APPROACHES TO AUTOMATION PROCESSING OF USER REQUESTS IN A MULTI-LEVEL SUPPORT SERVICE USING FEATURED MODELS

Bobryakov Alexander, Kurilyov Vladimir, Mokhov Andrey & Stefantsov Alexey

This Publication has to be referred as: Bobryakov, A[lexander]; Kurilyov, V[ladimir]; Mokhov, A[ndrey] & Stefantsov, A[leksei] (2019). Approaches to Automation Processing of User Requests in a Multi-Level Support Service Using Featured Models, Proceedings of the 30th DAAAM International Symposium, pp.0936-0944, B. Katalinic (Ed.), Published by DAAAM International, ISBN 978-3-902734-22-8, ISSN 1726-9679, Vienna, Austria
DOI: 10.2507/30th.daaam.proceedings.130

Abstract

The article is devoted to the development of the functional subsystem for automatic classification of incoming requests from users of Methodological, Technical and Consulting Support of the Ministry of Education and Science of Russia. Text mining tools are used to classify incoming requests. The paper considers two text representation models - Bag of words and Word2Vec, and compares different classification methods, such as logistic regression, random forest, and gradient boosting. To improve the results, optimization approaches are used - class system restructuration and introducing the technology of two-level classification.

Keywords: Text Mining; Classification; Technical and Consulting Support; User requests.

1. Introduction

Informatization and automation of economy budget sectors are currently the main factors determining their functioning and effectiveness. The successful development of the industries such as education, health care, social services, public administration, transport, energy, agriculture is currently impossible without modern digital technologies.

Automation and information support of management processes in the above mentioned areas involves the development and implementation of an integrated complex of Information Systems (IS) and services that automate the basic accounting functions, as well as the functions of statistical and analytical information processing and Management Decision Support (MDS) [1].

The main feature of the administrative structure of the budget sector is a centralized state administration. As a rule, this role is played by ministries that perform both financial and subject matter management of numerous Subordinate Organizations and Institutions (SOI).

That is why Information Systems (IS) at the industry level are distributed computer systems with:

- a powerful central server platform, which operates, as a rule, on the basis of the Ministry and solves the problems of collecting primary planning and reporting information, its statistical and analytical processing and MDS at the industry level;
- a number of peripheral nodes operating on the basis of individual SOI and solving the problems of delivering information to a central server platform and MDS at the organization level.

The effective operation of such industry complex of IS requires the single special centralized service of Methodological, Technical and Consulting Support (MTCS) to provide high-quality consulting and information services to managers and employees of specific SOI, both in terms of methodological support of budgeting process and subject management, and also in terms of technical support of IS. The structure and main functions of the MTCS implemented at the Ministry of Science and Higher Education, one of the largest budget departments in Russia, are described in [2]. The Methodological, Technical and Consulting Support (MTCS) service is a distributed hierarchical organizational and technical structure consisting of:

- the Center for Receiving and Routing Requests (CRRR) - the first level of MTCS;
- a group of Subject-oriented Centers of Competence (SCC) - the second level of MTCS.

The main operational principles of MTCS are the following:

- receiving and registration of user requests from CRRR coming through various communication channels
- formation of responses to typical user requests at the CRRR level using a specialized Knowledge Base (KB);
- routing (sending to the appropriate SCC) user requests for non-standard questions;
- formation of responses to atypical user requests at the SCC level

The main criterion for the effective performance of MTCS is the high speed and accuracy of responses to user requests. The main parameters determining the value of this criterion are the time of the initial analysis stage of the incoming requests and its processing time, which is either the time to search for a response to user requests in a specialized database at the CRRR level (in the case of typical request), or the time of its further routing to the corresponding SCC (in case of non-standard atypical request). The desire to achieve high values of the efficiency of the criterion requires, first of all, solving the problems of automating the implementation of these stages.

One of the approaches to improve the efficiency of solving the problems of analyzing and routing requests by the content of incoming requests is the use of models, methods and means of automatic classification of incoming requests. This task is a typical one, solved within the framework of the Text and Data Mining using Machine Learning methods, such as decision trees, naive Bayesian classifier, neural networks, support vector machine and others, as well as natural language processing methods [3],[4],[7].

