Prototype of Classifier for the Decision Support System of Legal Documents

A. Alekseev[1], A. Katasev[1], A. Kirillov[2], A. Khassianov[3] and D. Zuev [3]

[1]Kazan National Research Technical University, Kazan, Russia
[2]The Arbitration Court of the Republic of Tatarstan, Kazan, Russia
[3]Higher Institute of Information Technologies and Intelligent Systems of Kazan (Volga Region) Federal University, Kazan, Russia
dzuev11@gmail.com

Abstract. We propose a prototype of the classifier of electronic documents for the decision support system in the field of economic justice. The system uses both well-known text analytics algorithms and an original algorithm based on an artificial neural network. A text mining model has been developed to classify court documents to determine the category (class) of a statement of claim. A preliminary analysis of court documents and the selection of significant features were carried out. To choose the best way of solving problem of document classification we implemented Bayesian classification algorithm, k nearest neighbor algorithm and decision trees algorithm. All used algorithms show results with errors on the same sample corpus of texts. To improve the accuracy of classification, an original model based on an artificial neural network was developed, which shows an unmistakable determination of the type of document on a test sample for a number of classes of lawsuits in arbitration proceedings.

Keywords: Classification, Text Mining, Artificial Neural Network, Classification Algorithms, Decision Support System.

1 Introduction

The judicial system is an area where the amount of work with text-based documents is huge, and the decision-making process should always be clear and transparent. Therefore, especially in the face of growing workload for employees working in this area, automated intelligent tools for data analysis of the input information are required. Currently, the courts of the Russian Federation start implementation of electronic workflow in the field of legal proceedings based at e-documents instead of traditional ones [1]. Automated text analysis allow to outline important features of documents (jurisdiction, nature of the dispute, parties involved, etc.), search the judicial database and find similar documents for which decisions have already been made. Our research focused at the following aspect of the work of the judicial system: to reduce the burden on judges and reduce the time for considering economic disputes, we propose a model for classifying judicial documents based on Text Mining [2] that solves the

problem of determining the type of arbitration dispute. To determine the type class (category) of the judicial dispute, the following tasks were solved:

1. Creation of a text mining model for the classification of arbitration documents;
2. Modeling of the processing and classification of such documents in the Rapid Miner Studio software;
3. Programming of processing and classification modules in the R language;
4. Selection of the most effective classification algorithm for further testing in the arbitration proceedings.

Existing software solutions that used today in the legal field, usually, focused at automating workflow as a whole or implement databases of legal documents with a meager set of search tools. The semantic technologies and text analysis tools are used rarely. In this situation, improvement the quality and the efficiency of judges only possible with implementation of number automation tools into current workflow or with a significant increase amount of staff needed for manual operations.

Currently experts of the Stanford Center for Legal Informatics, the Chicago College of Law in Kent and the College of Law in South Texas created an intelligent system [3] based on machine learning and data analysis that predicts court verdicts with an accuracy of more than 70%. This system uses the records of the database of the US Supreme Court from 1816 to 2015 as input.

Another example of system similar to the topic of our study is the Case Cruncher Alpha system [4], developed at Sidney Sussex College, Cambridge, and focused on forecasting of the solution of legal problems in banks, insurance companies, and legal advice. Its main problem (as well as many other foreign systems) is the lack of support for the Russian language and Cyrillic transcription.

2 Legal Documents and Text Mining methods

The proposed classifier is one of the modules of the decision support system in the field of justice known as the intellectual information system (IS) “Robot Lawyer”. This information system allows participants of the legal process to prepare cases more effectively and plan judicial activities. The system is focused on arbitration courts. Goals and objectives, the general architecture of the Robot Lawyer system, the proposed modules and approaches to its development were presented in [5, 6]. In general, this system, based on artificial neural networks, is schematically depicted at Figure 1. On the scheme module 1 is a pre-processing module of text documents, module 2 is a module for determining the main class of a judicial document, module 3 is a module stands for determining the subclass of a judicial document, and module 4 – the decision making module. Our paper describes the first two of modules.

