УДК 004.588

Glazkova A.V.

Tyumen State University, Tyumen, Russia

AUTOMATIC DOCUMENT CLASSIFICATION ON THE BASIS OF TEXT AUDIENCE AGE GROUPS IN E-LEARNING SYSTEMS

ABSTRACT

The paper discusses the feasibility of automatic document classification mechanisms for e-learning systems. We suggest an intellectual system for text classification based on the age groups of text audience and represent the results of computational experiment characterizing the performance of the method.

KEYWORDS

Informational retrieval; document classification; natural language processing; e-learning systems.

Глазкова А.В.

Тюменский государственный университет, г. Тюмень, Россия

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ ДОКУМЕНТВ НА ОСНОВАНИИ ВОЗРАСТНЫХ ГРУПП ИХ АДРЕСАТОВ В СИСТЕМАХ ЭЛЕКТРОННОГО ОБУЧЕНИЯ

АННОТАЦИЯ

В статье обсуждаются возможности реализации механизмов автоматической классификации документов в рамках создания систем электронного обучения. Предлагается интеллектуальная система классификации текстов на основании возрастных групп их потенциальных адресатов. Приводятся результаты вычислительного эксперимента, характеризующего результативность предложенного метода.

КЛЮЧЕВЫЕ СЛОВА

Информационный поиск; классификация документов; обработка естественного языка; программирование; системы электронного обучения.

E-learning systems are rapidly changing the way that educational institutes prefer for training. The popularity of e-learning can be explained by the fact that there are a number of indisputable advantages compared with traditional learning.

E-learning is one of the most modern educational tools and a promising type of training. Students studying remotely can decide when and how much time during the semester they can allocate to studying. They give for themselves their individual training schedules. Students do not have to worry about the fact that they lag behind their classmates. They can always return to the study of more complex issues, several times watch video lectures, read the correspondence with the teacher. At the same time the already well-known topics can be skipped. Besides, students can learn from the comfort of their home or office from anywhere in the world. For starting training they only need to have a computer with Internet access. Students have no need to go to school or university every day. It is a definite advantage for people with disabilities, for those who live in remote areas, who are serving time in prison, for parents with young children. Distance learning does not require a release on the basic work and business trips.

In traditional forms of teaching it is quite difficult for tutor to give the necessary attention to the number of all students and adjust to everyone's pace of work. Using of distance technologies helps to organize individual approaches. In addition, sometimes consultations with the tutor via e-mail more efficiently and faster than the appointment of a personal meeting with the full-time or distance learning.

Studies show [1-3] that distance learning results do not concede to traditional forms of education. A student during his distance learning works with educational material by his own, that improves memorization and understanding. The opportunity to immediately apply the knowledge into practice at work helps to fix it. Futhermore, the modern technologies make education more interesting and lively.

In recent years, the acquisition and distribution of educational resources has been largely automated. In particular, e-learning developing was considered by Doneva R. and Gaftandzhieva S [4], Bodrow W. [5], Bazhenova I.Yu. [6], Parra B.J. [7], Shivdas P.A., Sivakumar S. [8].

One of the main difficulties of distance learning is the need of creating a volume electronic document library for storing a large number of educational texts [2, 9]. Documents in these libraries are oriented to students of various specializations, training level and age. Document classification according to these parameters requires a lot of human resources; therefore the solution of problems of automatic document classification is undoubtedly important for science and practice.

Various issues of automatic text classification are repeatedly discussed in scientific papers (in particular, the recent works of Onan A. et al. [10], Zhitomirsky-Geffet M. et al. [11], Le M.H. [12]). Fedotov A.M. et al. [13] proposed a technological approach for developing a model of information system to support the scientific and educational activities, organized in the form of a digital library. The problems of systematization of library documents are reviewed by Malki Z.[14] and Talla A. [15].

This study deals with the task of document classification on the basis of the age of their audience. This problem was affected by Akker R. and Traum D. [16], Choi D. [17], Lee H. [18]. Their studies were carried out for English corpuses. Using the same classification features for Russian documents is not correct due to the individual grammatical and lexical characteristics of Slavic languages. The ability to provide document classification on the basis of the age of their audience improves the relevance of the results of informational retrieval in electronic libraries and allows the system to eliminate unwanted resources from the query results.

