Method for Determining Information Proximity Based on Spectral Conversion of Text Documents

Maria A. Butakova
Dean, Rostov State
Transport University,
Rostov-on-Don, Russia
butakova@rgups.ru

Andrey V. Chernov
Department of Computer
Engineering and Automated
Control Systems, Rostov
State Transport University,
Rostov-on-Don, Russia
avcher@rgups.ru

Grigorii S. Miziukov
Center of monitoring
quality education, Rostov
State Transport University,
Rostov-on-Don, Russia
mgs_cmko@rgups.ru

Abstract

The process of identifying key information in unstructured sets of textual information is complex and multiple-aspect. In this regard various methods and technologies are being actively developed that can improve the analysis process and reduce the gap between the quality of the obtained results and the computational resources required for the analysis. This article provides an example of an alternative method for determining information proximity in large arrays of textual information. A distinctive feature of this method is the application of spectral conversion of the information and means of descriptive logic for the logical inference of analysis results of the text documents array. The main components of the method as well as conditions and statements of the logical inference of the analysis results are considered. The analysis of the obtained results based on the results of the approbation of the method is given. The obtained results clearly demonstrate the possibility of applying the method for semantic classification problems in information decision-making systems.

1 Introduction

In the context of the current trend of digitalization of various aspects of knowledge domain large amounts of information of different structure are accumulated. The predominant type in these arrays is unstructured information presented in the form of multiple multimedia and text files of different formats and linguistic affiliation. Intelligent analysis technologies are used to analyze these types of information [Lij10, Jia14]. Intelligent analysis is a complex of interdisciplinary links by means of which a basic model is built up that in future serves as a basis for application of various methods. The most commonly used methods are classification, prediction, clustering, association and time series modelling. However, semantic analysis is considered to be an important task within the framework of intelligent analysis of text data [Sar19]. Although there are multiple solutions and approaches in the field of semantic analysis of textual information, not all of them are able to fully provide a qualitative analysis process since there are a number of problems primarily related to the identification of semantic links between the analyzed objects. It is also worth noting the distinctive feature of unstructured information from structured or semistructured one which implies that this type of information does not have a structure that describes the stored data, and it has anthropogenic character. Such an abundance of heterogeneous information results in the need to apply combinations of several different methods to achieve the desired result. Therefore, this article proposes a method for determining information proximity in large arrays of text information, the distinctive feature of which is the use of spectral conversion of information and means of descriptive logic for the logical inference of the result of analysis of text documents array to classify emerging situations and identify redundancy in large arrays of text information. The article is arranged as follows. Section 2 includes information on currently available scientific studies in the selected field. Section 3 describes the proposed method. The main variables of the method, functions, as well as conditions and statements which the logical inference of the results is based on are also considered in the section. Section 4 describes the results of testing the method. Section 5 describes further scientific application of the method. Section 6 concludes the article.

2 Previous Work

The issue of determining information proximity in large arrays of unstructured textual information, for instance, the approach to semantic classification

In: A. Khomonenko, B. Sokolov, K. Ivanova (eds.): Selected Papers of the Models and Methods of Information Systems Research Workshop, St. Petersburg, Russia, 4-5 Dec. 2019, published at http://ceur-ws.org

[Bou14, Ma16] is of particular interest for scientific research. Currently, there are a large number of different methods and technologies used to analyze text information. Among the methods one can distinguish the method of extracting knowledge from information, methods of searching for information in coherent texts, clustering, classification and summarization [Tan19]. “Big Data” technology is the most promising and actively developing technology [Fad18, Ous17]. However, despite the constant development of these approaches there are difficulties that to one degree or another impede the qualitative analysis of textual information. In the article [Jus18] the author considers some of the most frequently encountered difficulties while analyzing .textual information. Particular attention should be paid to the methods of classifying and determining the information proximity between text documents. The most interesting approaches for solving problems in this area are presented in articles [Zha14, Fis08]. In the article [Zha14] the authors propose an approach based on the calculation of the semantic similarity of short texts through language-based network and word semantics. In [Fis08] the authors propose a group of auxiliary methods for determining the informational proximity and texts classification that supplement the formed semantic model with structural information for classification. The following sections propose an alternative approach for determining information proximity and semantic classification of texts based on spectral conversion of information and logical inference by means of descriptive logic.