For the practical implementation of models, methods and means of automatic classification of user requests, the MTCS of the Ministry of Education and Science of Russia has decided to introduce a Functional Subsystem of Automatic Classification (FS AC) of incoming user requests in the structure of the information complex.

It has been considered approaches to architecture design, selection and adaptation of model and algorithmic support for a functional subsystem of automatic classification of incoming user requests, and it has also been presented the results of practical testing of the proposed models and methods on the example of the work of MTCS on financial and economic activities of institutions and organizations subordinated to the Ministry of Education and Science of the Russian Federation.

2. The design of a operational subsystem for automatic classification of user requests

The proposed approach to automating the processing of user requests in MTCS of the Ministry of Education and Science of Russian Federation provides for the implementation of the following main stages of work:

1. design of the architecture of an operational subsystem for automatic classification of user requests;
2. selection and adaptation of model and algorithmic support for solving the problems of automatic classification of incoming user requests of MTCS, including:
 - analysis of the real array of user queries collected during the work of MTCS service in 2018, as well as the applicable class nomenclature in order to select the most effective mathematical and data processing methods;
 - selection and implementation of specific measures to improve the efficiency of the proposed models and methods.

For the design of FS AC, user requests are considered as short unstructured text documents received through various communication channels. The formal statement of the classification problem for such text documents at this stage is as follows:

There is X – a number of documents (user requests), and a fixed number of classes: $Q = \{Q_1, Q_2, \dots, Q_k, \dots, Q_K\}$ where Q_k is the label of k - class, $k=1, \dots, K$, K – is the known number of classes. A number of documents $\vec{X} (\vec{X} \in X)$ are described by features (terms, words). There is an unknown objective function (classifier) J ,

$$J: \vec{X} \rightarrow Q \quad (1)$$

Each class Q_k has a unique description. One and the same document $\vec{X} (\vec{X} \in X)$ is assigned on the basis of its terminological composition to several classes at once. Moreover, the classification of a document is determined by a certain degree of confidence of the classifier. At the learning stage, it is necessary to construct a classifier J^* , that is as close as possible to J in the selected system of features $(\vec{X}^T = [x^{(1)}, \dots, x^{(M)}]^T, M$ – is the number of terms, selected to describe the documents in X).

Classifier J^* is built on a set of documents marked up by an expert (“teacher”) and divided into two parts – training and test samples.

The designed architecture of the operational subsystem for automatically classifying incoming user requests in MTCS of the Ministry of Education and Science of Russia is shown in Figure 1 and it includes:

- an inbound request transport module that delivers an incoming request to the classification module;
- a classification module that, on the basis of attribute models and classification algorithms, assigns an incoming request to one or more classes from a given set;
- a database containing exact or “close” responses to user requests stored in a special structured form;
- a search module that finds exact or “close” information in the database based on the classes obtained after processing the request in the classification module. The found answer generally contains explanatory information upon request and regulatory, reference, administrative and other information necessary for the user;
- operator workstation, receiving and processing information found in the search module upon request. The operator generates a response, passes it to the user and sends it to the database for memorization. In the absence of a “suitable” answer, the operator sends a request to the expert
- a training module that provides training and retraining of the classifier. The training sample is constantly modified and based on the accumulated answers to users and it is used to iteratively refine the parameters of attribute models and classifiers;
- administrator workstation, managing the classifier training process, setting model parameters and the search module, optimizing the structuring of information in the database, monitoring the classification and search quality indicators, and administering the operation of the classification subsystem as a whole.

Fig. 1. The architecture of the operational subsystem for the automatic classification of incoming user requests in MTCS of the Ministry of Education and Science of Russia

3. Analysis of a real array of user requests and applicable class nomenclature

To solve the problem of designing the model and algorithmic support of FS AC and assessing the effectiveness of the proposed models and algorithms, there has been used a full sample of user requests received by MTCS of the Ministry of Education and Science of Russia in 2018 on the basis of 24,556 requests. User applications entered the OTRS user request processing system [10], which operates as part of the MTCS.

