-IDF method are related to the field of physics with a higher value of rank (the most important is the first word because of its high rank). Similar results were obtained for the biology domain (although this judgment is purely heuristic, as the rank of similarity is a value that is relative to the words within one algorithm).

Table 3

Results of TF-IDF and Word2Vec methods (the physics domain)

TF-IDF		Word2Vec	
N-Gram	Similarity rank	N-Gram	Similarity rank
Gravity	0.734	George	0.272
Interaction	0.244	Year	0.245
Quantum	0.229	Unit	0.225
Cause	0.171	Cold	0.22
Start	0.171	Meaning	0.201
Time	0.154	Matter	0.196
Universe	0.147	Ha	0.196
Force	0.129	Glass	0.195
Theory	0.129	Light	0.189
Mass	0.114	Solution	0.188

Table 4

Results of TF-IDF and Word2Vec methods (the biology domain)

TF-IDF		Word2Vec	
N-Gram	Similarity rank	N-Gram	Similarity rank
Hippocampus	0.377	Fruit	0.271
Memory	0.293	Site	0.264
Brain	0.153	Front	0.242
Damage	0.151	Inhibit	0.237
Part	0.129	Body	0.229
Cell	0.121	Time	0.220
System	0.102	Day	0.200
Lobe	0.102	Reptile	0.197
Navigation	0.091	Function	0.190
Immune	0.091	Arthropod	0.188

Comparing the TextRank and RAKE methods was a bit more difficult, as these methods are usually used with a single document. The following results of the processing were obtained on a random document from the subject area of physics (Table 5).

Table 5

Results of RAKE and TextRank methods work (the physics domain)

RAKE		TextRank Algorithm	
N-Gram	Similarity rank	N-Gram	Similarity rank
Semiconductor diodes begin conducting electricity	20.409	Electronic Device	0.119
Impatt diodes exhibit negative resistance	19.909	Power Device	0.110
Perform many different functions	16.000	Such Device	0.096
Reverse direction suddenly drops	14.000	Electronic System	0.090
Impatt diodes	9.909	Crystalline Solids	0.088
Doping impurities introduced	9.000	Wide Application	0.081
Zener diodes	8.909	Communications	0.060
Varactor diodes	8.909	Integration	0.058
Avalanche diodes	8.833	Computing	0.055
Reverse voltage across	8.500	Transistors	0.055

It should be noted that all the algorithms described above, worked with already processed and normalized texts. Among the phrases that RAKE and TextRank choose there are phrases that accurately represent physics terms, but the accuracy (according to heuristic data estimates) is lower. These algorithms show certain new and important phrases in their lists, but in the results of their work quite a lot of terms that are not a part of the dictionary.

From the results of the comparison of TextRank and RAKE, it was concluded that the algorithms get a large amount of key phrases that do not fit the requirements of the terms of the dictionary, which is why the key algorithm for further development of the system was chosen TF-IDF.

It is necessary to note in more detail about the procedure of normalization which documents passed. Since the goal is to create dictionaries of the subject area, an approach consisting of such stages was chosen:

1. Delete punctuation marks.
2. Bring the words of the text to lowercase.
3. Breaking words into *n*-grams.
4. Lemmatization.
5. Delete words that are not nouns.
6. Indexing.

Particularly important in this procedure was item number 5. The addition of this item to the procedure improved the performance of TF-IDF and RAKE methods in the early stages of the study. The main assumption on which this point is based – dictionaries should contain only nouns with adjectives.

After comparison and selection of the method, the MAS architecture was developed on its basis and the system was implemented. Methods using MAS are quite effective [3]. The system is quite simple at this stage of the study and has the following form, shown in Fig. 1.

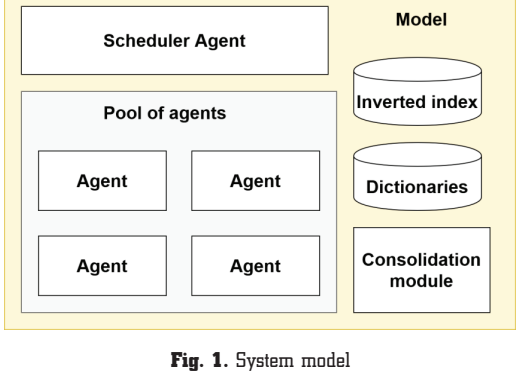

Fig. 1. System model

The main component in this system is the agent (Fig. 2), which performs the function of processing the data corpus depending on the set of keywords that were provided to it by the scheduler agent.

Each agent enters a certain number of words from the base dictionary, with which it selects the documents that will form its own body for processing via TF-IDF. The agent processes the corpus and publishes its own set of candidate words to be added to the main dictionary. The last stage in the system is the consolidation of the results of the work of agents (lists of candidate words) by the method of Schulze voting and further addition of the words chosen by voting to the main dictionary [10].