In this research we suggest a text classification method and give a list of classification features that we have used in our computational experiment.

The "ISACT'

We propose the intellectual system "ISACT" for text classification. The main task of the system is to ensure the automatic classification of texts based on the age of their audience destinations. The "ISACT" can be used both for determining the age category of the destination of the text and in order to conduct comprehensive research on large samples of texts on various subjects.

The "ISACT" consists of three modules (fig.1):

- the module of semantic and syntactic analysis;
- the classification module;
- the storage module.

The module of semantic and syntactic analysis parses texts and searches for syntactic and semantic text features.

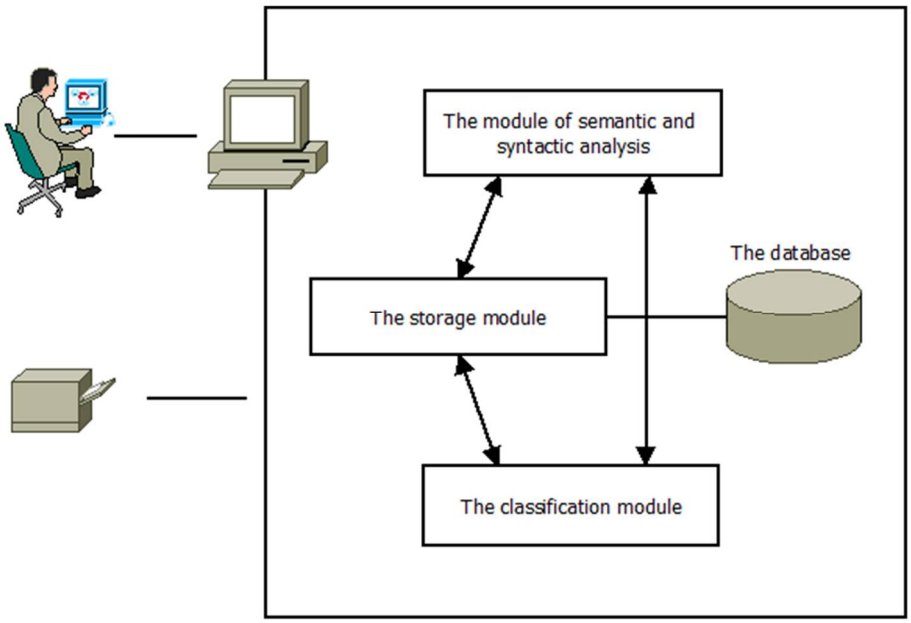

Fig.1. The "ISACT" architecture

The parser is present a text as a set of proposals and divide it into a plurality of pairs of tokens and corresponding frequencies.

Texts loading the system can be represented as in the markup or in the form of unallocated. The module of semantic and syntactic analysis creates a description of the processed text in XML-format for saving and uploading the results of semantic and syntactic text analysis, which allows us to present the

results of the analysis in a form understandable to both system and user. This description is used for the subsequent analysis of the data extracted. The results obtained in the module semantic-syntactic analysis, and discharged from the system may be re-added to the other modules to continue working with the text.

The classification module distributes the text by category. Correlation of the text with the categories is based on the results of the module of semantic and syntactic analysis working.

The module allows us to work with texts in two modes: the training mode and the control mode. Thus, depending on the software system parameters set by the user it is possible to use the system as a configuration of the classifier and to analyze and determine the category of text.

The classification module classifies texts in two ways: the method is based on the computation of Mahalanobis distance as a measure of the proximity of texts and neural method.

The classical form of Mahalanobis distance's formula is:

$$\rho\left(F_{T_i}, F_{T_j}\right) = \sqrt{(F_{T_i} - F_{T_j})^T \Lambda^T C^{-1} (F_{T_i} - F_{T_j})}\ , \qquad (1)$$

Λ – the matrix of weighting coefficients; C – the covariance matrix; F_{T_i}, F_{T_j} – sets of classification features for texts T_i and T_j.