The analysis of vocabulary and morphology in the subject area of financial and economic activity of SOI showed that the terminology used there has a clearly defined set of words, the number of options for constructing phrases in queries is also not very large, “unpredictability of expressions” in the query text is limited by computer and general cultural literacy of users. These facts significantly reduce the “fuzziness” of the user's request and increase the efficiency of solving the classification problem of such requests.

The applied classification system has included 31 classes:

- Typical requests processed by the request receiving and routing center (first level of MTCS);
- 27 classes of the second level (the level of the corresponding SCC), including non-standard atypical user requests for various subject areas of financial and economic activity;
- 3 service classes of the first level, including user requests related to system maintenance, as well as a small number of incorrect or very poorly formulated requests.

The distribution into classes is shown in Figure 2.

Fig 2. Classification of user requests. Designations of classes: first level of MTCS: 1 class - “CRRR”; second level of MTCS (SCC and service classes): 2 - “Navigator of the entrant”, 3 - IS SOI, 4 - SOI, 5 - ACS FEAP (Financial and Economic Activity Plan), 6 - Seminars-meetings, 7 - Incorrect questions, 8 - MR, 9 - Do it right 2.0 - MPEI, 10 - IS MR, 11 - Budgeting and planning of FEA, 12 - Salaries of MES, 13 - Performance reports, 14 - Calculation of subsidies, 15 - FM rating, 16 - Scholarships, 17 - AC of MPEI FMCS, 18 - IAC CBD, 19 - Do the right thing 2.0, 20 - Public obligations, 21 - Regulatory costs, 22 - Continuing education, 23 - Ongoing financing, 24 - Transfer to ACS FEAP, 25 - Lists of services, 26 - Postmaster, 27 - AC of FMCS, 28 - the Misc, 29 - Sandbox, 30 - IFC and IFA, 31 - IS assembly.

According to the results of the sample analysis at this stage, the specificity of the data were determined that can subsequently affect the resulting quality of the solution to the problem:

- Samples are unbalanced - class sizes vary greatly.
- The requests contain noisy data: links, stop words, confidential data, temporary data that do not carry any informative part.
- Additionally, we need to get rid of unshielded characters and English words

As a result of the preparation of the initial data, preliminary data processing was made, which included the removal of non-informative words, the removal of English words, the concatenation of prefixes (no, not, neither nor, without), the removal of single characters and the normalization of words.

As a result of pre-processing, test requests, as well as requests consisting of a random set of letters and queries written in English, were deleted. This led to a reduction in the number of requests to 21,637. The updated sample was divided into training and testing samples in the ratio of 70% to 30%. As a result, the entire sample of 21,637 requests was divided into a training sample of 15,145 requests and a testing sample of 6,492 requests. The analysis of the array of requests and the applied classification system allowed us to conclude that fixed class names and a certain “predictability” of requests allow us to apply characteristic models for the representation and analysis of text queries based on a simple mathematical apparatus.

4. Selected feature models and methods for presenting and analyzing text requests

The following models were selected as feature models for the presentation and analysis of text requests:

- “Bag of words” model (BOW, bag of words), in which the text is characterized by a set of its meaningful words, and only the fact of the presence of the word in the text is taken into account [4];
- Vector text model (Word2Vec, W2V), in which each term is characterized by a context-dependent vector [5],[6].

There are a significant number of classification methods, each of which shows the best efficiency depending on the characteristics of the initial set of text documents, such as selection size, text structure, text markup, selection of classification parameters (attributes), text semantics, the presence or absence of a training sample, and others.