3 Proposed Method

The process of determining the information proximity between the analyzed objects in large arrays of text information involves the identification of similar intersection points on the basis of which one can make an assumption about the information proximity of two objects. However, this process also determines the unique properties of the objects of analysis that act as preventive condition and do not allow to refer the objects of analysis to one category by the primary features thereby making the process of determining the information proximity more qualitative and effective. Thus, while designing the method the following tasks were set:

1. To identify common and unique properties of objects of analysis. In this case, this is the definition of common and unique lexical units in the text information flow in the process of analysis.
2. To form the structure of representation of the identified lexical units.
3. To determine the information proximity between the objects of analysis under the condition that the structures of the identified lexical units may be the same but the objects of analysis belong to different categories; the structures may be different but the objects belong to the same category.

To solve the above-mentioned problems we propose a method for determining the information proximity in text arrays of information based on the methods of spectral representation of information [Vas17, God77] and methods of logical inference by means of descriptive logic [Kri18]. The application of the spectral approach to the representation of textual information is determined by the high efficiency of the calculation process due to the operation with numerical values in the analysis process. Means of descriptive logic act as the main mechanism (the core of the method) that based on formulated statements determines the information proximity between the objects of analysis by logical inference. Conventionally, the method can be divided into three components. The first component of the method describes sets and basic functions performed after initialization of all objects. The second part is responsible for the process of spectral analysis and obtaining spectra of the analyzed objects. The final part is a set of criteria and statements of descriptive logic.

The process of determining information proximity begins with the initialization of all objects represented as a set of unstructured text documents $D = \{d_1, d_2,..., d_n\}$, where d_n is a text document. In turn, each d_n element of the set D is a set of lexical units $L = \{l_1, l_2,..., l_n\}$, where l_n is a lexical unit. The totality of lexical units of the set L forms meaningful semantic connections identifying the context K of each d_n element of the set D. Thus, the objects D, L and K are initialized at the first stage of the method and afterwards the performing of the functions *ReBuildTextStruct()* and *Intersection()* is followed. The purpose of *ReBuildTextStruct()* function consists of primary structuring of the set of lexical units and obtaining the data model as a set of “key-value” pairs. This is followed by the performing of *Intersection ()* function that returns the data dictionary – φ containing the same elements that are part of the primary data model obtained by *Re-BuildTextStruct()* function. Below there is an example of an algorithm fragment responsible for initializing and performing the first two functions.

```
$D \leftarrow \emptyset$
$D \leftarrow \Delta_D$
for each instance $d \in D$ do
    $firstDataModel_{RBTS(d)} \leftarrow ReBuildTextStruct(d)$,
    where $firstDataModel_{RBTS(d)} = \begin{bmatrix} fDM_{RBTS(d)_{x_1}}, \dots, fDM_{RBTS(d)_{x_n}}; fDM_{RBTS(d)_{y_1}}, \dots, \\ fDM_{RBTS(d)_{y_n}} \end{bmatrix} \Rightarrow$
    $fDM_{RBTS(d)_{x_n}} : fDM_{RBTS(d)_{y_n}}$
    function interpretation $ReBuildTextStruct(d)$:
  $return \leftarrow SELECT * WHERE \{ ?a\, V_{d_{criterion}}. ?a\, V_{d_{value}} ?b. OPTIONAL$
            $\{ ?a\, V_{d_{child}} ?c. ?c\, V_{d_{value}} ?d \}$
end for
for each instance $d \in D$ do
    for each instance $fDM \in firstDataModel_{RBTS(d)}$ do
        $\varphi \leftarrow Intersection(fDM, d)$, where $Intersection(fDM, d)$ query that checks $fDM_n \cap d_n \neq \emptyset$
        interpretation of the function $Intersection(fDM, d)$:
            $return \leftarrow fDM.SELECT\ x_k \rightarrow x_{k_i} \wedge x_v \rightarrow x_{v_i}.WHERE\ x_{v_i} = r_d.ToDictionary()$
    end for
end for
```

The initial data preparation is followed by the spectral conversion stage. The method of singular transformation was chosen as the main approach for obtaining the spectrum of information and its detailed operation can be found in the articles [Miz18, Mal19]. At this stage two functions are performed: *GetAdjacencyMatrix()* and *SVD().*The result of this stage is to obtain eigenvalues that in the terminology of spectral theory represent the spectrum of the analyzed data object of set *D.* A fragment of the second part of the algorithm is presented below.

```
for each instance $fDM \in firstDataModel_{RBTS(d)}$ do
    $M_{fDM} \leftarrow GetAdjacencyMatrix(fDM)$
    $SV_{M_{fDM}} \leftarrow SVD(M_{fDM})$, где SVD() – a method of singular transformation; $SV_{M_{fDM}}$ – singular values $\{SV_{M_{fDM_1}}, SV_{M_{fDM_2}}, \dots, SV_{M_{fDM_n}}\}$
end for
```