Similarly, the formula (1) can be calculated as the distance between the text and the center of mass of the category R, represented as a vector of weighted values of classification features:

$$\rho\left(F_{T_i}, R\right) = \sqrt{(F_{T_i} - R)^T \Lambda^T C^{-1} (F_{T_i} - R)}\ ,$$

$$R = \frac{\sum_{j=1}^{M} k_j F_{T_j}}{M},\ k_j > 0,\ \sum_{j=1}^{L} k_j = 1,$$

R – the vector describing the location of the center of mass of categories; M – the number of text of the category included in the training sample, $1 \le M \le L$; L – the total number of texts; k_j – the weight ratio of confidence training sample text.

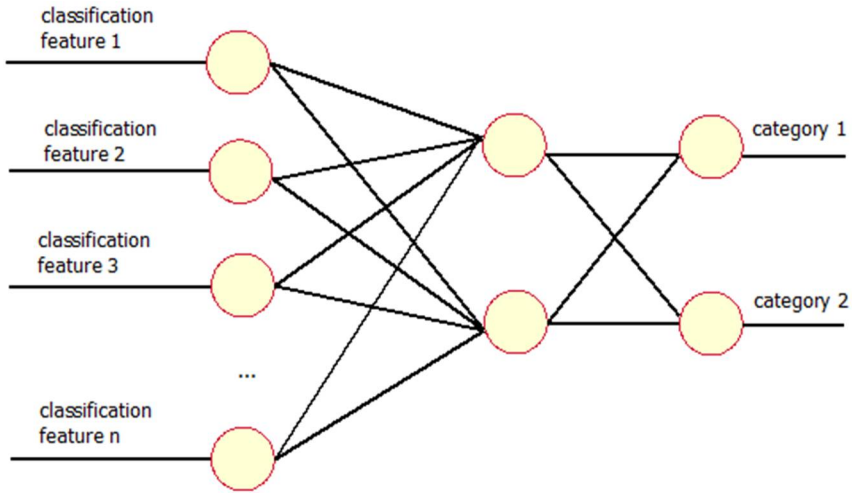

Fig.2. The structure of multi-layer perceptron

Neural network method implemented using a multi-layer perceptron (fig.2). Using of this network type is caused by its ability to solve problems of poorly formalized types based on existing examples and identifying patterns in the communication of input and output data. The input layer of the neural network comprises a number of neurons equal to the number of classifications, and the output layer contains the number of neurons corresponding to the number of categories.

The text is input to the classifier in the form of a set of classification features identified in the course of semantic and syntactic analysis module. The neural network implemented in the module has one hidden layer.

The network training was conducted on the basis of the back-propagation algorithm.

The storage module is intended for the organization of storing, editing and retrieval of texts and categories in relational database tables. In addition, the module has a mechanism of advanced text search, which allows us to limit the results of a search query filters (by author, publication, category), register of symbols and selected columns.

Features modular architecture allows flexible integration of the "ISACT" in electronic documents systems for solving problems of analysis, classification and storage of texts. For example, for integration of

the storage module to a relational database of the electronic library it is needed to change the connection string that is used to open the database and select the required record from the data source. The “ISACT” includes the following text data exchange formats: .xls, .xlsx, .xml, .txt.

The approbation and results

For the selection of classification features we used two samples of texts that are available on the Russian National Corpus website [19]. The first text sample is represented by literary texts of different genres (historical fiction, adventure, documentary prose, etc. – 5 902 documents, 9 332 659 sentences, 94 538 056 words), the second one includes children's literature (632 documents, 547 735 sentences, 4 742 627 words).

The corpus includes various types of texts representing modern standard (written) Russian. The age group of potential readers of texts - adult or child - is determined on the basis of expert evaluation.

In the computational experiments are used the Database “Morphological Standard of the Russian National Corpus” and “Database of meta tagging of the Russian National Corpus" (a collection of children's literature)”[20]. The sample size is 532 texts of modern fiction (from the middle of the 20 century) and 510 various texts of children’s literature.

This study deals with two categories – children’s and adults, according to the corpus provided for the experiments.

The classification result is the percentage of correctly classified records on the control sample The original texts of the sample was divided into a training and a test sample n times. Next, we calculated the average values for all partitions. The accuracy of the classification for the method based on Mahalanobis distance amounted to 74.16% (standard deviation - 5.88%), for the neural method - 72.07% (standard deviation - 6.62%).