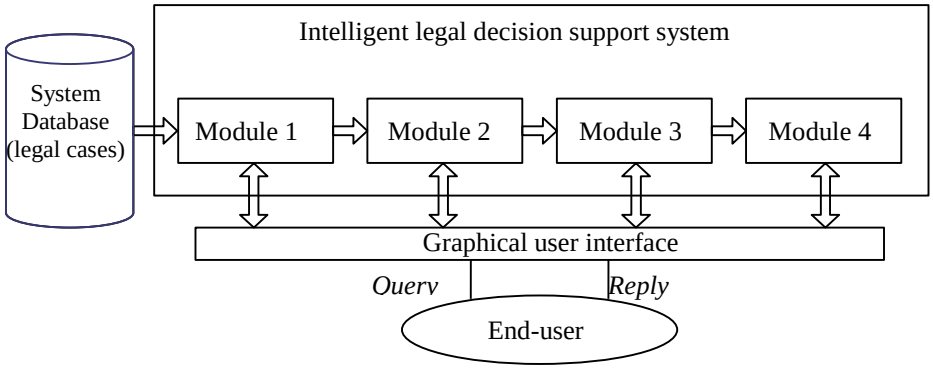

Fig. 1. Scheme of the proposed judicial decision support system

It is known that the analysis of text documents is performed in five steps [7]:

1. Information retrieval – discovering of the documents that are crucial for further processing and analysis. Usually users form the corpora of texts for analysis by themselves. With a growing number of documents manual selection becomes time-consuming, so we must use automated selection procedures.
2. Pre-processing of documents – selected documents are converted into a machine-readable format to apply formal algorithms of machine learning. Pre-processing usually is used to remove terms that don't affect text semantics, punctuation marks and to convert text into a normalized form. The methods of preprocessing used in our project are discussed below.
3. Information extraction – at this stage we extract all important information (features, key phrases) from the documents.
4. Machine learning – this is the main step in the analysis at which new knowledge is formed and patterns hidden are revealed.
5. Interpretation of the results - presentation of the analysis results in natural language in a user-friendly form.

2.1 **Preprocessing of documents**

Usually we use combination of several methods. One of them is the tokenization of the text, that is, the operation of breaking a document into separate words. As a result, an array of tokens is formed for further processing [8]. The next step is to convert all characters to upper or lower case. For example, all words "text", "Text", "TEXT" are reduced to lowercase "text" [9]. Then it is necessary to filter stop words that do not carry a significant informational meaning: conjunctions, prepositions, articles, interjections, particles, etc. The list of stop words must be compiled in advance. It depends on the text language and subject of the document being processed.

The next step is stemming [10], during which all words must be converted to the normal form. Main operations on this step are identifying cognate words, cutting off suffixes and endings, reducing the terms to the singular, nominative case of a noun,

adjective or indefinite form of the verb. The main problem in such transformations is a possible violation of the semantics of sentences and phrases, therefore, it is necessary to take into account the original language. Currently, there are a number of well-known implementations of the stemming and lemmatization algorithms for the Russian language in the form of plug-in software libraries – Snowball, Porter or MyStem [11].

We used the Snowball algorithm as a stemming library for modeling in RapidMiner Studio, and the library MyStem for software module written in R.

For further actions, it is necessary to present the text in a form convenient for analysis. We used the Document-Term Matrix (DTM) matrix, which is a table where each row corresponds to the document and the column to the terms found in the document body [12]. At the intersection of rows and columns, the values of the term weights are stored in the document.

During the creation of the prototype of the classifier, we analyzed court documents in four categories: contesting decisions of the antimonopoly authorities ("ANTIMONOPOLY"), contesting the actions of bailiffs ("BAILIFFS"), prosecuting for violation of licensing terms ("LICENSES"), disputes on non-fulfillment or improper fulfillment of obligations under supply contracts ("DELIVERY").

After the pre-processing stage of the documents, a DTM matrix of dimension 167 * 5419 was formed, where 167 is the total number of documents (38 for the class "BAILIFFS", 32 for the class "ANTIMONOPOLY", 61 for the class "DELIVERY", 36 for the class "LICENSES"). Total numbers of terms contained in all documents are 5419. For the matrix values we chose TF measure. After the initial filling the DTM matrix, we noticed that not all terms are valuable for determining the class of a document. So, before further actions, it is necessary to extract informative features from text processed.

2.2 Information extraction

As a rule, all methods for classifying of texts are based on the assumption that documents belonging to the same category (class) contain the same characteristics (words or phrases), and the presence or absence of such signs in the document indicates its belonging to certain class (see, e.g., [7]). To determine the group of features (terms) that characterize the categories of processed documents, we tested the entropy method of information growth (Information Gain), Chi Square and Gini index methods on our legal text corpora [14, 16].