The following three classification methods were chosen for use in the automatic classification of incoming messages: Logistic Regression, Random forest and Gradient Boosting (XG Boost library or XGB), as the most often used for solving text classification problems [4],[7],[8]. To assess the quality of classification, the following indicators were used [4]:

- precision - the ratio of the number of correctly classified requests to the total number of requests assigned by the classifier to this class;
- recall - the proportion of documents found by the classifier that belong to the class relative to all documents of this class in the sample;
- f1-score - is a harmonic mean between precision and recall.

$$f1 = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (2)$$

5. Results of processing the initial array of requests

The considered document presentation models Bag of words, Word2Vec and the classification methods under consideration Logistic Regression, Random forest, XGB belong to the class of machine learning methods with a teacher and suggest the presence of training and test data samples. Based on the processing of the training sample, model parameters are selected, classifier dictionaries are configured. The test sample is used to obtain various evaluation metrics, on the basis of which judgments are formed about the accuracy of the model and classifier. In order to make more accurate assessment of the quality of classification, the cross-validation method was used, which averages the readings of quality metrics over various partitions of the original set of documents [4]. We performed 5 folds of the initial selection into the training and test ones in the ratio of 70% to 30% and the training and classification procedure was carried out. The selection of the parameters of the methods Logistic Regression, Random forest, XGB was carried out by enumeration for each algorithm. For the implementation of text processing models, the Python programming language [8] was selected. The results obtained during the experiments are shown in table1.

Model Bag of words			
Classification method	Classification Quality Indicators		
	<i>precision</i>	<i>recall</i>	<i>f1-score</i>
Logistic Regression	0.77	0.78	0.77
Random forest	0.74	0.75	0.72
XGB	0.77	0.77	0.77
Model Word2Vec			
Classification method	Classification Quality Indicators		
	<i>precision</i>	<i>recall</i>	<i>f1-score</i>
Logistic Regression	0.76	0.78	0.76
Random forest	0.71	0.73	0.71
XGB	0.76	0.74	0.74

Table 1. Classification results for the Bag of words and Word2Vec models on the original sample

Thus, the bag of words model with the Logistic Regression classification algorithm showed the best results in the initial sample. These results on average accuracy are in the region of 75%, which is quite acceptable for this kind of tasks, however, to increase the efficiency of the created system of automatic classification, it is necessary to implement a number of measures to find the optimal parameters of the methods used.

6. Measures to improve the efficiency of the created operational subsystem of the automatic classification of incoming messages

In the framework of improving the efficiency of the created operational subsystem of automatic classification of incoming messages, a number of measures were proposed and implemented:

- Restructuring of the class system;
- Application of two-level classification technologies.

The main drawback of the original sample is that the classes are unbalanced and a number of requests from small classes did not fall into the training sample. As part of the restructuring of the class system in order to ensure its balance, special work was carried out to unify logically close classes and to remove small classes. Work on combining logically close classes included the development of questionnaires, conducting and processing the results of an expert survey. According to the results of an expert survey of 22 classes at the level of SCC, by combining classes, they were restructured into 11 classes; and 3 service classes were merged into one. The final structure of the classes and the distribution of user requests is presented in Figure 3.

Fig. 3. Distribution of user requests when combining logically close classes. Designations of classes: first level of MTCS: C0 - class "CRRR"; the second level of the MTCS (CRRR): C1 - Do the right thing, C2 - SOI, C3 - Budgeting and FEAP, C4 - SR, C5 - Advanced training and seminars, C6 - Calculation of subsidies, C7 - Slrs of MES, C8 - Performance reports, C9 - FM rating, C10 - AC of FMCS, C11 - Regulatory costs, C12 - Miscellaneous.

On the described class structure, experiments were carried out.

These experiments were similar to the ones on the original sample, the results of which are shown in table 2.

ModelBag of words			
Classification method	Classification Quality Indicators		
	<i>precision</i>	<i>recall</i>	<i>f1-score</i>
Logistic Regression	0.79	0.80	0.78
Random forest	0.79	0.80	0.79
XGB	0.81	0.80	0.80
Model Word2Vec			
Classification method	Classification Quality Indicators		
	<i>precision</i>	<i>recall</i>	<i>f1-score</i>
Logistic Regression	0.80	0.80	0.79
Random forest	0.78	0.79	0.77
XGB	0.76	0.71	0.73

Table 2. Classification results for the Bag of words and Word2Vec models on the restructured sample

The measures taken to unify logically close classes improved the classification quality. The best results for this sample were obtained for the Bag of words model and the XGB algorithm.

