Modernized Mathematical Model of Text Document Classification

Tetiana Golub[1][0000-0001-6024-008X]

[1]Zaporizhzhia National Technical University, Zhukovsky str.,64, Zaporizhzhia, 69063, Ukraine

golub.tv6@gmail.com

Abstract. The modernized mathematical model of the main stages of the text document classification is proposed. It takes into account the characteristics of certain categories. A mathematical description of the document data set creating stages, a document classification into categories is proposed. The principles of reducing the feature space dimension are described and the proposed method what used for determining the term weights is argued. The application of the method proposed in the article leads to reduce the analysis time of each document in order to make a decision about its category. This leads to decrease the resulting time for the analysis of the entire document set.

Keywords: text document classification, term vector, mathematical model, term weight, SLF parameter

1 Introduction

The information amount witch presented in text form increases continuously. Text information is accumulated in all areas of human activity. It is represented from data stored on personal computers to data in the form of Big Data. It covers such areas as business, research institutions, government and financial institutions that use technology intensively. Text information contains statistical data, control commands, reference information and principle laws of different processes. A feature of such information is the lack of its structuredness. It makes more complicated the process of its analysis [1].

Text analytics converts text into numbers. It allows organizing data and helps to identify patterns. Structured data are easier to analyze. Therefore, decisions made on their basis are more quality [2, 3].

If it is necessary to find the information in a data large amount, firstly it must be classified [4]. This process is the consideration subject in the proposed study.

Text classification refers to one of the computational linguistic tasks. It includes the definition of the text thematic affiliation, the text author, the statement emotional coloring and etc.

The task of organizing documents is solved to simplify the search for the necessary information. It is one of the most urgent tasks. Text classification is needed to solve

this problem. [5]. It is difficult to solve the classification problem because the data flow is constantly increasing. Therefore, its decision is relevant.

Many approaches to solving this problem are described in the literature. An overview and comparison of currently relevant methods are presented in accordance with the various stages of this process in [1, 6–8]. According to these sources one of the most important points of test classification is key feature selection. The works [9–15] were devoted to solving this problem. Various approaches, including statistical, frequency, latent-semantic and others are disclosed there. However, the described methods consider terms within the entire document collection. It is not possible to assess the importance of a separate term for each category separately.

The classification of text documents is the process of analyzing its content and automatically defining a document into one or several categories [16, 17]. Categories are sets of documents with a common theme. The set of categories is set by the expert or is determined automatically on the basis of the training sample. Automatic classifier is used in the information-analytical system at the stage of processing documents. An automatic classifier is a program that determines the subject of documents and assigns them to categories [6].

The inverse problem is also relevant. It consists of document selection from a document set according to the category defined by the user.

Presented in the literature mathematical models do not consider the term importance for certain categories. The author offers an improved mathematical model that takes into consideration this parameter.

The proposed in the article model considers this parameter which allows reducing the time for assessing the belonging of a document to certain categories by reducing the size of the term vector of certain categories for the text document classification.

2 Task formalization

The classifying document process in a formal form can be described as follows. The text document classification will be understood as the task of automatically defining a document into one or several categories based on its content. The category will be understood as a variety of documents with a general theme. Many categories are set by an expert or determined automatically using a training set. Automatic classifier is used in the information-analytical system at the document processing stage [6].

Mathematical models of the text document classification process given in [18, 19] are common. The author proposes the improvement of the existing variants of the term weight determining process as a part of the classification process with considering requirements of the task in this article.

It is proposed next designations to formally describe the process of text documents classifying:

— $T=\{t_1,...t_{|A|}\}$ – document term set;
— $B=\{b_1,...b_{|B|}\}$ – term set;
— $D=\{d_1,...d_{|D|}\}$ – documents set;
— $C=\{c_1,...c_{|C|}\}$ – category set;

— $E=\{e_1,...e_{|E|}\}$ – category term set.

In the general case, the searching task of documents which corresponding to a particular category is following.

A set of documents D, from which it is necessary to choose those documents d_i, which most likely belong to the category c_i determined in advance from the set of categories C exists. The solution of this problem is considered in this article.

$$\Phi(d_j, c_i) = \begin{cases} 0, & if \quad d_j \notin c_i \\ 1, & if \quad d_j \in c_i \end{cases}. \tag{1}$$

3 Document term set creating

The text document classification is performed using the analysis of the text document terms. A term is an intuitively defined expression of a formal language. It is the formal name of the object [2]. In this study, the term will be understood as the word obtained after stemming. Stemming is the reduction of a word to a certain normal form using the clipping of its endings and suffixes. The formation of terms is one of the tasks of the preprocessing stage.

