THE SYSTEM OF MULTILINGUAL TEXT DATA PROCESSING ON THE BASE OF THE MODIFIED ALGORITHM SVM

Oleg Moyseenko

Abstract: *The article describes the architecture and principles of work of the developed system of automated processing of large volumes of textual information. The system is based on the modification of the free software library LibSVM and on the implemented in it method of support vector machines. The system performs functions of search, classification, categorization, and clustering of text documents requested by the user.*

Keywords: *classification, categorization, clustering, text documents processing.*

ACM Classification Keywords: *H.3.3 Information Search and Retrieval, H.3.1 Content Analysis and Indexing.*

Introduction

The most widespread form of knowledge introduction is natural language (NL) texts. Intensive growth of text arrays is the reason why it is difficult to access target knowledge that is stored in them. Information retrieval systems are not designed to solve this problem, because they operate with syntactic component, not with the semantic one. Therefore there occurs the necessity of systems of knowledge extraction from NL texts with subsequent automated processing.

The disadvantage of existing developments of text data processing is their dependence on natural language representation of documents content and presence of inclusions in other languages, as well as a narrow focus on data processing, for example, only social networking, news feeds or documents relating to the same subject area (workflow tasks). Such disadvantages are characteristic to almost all modern and popular systems, e.g.: search systems of the network Internet, Intelligent Miner for Text, Oracle Text, Text Miner (SAS), Statistica and others.

So, development of the system of dynamic collections of text data processing without reference to the subject area and language-specific representation of these data is an important practical and scientific task.

Usage of SVM for the work with text data

We have tested some of these systems. It gave us the possibility to choose the most appropriate classification method that could help to decide one of the mentioned problems, in particular to classify multilingual text data collections.

SVM method (Supporting Vector Machines) [1] is one of the modern and successful methods of solving of a number of tasks:

- classification – assignment to the class with the specified properties/parameters;
- categorization – correlation with the hierarchical class system;
- clustering – creating subsets of thematically close data.

At present, one of the issues that occur while solving the mentioned tasks with the help of SVM, is the necessity to adapt the method to concrete goals, in our case, for the automation of analytical processing of dynamic collections of heterogeneous, multilingual, natural language text documents. There arises the need to change the standard set of internal parameters of the algorithm, namely such parameters that are defined by a user and are not changed while training [1]. It is also important to develop new software mechanisms and approaches for the modernization of the method to achieve necessary high results in this area..

Using SVM for the work with text data, the tasks it is supposed to solve become the following:

- Classification of multilingual documents by the following features:
- name;
- document language;
- presence of multilingual inclusions (they may more fully convey the content meaning);
- storage period (temporary and permanent storage period);
- requirements level (information and prescriptive – obligatory for execution);

1. Categorization – the next step to outline the thematic component of the document (semantics). For this it is necessary to divide the text into components using the headings and numbering. The system of categories includes headings of parts, chapters and sections, which are also numbered and, in their turn, are divided into paragraphs. A paragraph is an indent at the beginning of the first line of a certain part of the text, as well as the part of the text, which is between two such indents. Sentences related to each other in meaning are joined into a paragraph. Indents of one paragraph or a chapter also have to be connected by meaning and arranged in a logical sequence.
2. Clustering –selection of groups of documents that have similar meaning (thematic groups).

The choice of the SVM method to implement the system of text data processing is caused by the lack of ready-made solutions for the problem of clustering of dynamic collections of multilingual text documents that contain not only text bodies in different languages, but also allow the inclusion of foreign words, that are in fact noises, to which teaching methods of classification, and SVM among them, have poor stability.

The changes in the work of the basic SVM method to achieve the objective are described in the author’s article [2]. As a starting configuration was chosen a noncommercial project with an open software code, created for scientific research and experiments in the SVM method work. It was in the library SVMLight [3] that the algorithm was most fully and successfully implemented. The library is presented in different programming languages.

Due to the fact that an ideal classification method does not exist, further [2], in addition to the modified basic version of SVM, we have begun to work on the creation of additional, also changed program modules of classifiers based on a number of basic techniques (k-neighbors, decisions trees etc.), that will definitely allow us to achieve stable results in the processing of text data in different natural languages. In the architecture of the system being developed of NL texts processing, the modified algorithms are called the processing agents.

The architecture of the system being developed

Due to the large number of functional elements that are used in the process of analysis of a document and in search of the necessary document among the ones placed in the system, as well as to provide the quick change of the work pattern from the local to the distributive one, the following architecture of the system of processing of multilingual dynamic collections of text data has been implemented, as shown on the picture below.