Another area of work on the restructuring of the class system was an experiment to remove small classes, since such classes are not important, and, on the other hand, they can distort the results of classification by other classes.

In addition, the classification results obtained on the initial selection (both for the Bag of words model and for Word2vec) show that on small classes the values of classification quality indicators are small, but with an increase in class size, they increase, which indicates insufficient training of the classifier.

As part of the experiment, an estimate of the class size was formed, giving acceptable values of quality indicators. As an example, Figure 4 shows the dependence of f1-score on the class size for the Bag of words model and the most accurate XGB classifier.

Fig. 4. Dependence of f1-score on class size

An acceptable value of the f1-score (about 80%) is achieved on classes of size 700 requests and more. Accordingly, the proposed models and algorithms when working with the original selection will show good classification results when the number of requests in each class reaches the specified value.

To assess the impact of small classes on the classification results, an experiment was conducted. During that experiment the classes with the number of queries <300 were removed from the initial selection. As a result, a sample containing 9 classes with a volume of 20,262 queries was obtained, which was also divided into a training sample of 14,183 requests and test selection of 6,079 requests.

According to the results of the experiments, a significant increase in the values of the quality indicators of classification was not obtained (the best results were shown by the XGB method on the Bag of words model (precision = 0.81, recall = 0.81, f1-score = 0.81)).

As a result, the class structure shown in Figure 3 was made in the designed system of automatic classification of incoming messages.

The next step to improve the efficiency of the created system of automatic classification of incoming messages was an experiment on the use of two-level classification technologies.

The structure of the classes obtained as a result of the work described above has the following feature - the class "CRRR" ("C0", the first level of MTCS) is large (it contains a large number of requests). This contributes to the fact that this class can "pull" others onto itself and the accuracy of classification by other classes can be reduced. To verify this assumption, an experiment was conducted to assess the quality of classification without this class.

During the experiment using the XGB algorithm on the Bag of words model, the following results were obtained:

- Classification with the class "C0" has the following values of quality indicators: precision = 0.81, recall = 0.81, f1-score = 0.81
- Classification without the "C0" class has the following values of quality indicators: precision = 0.87, recall = 0.85, f1-score = 0.86

Based on the results of the experiment, an algorithm was proposed for the functional subsystem to automatic classification of incoming requests using two-level classification technologies, and it is shown in Figure 5.

Fig 5. The algorithm of the functional subsystem for the automatic classification of incoming requests using two-level classification technologies

At the first level, the request is classified into one of the following classes:

- "C0", if it refers to the named class;
- "Ci" if it refers to one of the subject-oriented classes of SCC.

If the request is assigned to the class "C0", it is processed according to conventional algorithms. If the request is assigned to the class "Ci", the second classification stage is performed, as a result of which the incoming request is related to a specific center of competence (one of the specific classes).

The use of two-level classification technologies allows to increase the values of the main quality indicators up to 85%.

7. Conclusions

The analysis of the processes of information support of users of departmental systems operated by the Ministry of Education and Science of Russia showed that the main reason for the low efficiency of the MTCS is the extremely much time spent on the analysis and routing of incoming requests to the corresponding center of competence. Within the framework of the automation of these operations, the architecture of a functional subsystem for the automatic classification of incoming user requests is proposed, the introduction of which will reduce tenfold the time and execution cost and increasing the overall efficiency of the service as well.

The difficulty in solving the problem of automatic classification was the presence of a large number of unbalanced classes, the size of which was varied from several requests for up to ten thousand. In addition, since the source data were real user requests containing, in some cases, incorrect, confidential, or not informative data, it was necessary to preprocess the sample during the preparation of the initial data, which led to an additional reduction in the available sampling selection.

Based on the analysis of a real array of user queries and the classification system used, the use of the Bag of words feature model and the Word2Vec vector text model, as well as the Logistic Regression, Random forest and Gradient Boosting classification methods, are proposed as the main mathematical apparatus.