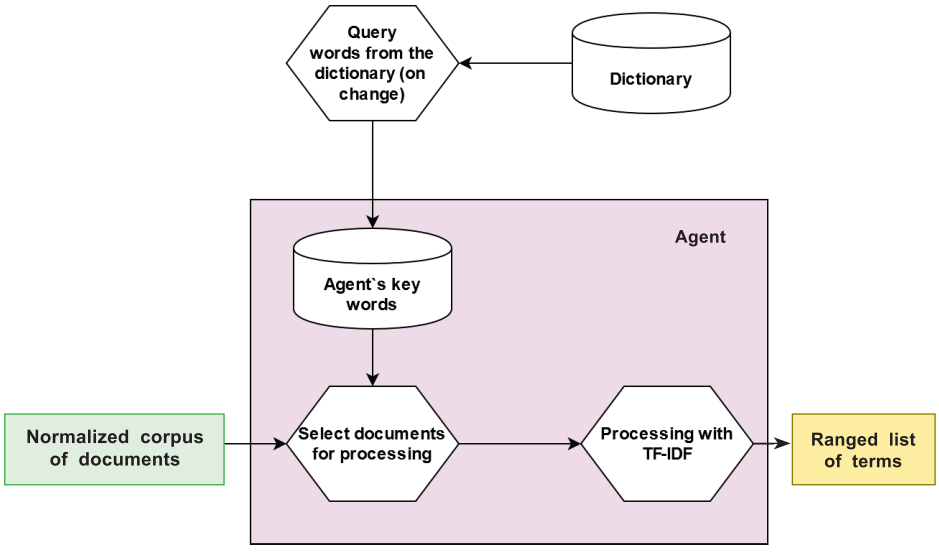

Fig. 2. Internal agent structure

The data corpora from Tables 1, 2 were used in the experiment. Each agent receives the same number of words from the dictionary (if possible) and the number of agents in the system is equal to the number of words in the dictionary of the subject area taken as a sum modulo the number of words agent to vote. Each agent offers 10 words to add to the dictionary (vote). Schultze wins 4 words in the voting process. The results of the system with this configuration are as follows:

1. For the body of physics selected: «gravity», «interaction», «quantum», «start».
2. For the body of biology selected: «brain», «vertebrate», «structure», «size».

From this example, it is possible to estimate that the accuracy of the system is about 75 %, as 3 of the 4 words can be attributed to the relevant subject areas.

4. Conclusions

Based on the study, it is possible to conclude that methods based on frequency analysis using TF-IDF show a high result of the accuracy of finding words related to a common subject area in data structures that have low semantic viscosity (dictionaries).

It can also conclude that the multi-agent system developed in the process of research has proved its correctness and has the right to exist, but requires certain modifications and research:

1. Investigation of the dependence of system efficiency on the size of the dictionary and the number of agents in the system. Investigation of the Amdal coefficient for the developed architecture [11].
2. Research of system operation on large sets of text data and MAS load testing.

ISSN 2664-9969

3. Study of the influence of the size of the case given to one agent on the accuracy of its work, and the accuracy of the system as a whole.

4. Investigation of the dependence of the size of the input data and the number of system agents on the efficiency of its work.

5. Development of a method for deleting words from the dictionary.

It will be important to conduct further research in the field of multi-agent systems, because at this stage, agents are elements of parallelization rather than full-fledged intellectual entities.

References

1. Mikolov, T., Le, Q. V., Sutskever, I. (2013). Exploiting Similarities among Languages for Machine Translation. *ArXiv*. Available at: https://arxiv.org/abs/1309.4168
2. Wu, H. C., Luk, R. W. P., Wong, K. F., Kwok, K. L. (2008). Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems, 26 (3),* 1–37. doi: http://doi.org/10.1145/1361684.1361686
3. Aref, M. M. (2003). A multi-agent system for natural language understanding. *IEMC'03 Proceedings. Managing Technologically Driven Organizations: The Human Side of Innovation and Change (IEEE Cat. No.03CH37502)*, 36–40. doi: http://doi.org/10.1109/kimas.2003.1245018
4. Fum, D., Guida, G., Tasso, C. (1988). A distributed multi-agent architecture for natural language processing. *Proceedings of the 12th conference on Computational linguistics,* 812–814. doi: http://doi.org/10.3115/991719.991801
5. Mihalcea, R., Tarau, P. (2004). TextRank: Bringing Order into Text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing,* 404–411.
6. Rose, S. R., Engel, D., Cramer, N., Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining*. doi: http://doi.org/10.1002/9780470689646.ch1
7. Twardowski, B., Ryzko, D. (2014). Multi-agent Architecture for Real-Time Big Data Processing. *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 3,* 333–337. doi: http://doi.org/10.1109/wi-iat.2014.185
8. Kiran, M., Murphy, P., Monga, I., Dugan, J., Baveja, S. S. (2015). Lambda architecture for cost-effective batch and speed big data processing. *2015 IEEE International Conference on Big Data (Big Data),* 2785–2792. doi: http://doi.org/10.1109/bigdata.2015.7364082
9. Singh, K., Behera, R., Mantri, J. (2019). Big Data Ecosystem: Review on Architectural Evolution. *Advances in Intelligent Systems and Computing,* 335–345. doi: http://doi.org/10.1007/978-981-13-1498-8_30
10. Schulze, M. (2018). The Schulze Method of Voting. *ArXiv*. Available at: https://arxiv.org/abs/1804.02973
11. Amdahl, G. (2007). Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities, Reprinted from the AFIPS Conference Proceedings, Vol. 30 (Atlantic City, N. J., Apr. 18–20). *IEEE Solid-State Circuits Newsletter, 12,* 19–20. doi: http://doi.org/10.1109/n-ssc.2007.4785615

Yaremenko Vadym, *Postgraduent Student, Assistant, Department of System Design, National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», Ukraine, e-mail: yaremenko.v.s@gmail.com, ORCID: http://orcid.org/0000-0001-8557-6938*

Syrotiuk Oleksandr, *Department of System Design, National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», Ukraine, e-mail: alexandr.syrotiuk.dev@gmail.com, ORCID: http://orcid.org/0000-0002-4531-6290*