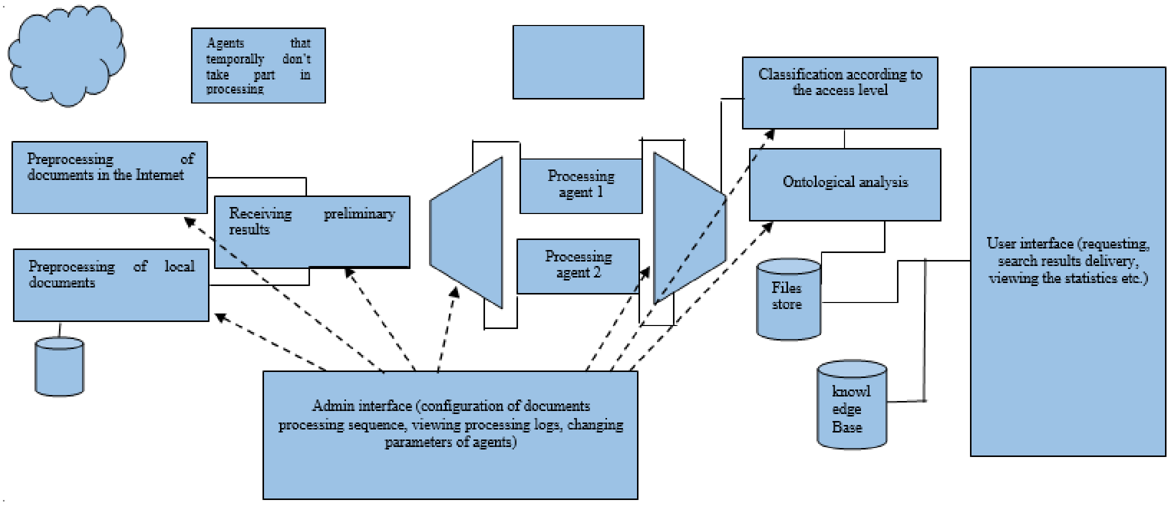
Agents that temporally don't take part in processing
Preprocessing of documents in the Internet
Receiving preliminary results
Preprocessing of local documents
Processing agent 1
Processing agent 2
Classification according to the access level
Ontological analysis
Files store
knowl edge Base
User interface (requesting, search results delivery, viewing the statistics etc.)
Admin interface (configuration of documents processing sequence, viewing processing logs, changing parameters of agents)

Fig. 1 – The architecture of the system of multilingual collections processing

The architecture offered consists of simple blocks that are common programs - executables. Their functional possibilities are quite narrow, that is why they are small enough and easily replaceable when necessary.

The issue of the need to adapt the proposed methods and algorithms for working with new input data can be solved using multiagent approach.

The main idea of the mentioned approach is in the division of a set of tasks of data processing between the agents-analysts, identifying of local regularities on the received subsets and bringing together of the acquired knowledge to form the final result.

The so called data collection agents are used to read data from different sources and transfer them to analysis agents. In addition, the managing agent has been implemented, which is placed outside the architecture in the picture and is the coordinator of the work of other agents and is responsible for the distribution of work between them. The organization of interaction between the agents depends primarily on the task, the methods used and algorithms (language detection, categorization, classification etc.).

Blocks of the system interact with each other using the interface of input-output channels, namely linking the standard output of the program-source with the standard input of the program-consumer. Such communication enables without any change of the existing components to analyze data that is sent from a previous agent-processor to the next, extend the set of data that is transmitted or even completely replace them.

As the data structures, used for the interaction between the system elements, text lines in the form "Key: Value" have been used. If it is necessary to extend the meaning for several lines, one can use the indentation (blank spaces or tabs) in the beginning of the line, which should be appended to the previous one. After the lines with keys and values it is allowed to transfer arbitrary data after a blank line. In this case the lines with keys and values form the heading.

In order to reduce the flow of data that is transmitted through the conveyor of agents, large amounts of data that come from outside (the original text files of different presentation formats) or created by intermediate handlers are stored in files and transmitted only by their names. It is important to be able to use the same structure of the intermediate processing (to simplify errors searching and audit of the system). For this purpose, when a new file comes, a catalogue is created in the local file system, and later – in the global store, and the file is processed only in it. Since the agent list can be expanded at any time, it is possible to customize the names of source files by using namespaces (name of the agent is used as a prefix).

If necessary, separate agents are transferred almost unchanged from the model of a separate process to the model of the web-service (SOAP or REST with data transfer of JSON or XML formats) or to the model of the agent connected to a string of messages.

Another essential function of each agent is output of operation log. The level of detail of the log is defined by command line arguments and/or an environment variable. This log is particularly needed in the process of refining of the system and for the initial setup.

Conclusion

The processing system of dynamic arrays of multilingual text data, some aspects of which have been stated in this article, needs further development and improvement of the operating results. The results of comparative experiments of the work evaluation of the classifier of the system and a commercial product StatSoft Statistca will be given in the next article.

Several processing steps may require different approaches for the solution of the same issue. For example, a comparison of documents in different languages with a particular subject domain requires different vocabularies, different areas of lines processing (LTR or RTL). To simplify the implementation of the system it is proposed to create a base program (and a class of agents) of switching outputs (demultiplexer), which, depending on the values of a certain key will cause one or other next element. Of course, it may take a class of multiplexers which will unify the original data depending on the previous agent.

Gratitude

Professor V.V. Lytvynov for the invitation to publish and help in reviewing

Bibliography

1Vapnik V. Statistical learning theory. Wiley, New York, 1998. – [Electronic resource]. – Access mode to the resource: http://books.google.com.ua

2 Мойсеенко О.П. SVM при классификации мультиязычных текстов / В.В. Литвинов, О.П. Мойсеенко// Вестник ЧНТУ. Технические науки. –2013. – №4 (66).– С. 119–123

3 SVM-Light Support Vector Machine. – [Electronic resource]. – Access mode to the resource: http://www.svmlight.joachims.org

4 Мойсеенко О.П. Автоматизованная система обработки динамических коллекций разноязычных текстовых документов по морскому и речному делу / Литвинов В.В., Мойсеенко О.П. // Математические машины и системы. – 2014. – № 2. – С. 59 – 64.

5 Ю. Сімакін Методи і алгоритми багатокористувацької обробки розподіленої просторової інформації регіону на основі технології геопорталів/ С. Кривоберець, Ю. Сімакін // Вісник ЧДТУ – 2010. - №42. - С. 213-221.

Author Information

***Oleg Moyseenko** – Postgraduate, the assistant lecturer, Chernihiv National University of Technology, 95, Shevchenko street, Chernihiv, Ukraine, 14027; e-mail: OlegMoyseenko@gamil.com*

Major Fields of Scientific Research: expert systems, automated speech processing