Conclusion

The results of the study show the possibility of using automatic text classification for text age audience detection tasks. In particular, this type of text classification will be useful in e-learning systems and in digital libraries. In prospect, we plan to conduct experiments for more age categories of users. This requires obtaining a training sample with the expert division of texts into the required number of categories.

The software package “ISACT” implemented in educational institution "Ugra Training Center" as the subsystem of e-learning system.

Литература

1. Ni A.Y. Comparing the effectiveness of classroom and online learning: teaching research methods// Journal of Public Affairs Education. – 2014. – №2. – P. 199-215.
2. Nguyen T. The effectiveness of online learning: beyond no significant difference and future horizons // MERLOT Journal of Online Learning and Teaching. – 2015. – №2. – P. 309-319.
3. Arkorful V., Abaidoo N. The role of e-learning, the advantages and disadvantages of its adoption in higher education // International Journal of Education and Research. – 2014. – №12. – P. 397-410.
4. Doneva R, Gaftandzhieva S. Automated e-learning quality evaluation // Proceedings of International Conference on e-Learning. – Bratislava, Slovakia, 2014. – P. 169-174.
5. Bodrow W. Evaluation of enterprise skills from the perspective of university education // Proceedings of International Conference on e-Learning. – Bratislava, Slovakia, 2015. – P. 130-135.
6. Баженова И.Ю. Современные подходы к формированию профессиональных компетенций в области применения языков программирования // Современные информационные технологии и ИТ-образование. – 2015. – №11. – С. 130-134.
7. Parra B.J. Learning strategies and styles as a basis for building personal learning environments // International Journal of Educational Technology in Higher Education. – 2016. – №11. – P. 70-77.
8. Shivdas P.A., Sivakumar S. Innovation in services: A lancastrian approach to the field of e-learning // Education and Information Technologies. – 2016. – №6. – P. 1913–1925.
9. Sharifabadi S.R. How digital libraries can support e-learning // The Electronic Library. – 2006. – №3. – P. 389-401.
10. Onan A., Korukoğlu S., Bulut H. Ensemble of keyword extraction methods and classifiers in text classification // Expert Systems with Applications. – 2016. – Vol. 57. – P. 232-247.
11. Zhitomirsky-Geffet M., David E., Koppel M., Uzan H. Utilizing overtly political texts for fully automatic evaluation of political leaning of online news websites // Online Information Review. – 2016. – №3. – P. 362-379.
12. Ле М.Х. Оптимизация алгоритма knn для классификации текстов // Труды Московского физико-технического института. – 2016. – № 1. – С. 92-94.
13. Fedotov A.M., Barakhnin V.B., Fedotova O.A., Zhizhimov O.L. A model of digital library to support research activities // Современные информационные технологии для фундаментальных научных исследований в области наук о земле: материалы Международной конференции. – Петропавловск-Камчатский, 2014. – С. 22.
14. Malki Z. Information and communication technologies role in developing electronic libraries and information centers // Journal of Theoretical and Applied Information Technology. – 2015. – № 2. – P. 167-183.
15. Tella A. Electronic and paper based data collection methods in library and information science research: A comparative analyses // New Library World. – 2015. – №9-10. – P. 362-379.
16. Akker R., Traum D. A comparison of addressee detection methods for multiparty conversations // Proc. of methods for multiparty conversations. – Amsterdam, 2009. – P. 99-106.
17. Choi D., Ko B., Kim H., Kim P. Text analysis for detecting terrorism-related articles on the Web // Journal of Network and Computer Applications. – 2013. –№5. – P. 37-46.

18. Lee H., Stolcke A., Shriberg E. Using out-of-domain data for lexical addressee detection in human-human-computer dialog // Proc. North American ACL/Human Language Technology Conference. – Atlanta, 2013. – P. 215-219.
19. Национальный корпус русского языка [Электронный ресурс]. 2015. URL: http:// ruscorpora.ru/ (дата обращения: 05.10.2016).
20. «База данных метатекстовой разметки Национального корпуса русского языка» (коллекция детской литературы)». 2014.