The final step in the method is performing the *IsSumilar ()* function that returns the result – ψ, which contains the response on the informational proximity between the objects of analysis. The *Is Similar ()* function is a set of conditions for the feasibility of the process of determining the information proximity between the objects and a set of statements suggesting the possibility of information proximity between the objects. Below the final fragment of the algorithm is given that includes a description of the *IsSumilar ()* function.

$\psi \leftarrow IsSimilar(SV_{M_{fDM}}, \varphi)$

Interpretation of the function *IsSimilar*():

$K = T \cup A$, where K – knowledge base*; T – Tbox, A – Abox.*

Conditions for *IsSimilar()* function feasibility:

1. Termination. For any $(SV_{M_{fDM}}, \varphi, T)$ function Θ gives a response $\Theta(SV_{M_{fDM}}, \varphi, T)$ in finite amount of time t^m, where *m*–$SV_{M_{fDM_x}}$ x $SV_{M_{fDM_y}} = \{(j, h) \mid j \in SV_{M_{fDM_x}}, h \in SV_{M_{fDM_y}} \wedge \varphi_x$ x $\varphi_y = \{(s, e) \mid s \in \varphi_x, e \in \varphi_y\}$.
2. Correctness. For any$(SV_{M_{fDM}}, \varphi, T)$, if concepts $SV_{M_{fDM}}, \varphi$ are feasible relative to T, then $\Theta\left(SV_{M_{fDM}}, \varphi, T\right) = 1$.
3. Completeness. For any $(SV_{M_{fDM}}, \varphi, T)$, if $\Theta\left(SV_{M_{fDM}}, \varphi, T\right) = 1$, then concepts $SV_{M_{fDM}}, \varphi$ are feasible relative to *T.*

Feasibility conditions 2 and 3 come to $U(T) = \begin{cases} \top, if\ U \vDash \top \\ \bot, if\ U \nvDash \top \end{cases}$

Statements:

1. For any concepts $SV_{M_{fDM_x}}, SV_{M_{fDM_y}}$ and terminology T there is a concept $SV_{M_{fDM_y}} \subseteq SV_{M_{fDM_x}}$, that is $T \vDash SV_{M_{fDM_y}} \equiv SV_{M_{fDM_x}} \Leftrightarrow T \vDash SV_{M_{fDM_y}} \subseteq SV_{M_{fDM_x}}$ и $T \vDash SV_{M_{fDM_x}} \subseteq SV_{M_{fDM_y}}$ и $\vDash \varphi_y \subseteq \varphi_x$.
2. There is at least one individual $i_{SV_{M_{fDM_y}}}$ such that belongs to the concept $SV_{M_{fDM_x}} \Leftrightarrow \exists i_{SV_{M_{fDM_y}}} \in SV_{M_{fDM_x}}$, that is $K \vDash i_{SV_{M_{fDM_y}}} : SV_{M_{fDM_x}} \Leftrightarrow (T, A \cup i_{SV_{M_{fDM_y}}} : SV_{M_{fDM_x}})$.

For the concept φ the statements will be similar.

Thus, if the statement 2 = ⊤, then the statement 1 = ⊤, that is there is such an interpretation $I = (\Delta, \cdot^I)$, for which $K \vDash i_{SV_{M_{fDM_y}}} : SV_{M_{fDM_x}}$.

The proof of statements is reduced to the following rules:

1. $\forall \vec{z}\{\vec{z} \vDash T \mid \vec{z} \in \Delta^I\}$
2. $\exists \vec{z}\{\vec{z} \vDash T \mid \vec{z} \in \Delta^I\}$, где $\vec{z} = \{\vec{z}_1, \vec{z}_2, \dots, \vec{z}_n\}$ – singular meanings of the concepts $SV_{M_{fDM_x}}$ и $SV_{M_{fDM_y}}$

```
if $\top^\psi$
    return ← (bool)similar ⇒ true
else if $\bot^\psi$
    return ← (bool)not similar ⇒false
end if
```

4 Example and Discussion

To test the proposed method an array consisting of more than 10 000 unstructured text documents of different subject orientation was formed. Each document in the array had a different extension and language affiliation. The server of the following configuration was selected as the test environment:

- CPU'S: 40 * Intel(R) Xeon(R) CPU E5-2690 v2 @ 3.00GHz;
- RAM: 257826 Mb.;
- OS: Ubuntu Server Edition;
- Apache: 2.4.10;
- MySQL: 5.7.21-20;
- Nginx: 1.13.4;
- PHP: 7.3.