The text is presented in the form of a document term set model for solving the classification problem. Each term has its own weight.

Text preprocessing is performed when determining whether a document belongs to any category, considering the importance of each term.

The preprocessing process has the following characteristics suggested by the author:

— $T \in B$ - all terms of the document are included in the set of possible terms;

— $E \in B$ - all terms of the category are included in the set of possible terms;

The set of elements of the sets T and E forms the set B. The sets T and E constitute the set B.

— $T_M = <n_1(t_1), n_2(t_2), ... n_{/T/}(t_{/T/})>$ – a multiset of the set T. It allows collecting the occurrence of the set elements several times.

The formation of a document multiset of one group, category, allows its power for each term determines. This parameter estimates the quantitative index of the term occurrence.

— $E_M = <n_1(e_1), n_2(e_2), \ldots n_{/E/}(e_{/E/})>$ – a multiset of the set E. It allows collecting the occurrence of the set elements several times.

Based on the category multiset, it is possible to determine the term indicators by the power of their occurrence. Analyzing this parameter it is possible to determine in how many categories of the collection the considered term occurs at least once. It allows distinguishing terms that are characteristic for all categories and are not characteristic to a particular category. These terms do not contain information for classification. Therefore it is possible to exclude them from the analyzed set.

Subsequent text processing is performed based on these characteristics.

All words that appear in the documents can be ordered in some way, for example alphabetically. Then, for each document it is possible to write out the entire set of weights matching the dictionary words. If some term is out of the document, then the weight will be zero. That is the vector will be:

$$d_i = (w_1, w_2, \ldots, w_n), \quad (2)$$

where d_i - i-th document vector representation, w_i - weight of the i-th document term, n - the total number of different terms in all the documents of the collection –the power of set B [21]

4 Term weight identification

The term weight values of the set E for each category of the incoming set B are determined to assess the occurrence to the document category.

There are many methods to determine the term weights in the literature are presented. Some of them are:

— Boolean weight. w = sign(tf), i.e. 1 - if word occurs in the document, 0 - otherwise;
— w=tf – number of word duplications in the document [3];
— w=tf/df - the coefficient «tf•idf», i.e. the multiplication of the words occur-rence frequency (tf), to the reciprocal value of the words occurrence frequency in all documents of the collection (inverse df). There are many options to define the weight value of the i-th term (wij) in the document dj. One of the simplest options is the following: wij= tf•log10(1/df). When the formulas «tf•idf» are used the problem of common words is solved – when the words with no meaning are of high weight;
— SLF parameter [3];
— Latent semantic analysis [7, 8].

The mentioned methods for determining the term weight values characterize that terms within a single document or within the entire collection as a whole. The importance and significance of a term within a single category is ignored in both cases.

The SLF parameter [3], used to determine the weight values of each term of the set E, compensates for this disadvantage. The parameter SLF is a coefficient that characterizes the assessment of terms with regard to their inclusion in the category. This

method considers the importance of each term for a particular category, unlike many other approaches to determining weight values.

The following parameters were defined to find the SLF parameter:

1. df_{tc} - the number of documents of category c, in which the term t occurs at least once;
2. N_d - the number of documents in the category c;
3. NDF_{tc} - normalized frequency of occurrence of the term t in the category c. It is found as the ratio of the document number of the category c, in which the term t occurs at least once, to the number of documents in the category c. This estimate is local to the category.

$$NDF_{tc} = df_{tc}/N_c \tag{3}$$

4. SLF_t - logarithmic sum of the term t frequencies:

$$SLF_t = log(|C|/\sum(NDF_{tc})) \tag{4}$$

The SLF_t indicator eliminates the imbalance between categories with small and with a large number of documents.

The SLF parameter for each term within the collection is determined according to the formula (5)

$$TFSLF_t = TF_t(E_{/t/}) \cdot SLF_t, \tag{5}$$

where $TF_t(E_{/t/})$ – the frequency of the term belonging to the set B. It is defined as the ratio of the certain term number occurrences to the total number of the document terms. Thus, the importance of the term t_i within a separate document d_j is estimated [9].

Vector B_T, with the weight coefficient values of the set T terms within the entire collection as a whole, will be obtained. In this case, the significance of terms of a particular category is not fully considered. It reduces the quality indicators of the classification implementation of texts belonging to similar in meaning and used words topics.

