Conceptual Scheme for Text Classification System

Nicolay Lyfenko

Russian State University for the Humanities, Moscow
lyfenkoNick@yandex.ru

Abstract. The paper describes an application of classification algorithms to the text categorization problem. Author proposes a conceptual scheme for an automatic text categorization system. This system must operate with various text representation models and data mining methods. The novelty of this system consists in advanced implementation of JSM method for automatic hypothesis generation — an original logical-combinatorial technology of data mining, which is developed in Russia by several research groups.

Keywords: text classification system, machine learning, data mining, natural language processing

1 Introduction

Due to an increasing number of text documents in digital form and the extension of a data stream in different fields of professional activities the interest in a text categorization task has essentially increased. The main goal of classifying a new text is to assign a predefined class or classes to it [1]. It is being solved with the help of the text classification system ADC (*automatic document classifier*). Our system includes: different text representation models, a number of text mining methods and some text similarity metrics.

The main goal of the system is to compare various classical text classification methods to JSM method for automatic hypothesis generation and choose the best one for a particular task [2, 3].

This research is in progress so the main purpose of this work is to build a conceptual scheme for the ADC system, develop a project scheme for ADC system and represent its current state of work.

There is a great variety of machine learning methods to make a text classification. The most popular of are: *k-nearest neighbor, Rocchio classifier, neural network, decision trees, naive Bayes classifier,* and *support vector machine* [4–6]. There are not only algorithms but ready to use frameworks and IDE's for text classification problem (e.g. Rapidminer[1], Gate[2]). But none of them has the JSM method implemented.

This method was proposed by V.K. Finn at the beginning of the 1980s. The abbreviation JSM is given in honor to John Stuart Mill. The JSM method uses the Mill's idea

[1] http://rapidminer.com/products/rapidminer-studio/
[2] https://gate.ac.uk/

that common effects are more likely to have common causes. The JSM method for automatic hypothesis generation is known as an original set of logical combinatorial technologies for data mining using rules of plausible reasoning [7].

The JSM method includes three cognitive procedures: *induction, analogy, abduction* [2] and two main stages: *learning* (to identify data patterns using Mill's agreement) and *prediction*. By means of *induction* the JSM method generates casual hypotheses. With the help of *analogy* additional definition to unknown examples is formed (prediction). The *abduction* procedure evaluates the plausibility of the generated hypothesis.

This logical-combinatorial method for intelligent data analysis has shown good results on level with *SVM* method in the work [8] for the task of sentiment analysis. So we have a proposal to apply it in the task of automatic topic and authorship classification.

2 Conceptual Scheme for ADC System

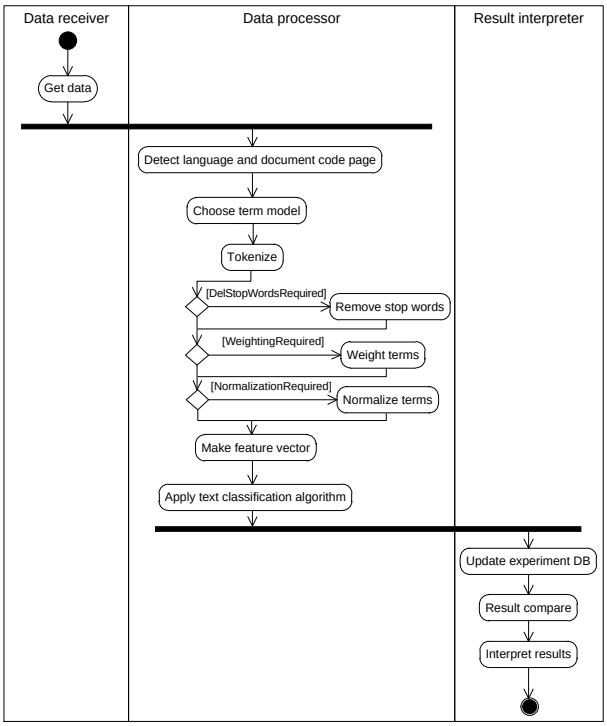

Fig. 1. Conceptual scheme for ADC system

Fig. 1 shows the key steps for automatic document classification used in the ADC system: to get data, to process it and to analyze results.

Reasoning from the fact that a document to analyze can be written in different code pages and various languages (Russian and English currently supported) a character set

and a text language should be identified. We are using statistical analysis as in [9]. In our research we normalize terms with the help of a made inverse dictionary based on Zaliznak's for the Russian language[3]. English words are stemmed.

We use some classical IR text models: *frequent* model, *tf-idf* model for text representation as an *n*-dimensional vector (*vector space model*) and not so popular but promising ones are investigated: *LOWBOW* (Locally Weighted Bag of Words Framework) [9], *MFS* (Maximal Frequent Sequences) [6], *Document Occurrence Representation* (DOR) & *Term Co-occurrence Representation* (TCOR) [9].