The quality of the obtained results was assessed according to the following criteria:

- percentage of determining information proximity;
- number of identified classification groups;
- possibility of information proximity at the same spectrum and context;
- possibility of information proximity at the same spectrum at a different context;
- possibility of information proximity at the different spectrum but the same context;
- possibility of information proximity at different spectrum and context.

Thus, based on the above-mentioned criteria the results of the work (Fig. 1, 2) of the proposed method for determining the information proximity were obtained. Figure 1 shows the dynamics of determining information proximity. The diagram shows that the percentage of information proximity varies from 20% to 90%, while the average boundary for determining information proximity is ~ 61%. It is also worth noting that both curves have the same distribution that indicates the consistency of the results obtained after comparing the two operating modes of the algorithm (statements 1, 2).

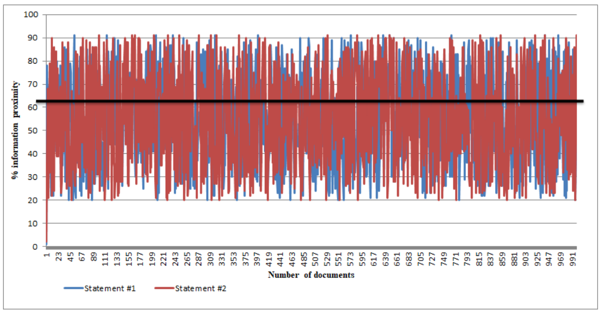

Figure 1: Dynamics of determining information proximity

Figure 2 shows the final distribution of unstructured documents array. This distribution shows that 10 highest priority categories were identified according to the results of the algorithm among which there was a further classification of the analyzed documents. Each identified classification group included from 6% to 10% of the documents out of the total number of contained in the array. Each identified classification group included from 6% to 10% of the documents from the total number contained in the array.

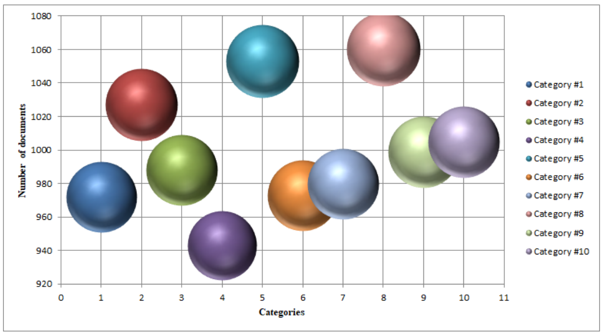

Figure 2: Distribution of documents by classification categories

5 Future Research

The process of determining the information proximity between the analyzed text documents in large arrays of information has significant potential in the problems of semantic classification. The data sets obtained as a result of testing can be used to build up more comprehensive thematic dictionaries of the subject areas that can be used in management decision-making systems and situational management. In addition, the process of deriving the results of logical statements can be accompanied by visualization to reflect the map of semantic relations between various text documents in information arrays more fully.

6 Conclusion

The article considers a method suggesting an alternative approach to the problem of determining information proximity between sets of objects represented as a set of unstructured text documents. The obtained results of the experiment show a high degree of information proximity determining with an optimal ratio of the execution time of all operations and the use of computational resources. Based on this it is proposed to apply this method to problems of semantic classification in decision-making information systems to classify emerging situations and identify redundancy in large arrays of textual information, thereby reducing the amount of necessary stored information and response time to an incoming query.

Acknowledgment. The reported study was funded by the Russian Foundation for Basic Research, according to the research projects No. 19-01-00246-a, 18-01-00402-a.

References

[Lij10] C. Lijun, Y. Hongkui, L. Yuxiang & L. Xiyin: Research and exploration of text mining technology. 2nd International Conference on Advanced Computer Control, vol. 5, pp. 435-439 (2010).