The experiments showed that these models are acceptable for solving the task (the main indicators of the quality of classification are in the range of 0.71 - 0.77), but the further increase in the values of the indicators of quality requires the development and implementation of a number of measures.

It is shown that as a result of applying two optimization approaches: restructuring class systems and applying two-level classification technologies, it is possible to increase the main indicators of classification quality to values of 0.85 - 0.86.

Based on the results of the experiments, the Bag of words model and the Gradient boosting classification algorithm (XGB) are proposed as the basic mathematical apparatus of the designed functional subsystem.

One of the vectors for further research is the analysis of the effectiveness of using special algorithms to solve the problem of unbalanced sampling (for example, SMOTE [11], [12]).

In addition, based on the results of the MTCS in 2019, the training sample will be significantly expanded due to new requests received from users of information systems, as well as a deeper adjustment of the parameters of the classification methods used, which will further increase the classification accuracy.

8. References

- [1] Alashkevich M., Bobryakov A., Klimenko A., Stefantsov A. (2015) Automation and Informational Support of Budgetary Institution Financing Processes, Chapter 27 in DAAAM International Scientific Book, 2015. – Pp. 319–328. B. Katalinic (Ed), Published by DAAAM International, ISBN: 978-3-902734-05-1, ISSN 1726-9687, Vienna, Austria. DOI: 10.2507/daaam.scibook.2015.27.
- [2] Alashkevich M., Bobryakov A., Klimenko A., Stefantsov A., Shcherbakov D. (2016) Organization of Complex Information Support of Budgetary Institutions' Financial and Economic Activity, Chapter 18 in DAAAM International Scientific Book 2016, pp. xxx–xxx. B. Katalinic (Ed.), Published by DAAAM International, ISBN: 978-3-902734-09-9, ISSN 1726-9687, Vienna, Austria. DOI: 10.2507/daaam.scibook.2016.18
- [3] Da Chen, Hong Wang (2018) Research on Short Text Classification Algorithm Based on Neural Network, Proceedings of 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC) 25-27 May 2018, Xi'an, pp. 14-20, IEEE, DOI: 10.1109/IMCEC.2018.8469376
- [4] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, (2008) Introduction to Information Retrieval, Cambridge University Press, ISBN: 1139472100, 9781139472104
- [5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean (2013) Distributed Representations of Words and Phrases and their Compositionality, in proceedings of NIPS, NIPS.
- [6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. (2013) Efficient Estimation of Word Representations in Vector Space, in Proceedings of Workshop at ICLR, ICLR
- [7] Batura T.V. (2017) Metodi avtomaticheskoy klassifikatsii tekstov // Software & Systems. №1(30), pp. 85-89, DOI: 10.15827/0236-235X.117.085-099
- [8] Friedman J. (1999) Greedy Function Approximation: A Gradient Boosting Machine. — IMS 1999 Reitz Lecture.
- [9] Andreas C. Müller and Sarah Guido. (2016) Introduction to Machine Learning with Python. O'Reilly Media, Inc., ISBN: 1449369898, 9781449369897.
- [10] Isacov D.N. (2017) Obzor i sravnenie besplatnikh ITSM-sistem // in Proceedings of XIX conference Nauchnoe soobshchestvo studentov: mezhdisciplinarnie issledovaniya, #8(19), 17 April 2017, Novosibirsk, pp. 62-75, ISSN 2542-0011. Available from: [https://sibac.info/archive/meghdis/8\(19\).pdf](https://sibac.info/archive/meghdis/8(19).pdf), Accessed on: 09.09.2019
- [11] Chawla, Nitesh V., et al. (2002) SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, pp. 321-357.
- [12] Fithria Siti Hanifah, Hari Wijayanto, Anang Kurnia (2015) SMOTE Bagging Algorithm for Imbalanced Data Set in Logistic Regression Analysis. Applied Mathematical Sciences, Vol. 9, No. 138, pp. 6857-6865.