Table 1. Features extraction results

Information Gain		Gini Index		Chi Square	
Term	Coef.	Term	Coef.	Term	Coef.
истц	1.0	Истц	1.0	судебн	1.000
истец	0.899	Истец	0.925	заявител	0.935
исполнител	0.898	обязательств	0.923	ответствен	0.921

пристав	0.898	заявител	0.878	Рф	0.890
обязательств	0.888	Ответчик	0.874	ответчик	0.861
заявител	0.855	Накладн	0.873	исполнител	0.806
договор	0.846	Договор	0.849	пристав	0.806
незакон	0.840	исполнител	0.813	Са	0.801
накладн	0.835	Пристав	0.813	Коап	0.783
ответчик	0.831	Приста	0.758	исполнительн	0.766
взыскател	0.799	Взыскател	0.754	Привлека	0.759
исполнительн	0.799	исполнительн	0.736	производств	0.749
приста	0.793	Коап	0.731	Истц	0.744
взыскан	0.790	Иск	0.729	предусмотрен	0.725
антимонопольн	0.784	производств	0.721	Взыскан	0.724

As a result we found that all methods used revealed almost identical terms, slightly differing in their coefficients.

For the training and test samples, we used 40 terms obtained at the stage of extraction of informative terms, thus, the final document-term matrix received a dimension of 167 * 40. At the production stage number of terms will be increased, however, at the stage of developing of a prototype system, the selected limited set of terms is enough. Obviously, the number of terms used affects the dimension of the matrix and affects the time required to train the classifier. In order to create all samples (test and training ones) we divide the resulting document-term matrix into 117 * 40 and 50 * 40 matrices (70/30 ratio). The matrix values were used as input for classification algorithms.

3 Classification of Electronic Documents

The problem of the classification [15] is known as follows. Assume that we have a set of text documents $D = \{X_1, ..., X_n\}$ and a set of k different discrete values $\{1, ..., k\}$, each of which corresponds to a label of a class (category). To solve the problem it is necessary to determine category (corresponding to the label value) for each document X_i.

Usually such problems solved with the help of machine learning algorithms with a teacher, where a training set of documents (i.e., documents with well-known category labels) is used to build a classification model that determines the relationship of features in a particular document with one of the class labels. For elements of a test sample where the class of the documents is unknown, the developed and trained model should determine the class label. To clarify the algorithms of the classifier, the model should be retrained periodically.

We tested several classification algorithms to determine the method that is most optimal on the available data sample. The following classification methods were tested: the naive Bayes classifier [7], the method of k-nearest neighbors [8] and decision

trees method [15]. The results of usage of these classifiers on the test sample are listed in Tables 2–4 (classification matrices).

Table 2. Results of usage a naïve Baes classifier

	True BAILIFFS	True ANTIMONOPOLY	True DELIVERY	True LICENCES	Class precision
pred.BAILIFFS	11	0	0	0	100%
pred. ANTIMONOPOLY	0	10	0	2	83.33%
pred. DELIVERY	0	0	18	0	100%
pred. LICENCES	0	0	0	9	100%
Class recall	100%	100%	100%	81.82%	

Table 3. Results of usage the k-nearest neighbors algorithm

	True BAILIFFS	True ANTIMONOPOLY	True DELIVERY	True LICENCES	Class precision
pred.BAILIFFS	11	0	0	0	100%
pred. ANTIMONOPOLY	0	10	0	1	90.91%
pred. DELIVERY	0	0	18	0	100%
pred. LICENCES	0	0	0	10	100%
Class recall	100%	100%	100%	90.91%	

Table 4. Results of usage Decision trees algorithm

	True BAILIFFS	True ANTIMONOPOLY	True DELIVERY	True LICENCES	Class precision
pred.BAILIFFS	11	0	0	0	100%
pred. ANTIMONOPOLY	0	10	0	2	83.33%
pred. DELIVERY	0	0	18	1	94.74%
pred. LICENCES	0	0	0	10	100%
Class recall	100%	100%	100%	90.91%	

None of the applied algorithms showed 100% classification accuracy. Here we can see classification errors and accuracy obtained. Bayes classifier – 4%, classification accuracy 96%; kNN –2%, classification accuracy 98%; decision trees – 2%, classification accuracy 98%. Since the data sample is very small, the results cannot be considered as satisfactory.

To improve accuracy of the classification of legal documents, we developed a model based on artificial neural networks (ANNs) [16]. Proposed neural network has the following parameters:

1. 40 neurons in the input layer, 1 hidden layer with 4 neurons, 4 output neurons;
2. activation function: sigmoid.

As a result the classifier based on the neural network showed 100% classification accuracy on the test sample for classification according to four criteria (see Table 5).