References

1. Ni A.Y. Comparing the effectiveness of classroom and online learning: teaching research methods// Journal of Public Affairs Education. – 2014. – №2. – P. 199-215.
2. Nguyen T. The effectiveness of online learning: beyond no significant difference and future horizons // MERLOT Journal of Online Learning and Teaching. – 2015. – №2. – P. 309-319.
3. Arkorful V., Abaidoo N. The role of e-learning, the advantages and disadvantages of its adoption in higher education // International Journal of Education and Research. – 2014. – №12. – P. 397-410.
4. Doneva R, Gaftandzhieva S. Automated e-learning quality evaluation // Proceedings of International Conference on e-Learning. – Bratislava, Slovakia, 2014. – P. 169-174.
5. Bodrow W. Evaluation of enterprise skills from the perspective of university education // Proceedings of International Conference on e-Learning. – Bratislava, Slovakia, 2015. – P. 130-135.
6. Bazhenova I.Yu. Sovremennye podkhody k formirovaniyu professional'nykh kompetentsiy v oblasti primeneniya yazykov programmirovaniya // Sovremennye informatsionnye tekhnologii i IT-obrazovanie. – 2015. – №11. – P. 130-134.
7. Parra B.J. Learning strategies and styles as a basis for building personal learning environments // International Journal of Educational Technology in Higher Education. – 2016. – №11. – P. 70-77.
8. Shivdas P.A., Sivakumar S. Innovation in services: A lancastrian approach to the field of e-learning // Education and Information Technologies. – 2016. – №6. – P. 1913–1925.
9. Sharifabadi S.R. How digital libraries can support e-learning // The Electronic Library. – 2006. – №3. – P. 389-401.
10. Onan A., Korukoğlu S., Bulut H. Ensemble of keyword extraction methods and classifiers in text classification // Expert Systems with Applications. – 2016. – Vol. 57. – P. 232-247.
11. Zhitomirsky-Geffet M., David E., Koppel M., Uzan H. Utilizing overtly political texts for fully automatic evaluation of political leaning of online news websites // Online Information Review. – 2016. – №3. – P. 362-379.
12. Le M.Kh. Optimizatsiya algoritma knn dlya klassifikatsii tekstov // Trudy Moskovskogo fiziko-tekhnicheskogo instituta. – 2016. – № 1. – P. 92-94.
13. Fedotov A.M., Barakhnin V.B., Fedotova O.A., Zhizhimov O.L. A model of digital library to support research activities // Sovremennye informatsionnye tekhnologii dlya fundamental'nykh nauchnykh issledovaniy v oblasti nauk o zemle: materialy Mezhdunarodnoy konferentsii. – Petropavlovsk-Kamchatskiy, 2014. – P. 22.
14. Malki Z. Information and communication technologies role in developing electronic libraries and information centers // Journal of Theoretical and Applied Information Technology. – 2015. – № 2. – P. 167-183.
15. Tella A. Electronic and paper based data collection methods in library and information science research: A comparative analyses // New Library World. – 2015. – №9-10. – P. 362-379.
16. Akker R., Traum D. A comparison of addressee detection methods for multiparty conversations // Proc. of methods for multiparty conversations. – Amsterdam, 2009. – P. 99-106.
17. Choi D., Ko B., Kim H., Kim P. Text analysis for detecting terrorism-related articles on the Web // Journal of Network and Computer Applications. – 2013. –№5. – P. 37-46.
18. Lee H., Stolcke A., Shriberg E. Using out-of-domain data for lexical addressee detection in human-human-computer dialog // Proc. North American ACL/Human Language Technology Conference. – Atlanta, 2013. – P. 215-219.
19. The Russian National Corpus [Electronic resource]. 2015. URL: http:// ruscorpora.ru/ (date of access: 05.10.2016).
20. «Baza dannykh metatekstovoy razmetki Natsional'nogo korpusa russkogo yazyka» (kollektsiya detskoy literatury)». 2014.

Поступила 07.10.2016

Об авторах:

Глазкова Анна Валерьевна, ассистент кафедры программного обеспечения Тюменского государственного университета, anna_glazkova@yahoo.com.