The *SLF* parameter considers the term importance for categories within the collection, but does not take into account the importance of terms for each category separately. The following modification of the term weight definition based on the given parameter is proposed by the author for solving this problem.

The author proposes a sequence of actions for defining non-informative terms for each category individually based on the *SLF* parameter and statistical data. And further removal of these terms from the term vector of a separate category.

The sequence of actions to determine the weight values of the terms of the set E of each category for each category term e_i:

- the coefficient *tf/df* for each category term e_i is determined;
- the value of the weight of each term by categories is determined;
- uncharacteristic terms for each category are identified and removed.

The coefficient *TF* for each category term e_i within the collection as a whole is defined as the ratio of the total number of each term within a separate category to the total number of each term within the collection as a whole (6).

$$TF(t_i, c_j) = \frac{fr_{ij}}{\sum_i fr_{ij}}, \tag{6}$$

where $0 \le i \le |E|$, $0 \le j \le |C|$.

The importance of the term τ_i within a single document d_j is evaluated. [14]

The weight of each term by category, taking into account its occurrence in collection categories (the set *E*, containing the $CTFSLF(t_i c_j)$ values of each category term) is defined as the product of the *TF* coefficient for each term of individual categories and the *SLF* parameter:

$$CTFSLF \quad (t_i, c_j) = TF\ (t_i, c_j) * SLF_k, \tag{7}$$

where $0 \le i \le |E|$, $0 \le j \le |C|$, $0 \le k \le |B|$.

The *CTFSLF* method for determining the term weights makes it possible to take into consideration the term importance within a particular category.

5 The feature space dimension reducing

The computational complexity of various classification methods directly depends on the feature space dimension. Therefore, the stage of the used term number reducing, or the stage of reducing the dictionary size of the category $|B|$ to $|B'| << |B|$, often is performed for classification problem solving.

The purpose of this stage is to reduce the data set dimension. This goal is achieved by removing uninformative for classifying terms. It allows decreasing the data size, to reduce the computing power requirements of the algorithm [4].

In this case, each documents terms vector undergoes the following preliminary processing:

- elimination of stop words (often used and not carrying a semantic load such as unions) [5];
- performing a morphological analysis of words [5];
- using clustering methods [6].

The following method of terms vector size reducing on the basis of the modernization described previously is proposed by author. It consists of the stage of determining non-characteristic terms for separate categories and the stage of their remove.

The value of K_j is calculated to determine the threshold value. The value of K_j is calculated as the inverse value of the number of documents which belongs to the analyzed categories. It is used to remove non-informative terms.

The term weight describes the property of its belonging to certain category. Terms that are found in all categories are low weight. Terms whose weights are below threshold are excluded.

$$K_j = \frac{1}{|D_j|} \tag{8}$$

where $0 \le j \le |C|$.

Further, the weight value is compared with the threshold value for each collection term.

If the value of the term weight is less than the threshold, this value is equaled to zero:

$$\Psi(e_i, c_j) = \begin{cases} 0, & if \quad e_i < k_i \\ CTFSLF\,(t_i, c_j), & if \quad e_i \ge k_i \end{cases} \tag{9}$$

where $0 \le i \le |E|$, $0 \le j \le |C|$.

The given analysis allows us to identify and exclude from the analysis such terms with low informativeness, as often encountered in the categories in the document corpus, and which are not informative for classification.

Thus, the removal of the terms distinguished from the feature space as a result of the analysis will reduce the length of the analyzed set and simplify the classification task.

The resulting term vector is used to search for documents belonging to a particular category, using the classification process.

6 Document classification into categories

In general, the task of classifying documents into categories is to find the maxi-mum sum value of the term weighted coefficients that coincide with the terms characterizing a separate category.

The following parameter is introduced by author to evaluate this indicator.

W – a set that indicates the degree which shows this document falls into a separate category. A set is defined as the intersection of the document set T and the corresponding categories set E. All terms that are included in both sets are included in the set W.

$$W = T \cap E \tag{10}$$

The estimated value of the belonging degree of document to a separate category can be defined as the sum of the products of the set W elements by the corresponding weight values Ψ for terms belonging to the set T.

Then the degree of document compliance to a separate category can be determined as follows.

$$NW_d = \sum_i W(t_i) \cdot TFSLF\ (t_i, e_j) \tag{11}$$

where NW_d – the normalized value, the degree of coincidence of the term set belonging to category T to the term set of category *E*.