[Jia14] M. Jiang, Y. Zhou, X. Fan, O. Wang, X. Zhang, Z. Zhang, J. Lian and Z. Pei.: A Variety of Text Mining Technology and Tools Research. 2014 International Conference on Mechatronics, Electronic, Industrial and Control Engineering, pp. 918-921 (2014).

[Sar19] D. Sarkar: Semantic Analysis. In: Text Analytics with Python. Apress, Berkeley, CA (2019). https://doi.org/10.1007/978-1-4842-4354-1_8

[Bou14] A. Bouaziz, C. Dartigues-Pallez, C. da Costa Pereira, F. Precioso, P. Lloret: Short Text Classification Using Semantic Random Forest. In: Bellatreche L., Mohania M.K. (eds) Data Warehousing and Knowledge Discovery. DaWaK 2014. Lecture Notes in Computer Science, vol. 8646, pp. 288-299. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10160-6_26

[Ma16] H. Ma, R. Zhou, F. Liu, X. Lu: Effectively Classifying Short Texts via Improved Lexical Category and Semantic Features. In: Huang DS., Bevilacqua V., Premaratne P. (eds) Intelligent Computing Theories and Application. ICIC 2016. Lecture Notes in Computer Science, vol. 9771, pp. 163-174. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-42291-6_16

[Tan19] S. Tandel, J. Abhishek and D. Siddharth: A Survey on Text Mining Techniques. 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 1022-1026 (2019).

https://doi.org/10.1109/ICACCS.2019.8728547

[Fad18] S. Fadiya and A. Sari:. The importance of big data technology. International Journal of Engineering & Technology, vol. 7, pp. 485 (2018).

[Ous17] A. Oussous, F-Z. Benjelloun, A. Ait Lahcen & S. Belfkih: Big Data Technologies: A Survey. Journal of King Saud University - Computer and Information Sciences, vol. 30, pp. 431-448 (2017). https://doi.org/10.1016/j.jksuci.2017.06.001

[Jus18] C. Justicia, D. Sánchez, I. Blanco and M. Martin-Bautista: Text Mining: Techniques, Applications, and Challenges. International Journal of Uncertainty Fuzziness and Knowledge-Based Systems, vol. 26, pp.553-582 (2018). https://doi.org/10.1142/S0218488518500265

[Zha14] Z. Zhan, F. Lin, X. Yang: Semantic Similarity Calculation of Short Texts Based on Language Network and Word Semantic Information. In: Wu J., Chen H., Wang X. (eds) Advanced Computer Architecture. Communications in Computer and Information Science, vol. 451, pp. 215-228. Springer, Berlin, Heidelberg (2014). https://doi.org/10.1007/978-3-662-44491-7_17

[Fis08] J.M. Fishbein, C. Eliasmith: Methods for Augmenting Semantic Models with Structural Information for Text Classification. In: Macdonald C., Ounis I., Plachouras V., Ruthven I., White R.W. (eds) Advances in Information Retrieval. ECIR 2008. Lecture Notes in Computer Science, vol. 4956, pp. 575-579. Springer, Berlin, Heidelberg (2008) https://doi.org/10.1007/978-3-540-78646-7_58

[Vas17] H.L. Vasudeva: Spectral Theory and Special Classes of Operators. In: Elements of Hilbert Spaces and Operator Theory. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-3020-8_4 [God77] C. Godsil, D.A. Holton, B. McKay: The spectrum of a graph. In: Little C.H.C. (eds) Combinatorial Mathematics V. Lecture Notes in Mathematics, vol. 622. Springer, Berlin, Heidelberg (1977). https://doi.org/10.1007/BFb0069184

[Kri18] A. Krisnadhi, P. Hitzler: Description Logics. In: Alhajj R., Rokne J. (eds) Encyclopedia of Social Network Analysis and Mining, pp. 572--581. Springer, New York, NY (2018). https://doi.org/10.1007/978-1-4939-7131-2_108

[Miz18] G.S. Miziukov: Finding similarity between unstructured data objects on the basis of the method of singular decomposition of the spectrum of a graph (2018). http://www.ivdon.ru/uploads/article/pdf/IVD_19_Miziukov_N.pdf_e0a3d9ae84.pdf. Web-Accessed 10 Nov 2019

[Mal19] M.M. Malamud: On Singular Spectrum of Finite-Dimensional Perturbations (toward the Aronszajn–Donoghue–Kac Theory). Dokl. Math, vol. 100,pp 358–362 (2019). https://doi.org/10.1134/S1064562419040124