Table 5. Results of usage ANN algorithm

	True BAILIFFS	True ANTIMONOPOLY	True DELIVERY	True LICENCES	Class precision
pred.BAILIFFS	11	0	0	0	100%
pred. ANTIMONOPOLY	0	10	0	0	100%
pred. DELIVERY	0	0	18	0	100%
pred. LICENCES	0	0	0	11	100%
Class recall	100%	100%	100%	100%	

4 Conclusion

We applied and tested known methods of the text mining algorithms in a separate subject area – arbitration proceedings. A neural network model for classifying text documents into standard categories is proposed. To solve the classification problem, a preliminary analysis of court documents and the extraction of informative features for certain categories of litigation were done. We applied Bayes classification, k nearest neighbor and decision trees algorithms on our corpus. To increase the absolute accuracy of the classification, a model based on an artificial neural network in a test sample was suggested. Proposed model showed 100% classification accuracy on test sample. For preprocessing procedure we developed a software package in the R language. All tests were done at the legal documents corpus of the Arbitration Court of the Republic of Tatarstan.

At the next step, we plan to increase the amount of the documents, to consider a larger number of types of possible classes, and to develop software modules that implementing the steps of selecting informative features and classification. After performing all tests, the module will be included as a service in the "Robot-Lawyer" system [5, 6].

This work was funded by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities, grant agreement no. 1.2368.2017) and with partial financial support of the Russian Foundation for Basic Research and the Government of the Republic of Tatarstan, within the framework of scientific project No. 18-47-160012.

References

1. Postanovleniye Plenuma Vysshego Arbitrazhnogo Suda RF ot 25 dekabrya 2013 g. No. 100 "Ob utverzhdenii Instruktsii po deloproizvodstvu v arbitrazhnykh sudakh Rossiyskoy Federatsii (pervoy, apellyatsionnoy i kassatsionnoy instantsiy)" (2013).

2. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery: an Overview, Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, (1996).
3. Katz, D.M., Bommarito, M.J. II, Blackman, J.: A general approach for predicting the behavior of the Supreme Court of the United States, PLoS ONE.
4. Case Crunch Alfa. http://www.case-crunch.com.
5. Zuev, D.S., Marchenko, A.A., Khassianov, A.F.: Text mining tools in legal documents. In: CEUR Workshop Proceedings, pp. 214–218. (2017).
6. Alekseev, A.A., Zuev, D.S., Katasev, A.S., Tutubalina, E.V., Khassianov, A.F.: Intellectual information decision support system in the field of economic justice. In: Nauchnyy servis v seti Internet: trudy XIX Vserossiyskoy nauchnoy konferentsii (17–22 sentyabrya 2018 g., g. Novorossiysk), Moscow, Keldysh Institute of Applied Mathematics (2018).
7. Barsegyan, A.A., Yelizarov, S.I., Kupriyanov, M.S., Kholod, I.I., Tess, M.D.: Analiz dannykh i protsessov. 3 edn., BKHV-Peterburg, S.-Petersburg (2009).
8. Aggarwal, C.C.: Machine Learning for Text. Springer International Publishing AG, part of Springer Nature (2018).
9. Hofmann, M., Klinkenberg, R.: RapidMiner: Data Mining Use Cases and Business Analytics Applications. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, CRC Press (2016).
10. Manning, C.D., Raghavan, P., Schutze, H. S.: Introduction to Information Retrieval. Cambridge University Press, Cambridge, England (2008).
11. Segalovich, I.: A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In: Proceedings of the International Conference on Machine Learning, Las Vegas, Nevada, USA (2003).
12. Williams, G.: Hands-On Data Science with R, Text Mining, (2014).
13. Feinerer, I., Hornik, K., Meyer, D.: Text Mining Infrastructure in R.: Journal of Statistical Software. 25 (5), pp. 54–64 (2008).
14. Kotu, V., Deshpande, B.: Predictive Analytics and Data Mining. Concepts and Practice with RapidMiner. Morgan Kaufmann, 225 Wyman Street, Waltham, MA 02451, USA (2014).
15. Aggarwal, C.C.: Data Classification: Algorithms and Applications. Chapman& Hall/CRC Data Mining and Knowledge Discovery Series, CRC Press (2015).
16. Glova, V.I., Anikin, I.V., Katasov, A.S., Krivilov, M.A., Nasyrov, R.I.: Myagkiye vychisleniya, Uchebnoye posobiye. Izd-vo Kazan. gos. tekhn. un-ta, Kazan, Russia (2010).