When a document and category match, this parameter will have a maximum value relative to other categories, and when comparing a document with a foreign category, the match will be observed mainly only for common words that can be attributed to several categories and whose significance decreases with increasing number of these categories.

7 Classification stage time reducing

The application of the method proposed by the author will reduce the spent time at the classification stage.

According to the property of additivity, the resulting value of time spent on the classification of the n documents is equal to the sum of time spent on the classification of each document separately. That is, the resulting value of time is determined by adding the individual time spent on the classification of each document. It is proposed next designations:

- A –total number of documents for classification;
- $S=\{s_1,...s_{|A|}\}$ –the set containing the time spent on the classification of each document analyzed sample;
- *S1* - the set containing the time spent on the classification of each document analyzed sample using based method;
- *S2* - the set containing the time spent on the classification of each document analyzed sample using proposed method.

The total time to perform classifications of all documents using the methods *S1* and *S2* is determined:

$$S_\Sigma = \sum_i s_i \tag{12}$$

According to the properties of commutativity and associativity for the addition operation, the elements of the sets *S1* and *S2* can be grouped into two groups. The first group consists of the sum of expenditure time equal in total value for both sets. The second group consists of the summands whose total values differ. If the different total values from the second group of the sample $S2_i$ are less than the different total values of the sample $S1_i$, then it can be argued that the sum of the sample *S2* is less than the sum of the sample *S1* that is presented in (13).

$$if \quad s1_i > s2_i \quad than \quad S1_\Sigma > S2_\Sigma \tag{13}$$

Thus, analyzing the obtained results, it can be argued that the shorter the time spent on implementing the classification process of each document separately, the shorter the time value of implementing the classification as a whole. Since reducing the time spent on the classification of a certain document leads to a decrease in the time spent on the classification as a whole. So, this task is relevant.

8 Proposed method testing

The task of document classifying by individual categories of class 004 " Computer science and technology. Computing. Data processing" of the UDC classifier was selected for testing the proposed method. Certain categories are:

- 004.0 " Special auxiliary subdivision for computing",
- 004.2 " Computer architecture",
- 004.4 "Software",
- 004.9 " Application-oriented computer-based techniques".

30 documents of each category were used as a training sample. Categories of documents were determined by their authors. Testing was conducted on unused for training documents for each category. The training and testing results are shown in tables 1-2.

Table 1. Term vector size after learning stage

Category	Words in docum	SLF		CTFSLF		Excluded words	Decreasing part
		Terms in vector	Term part	Terms in vector	Term part		
004.0	148419	22118	14,90%	18450	12,43%	3668	16,58%
004.2	111213	12510	11,25%	8978	8,07%	3532	28,23%
004.4	108077	18752	17,35%	14652	13,56%	4100	21,86%
004.9	104207	17411	16,71%	13473	12,93%	3938	22,62%
Average result	–	–	15,05%	–	11,75%	3809	21,53%

Table 2. Spent time for testing stage

Category of document	Time for SLF, s	Time for CTFSLF, s	Decreasing time, s	Decreasing part of time
004.0	0,03125	0,02500	0,006251	20,00%
004.2	0,018751	0,01250	0,006249	33,33%
004.4	0,021877	0,021875	0,000002	0,01%
004.9	0,028126	0,015627	0,012499	44,44%
Summary /	0,100004	0,075003	0,025001	24,44%

average result				

As can be seen from table 1, the terms average proportion of the words in documents total number according to the original SLF method is 15.05%. The proposed CTFSLF method shows a result of 11, 75%. The average number of terms excluded from each category is 21.53%. As a result, the average time for determining the category of a document was reduced by 24.44% (table 2). This shows the promise of the proposed method.

9 Conclusions

Thus, this article a modernized mathematical model of the text document classification main stages taking into account the characteristics of certain categories proposed. A mathematical description of the document data set creating stages for a document classification into categories is proposed. The principles of reducing the feature space dimension are described and the proposed method using for determining the weights of terms is argued.

The purpose of the proposed approach is to identify and exclude non-informative terms for a particular category, i.e. leave inherent informative terms that characterize the category. The using of this approach leads to reduce the amount of computations performed for searching in the general collection of documents belonging to a particular category. As a result, the analysis time to classification of certain document is reduced. This leads to reduce the resulting time for analyzing the entire set of documents.