2.1 Project Object Model

In order to choose the best technic for a certain text classification approach we have to compare all the methods and have a log of our experiments. That is why it is proper to have well-structured and a user-friendly GUI for an experiment and logically organized project scheme for ADC system and data base for experiments.

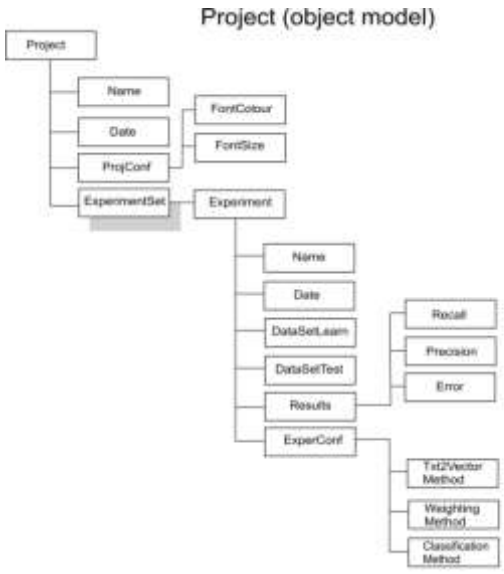

Fig. 2. Project model for ADC system

A project scheme for ADC system is represented in Fig. 2. It has *a name*, *a date* and *a project configuration* (for user's visualization preferences) properties and experiment set as a collection of experiments. It is useful to know which piece of data is used for a learning phase and a test one and what results should be shown in a log file. The property experiment configuration (*ExConfiguration*) gives the information about the text representation model, term weighting and the classification method.

[3] With the help of the COM object from www.aot.ru

3 Conclusions

In the article we suggest a conceptual scheme for an automatic document classification system (ADC). The main goal of which is to choose the best text representation model and classification algorithm for a certain application. In more detail: to compare JSM method for automatic hypothesis generation to text classification methods. That is why a project object model and its conceptual scheme are developed. The current state of the system is the following: the task of converting a text to an *n*-dimensional vector is solved. *Frequent* and *tf-idf* models for text representation are implemented. Term normalization (using the dictionary for Russian and stemming for English languages) is done.

Later the JSM method should be implemented and examined; data base scheme should be developed; experiments should be carried out and the results should be compared.

References

1. Sebastiani, F.: Machine Learning in Automated Text Categorization. J. ACM Computing Surveys vol. 34(1), pp. 1–47 (2002)
2. Finn, V.K.: Plausible inference and plausible reasoning. J. Sov Math, vol. 56(1), pp. 2201–2248 (1991)
3. Finn, V.K.: The synthesis of cognitive procedures and problem of induction. Autom Doc Math Lingust, vol. 43(3), pp.149–195 (1999)
4. Lyfenko, N.: Avtomaticheskaja Klassifikacija Tekstovyh Dokumentov na Russkom i Anglijskom Jazykah s Pomoshh'ju Metodov Mashinnogo Obuchenija. J. Molodezhnyj nauchno-tehnicheskij vestnik, vol. 4, (2013) (in Russian)
5. Cabera, J.M., Escalante, H. J., Montes-y-Gómez, M.: Distributional Term Representations for Short-Text Categorization. 14th International Conference on Text Processing and Computational Linguistics. Samos, Greece, (2013)
6. Ahonen-Myka, H.: Finding All Maximal Frequent Sequences in Text. Proceedings of the 16th International Conference of Machine Learning ICML-99 Workshop on Machine Learning in Text Data Analisys, eds. D. Mladenic and G. Grobelnik, pp.11-17, J. Stefan Institute, Ljubljana, (1999)
7. Anshakov, O.M. The JSM method: A set-theoretical explanation. Automatic Documentation and Mathematical Linguistics 46 (5),pp. 202-220,(2012)
8. Kotelnikov, E. V.: Using JSM Method for Sentiment Analysis. 3rd International Conference on Science and Technology Held by SCIEURO in London, pp.56 (2013)
9. Lebanon, G., Mao, Y., Dillon, M.: The Locally Weighted Bag of Words Framework for Document Representation. J. Machine Learning Research. vol 8, pp.2405–2441, (2007)

Концептуальная схема системы классификации текста

Николай Д. Лыфенко

Российский государственный гуманитарный университет
lyfenkoNick@yandex.ru

Аннотация. Предлагается концептуальная схема для решения задачи автоматической классификации текста. Рассматриваются различные представления текстов на естественном языке, а также статистические и логико-комбинаторные методы анализа текстов. Новизна система заключается в имплементации ДСМ метода автоматического порождения гипотез – оригинальной технологии интеллектуального анализа данных, разрабатываемой в России различными группами исследователей.

Ключевые слова. Классификация текста, машинное обучение, обработка естественного языка, интеллектуальный анализ данных.