References

1. Thangaraj M., Sivakami M.: Text classification techniques: A literature review. Interdisciplinary Journal of Information, Knowledge, and Management, 2018, vol. 13, pp. 117-135 (2018)
2. Brindha S., Sukumaran S., Prabha, K.: A survey on classification techniques for text mining. 3rd International Conference on Advanced Computing and Communication Systems. IEEE. Coimbatore, Indi,. (2016) doi: 10.1109/ICACCS.2016.7586371
3. Daud A., Li J., Zhou L.: Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey. Frontiers of computer science in China, 2010, vol. 4, no. 2, pp. 280–301 (2010)
4. Korde V., Mahender N.: Text classification and classifiers: a survey. International Journal of Artificial Intelligence & Applications (IJAIA), 2012, Vol. 3, no. 2, pp. 85–99 (2012)
5. Pankov S. V., Shebanin S. P., Ribakov A. A.: Thematic classification of text. ROOKEE, ROMIP 2010, Kazan', Russia, 2010, pp. 142-147 (2010)
6. Golub T. The Analysis of text documents classifiers constructing methods, Modern problems of radio engineering, telecommunications, and computer science, 2016, pp.742-745 (2016)
7. Yang Y., Zhang J., Kisiel B.: A scalability analysis of classifiers in text categorization. ACM SIGIR'03, (2003)

8. Sebastiani F.: Machine learning in automated text categorization. ACM computing surveys (CSUR), 2002, vol. 34, pp. 1-47 (2002)
9. Karpovich S.N.: Multi-valued text documents classification using probabilistic thematic modeling ml-PLSI. SPIIRAS Proceedings. 2016. vol. 4(47), pp.92 – 104. (2016) doi: 10.15622/sp.47.5.
10. Kuralegov I.: Automatic classification of documents based on latent semantic analysis. 1st International Conference Digital Libraries: Advanced Methods and Technologies, Digital Collections, St-Petersburg, Russia, 1999, pp. 89-96. (1999)
11. Andreev A. M.: Automatic classification of text documents using the neural network algorithms and semantic analysis. Advanced Methods and Technologies, Digital Collections, St-Petersburg, Russia, 2003, pp. 76-86. (2003)
12. Krasnov A., Ilatovskiy A.S., Khomonenko A.D., Arsen'yev V.N.: Evaluation of documents semantic proximity based on latent-semantic analysis with automatic selection of rank values. SPIIRAN proceedings, 2017. no. 5(54), pp. 185-204 (2017)
13. Rehman Abdur, Barbi H., Saeed M., Feature Extraction for Classification of Text Documents. International Conference on Communications and Information Technology (ICCIT 2012), Hammamet, Tunisia, 2012, pp. 234 - 239. (2012)
14. Budanitsky A. Hirst G.: Evaluating WordNet-based Measures of Lexical Semantic Relatedness Computational Linguistics. 2006. Vol. 32. pp. 13-47 (2006)
15. Bondarchuk D.V. Vector model of knowledge representation based on semantic proximity of terms]. Bulletin of SUSU. Series: Computational Mathematics and Computer Science. 2017. vol. 6 no. 3. pp. 73–83 (2017) doi: 10.14521/cmse170305.
16. Tsoumakas G., Katakis I.: Multi-label classification: an overview. International Journal of Data Warehousing & Mining. 2007. vol. 3(3). pp. 1–13 (2007)
17. Rubin T.N., Chambers A., Smyth P., Steyvers M.: Statistical topic models for multilabel document classification. Machine Learning. 2012. vol. 88. no. 1–2. pp. 157–208 (2012)
18. Erpev A.S.: Automatic classification of text documents. Mathematical Structures and Modeling. 2010, vol. 21, pp. 65-81 (2010)
19. Zyuz'kov V. M.: Mathematical logic and theory of algorithms. Tomsk, El Content (2015)
20. Willett P. The Porter Stemming Algorithm: Then and Now Program: Electronic Library and Information Systems. 2006. vol. 4, no. 4. pp. 219-223 (2006)
21. Golub T.V., Tyahunova M.YU.: The method of Ukrainian language stitemming for the classification of documents based on Porter's algorithm. Scientific papers of the Donetsk National Technical University. Series: Informatics, Cybernetics and Computing 2017, no. 1, pp. 59 – 63 (2017)
22. Oliynyk YU. O., Katyushchenko D. O.: Analysis of the methods of determining the text documents signs weight. Scientific Review, 2018, 3(46), pp. 112 – 123 (2018).