Rubryx: Technology of Text Classification Using Lexical Meaning Based Approach[1]

Vladimir Polyakov[2]
vladimir_polyakov@yahoo.com

Vladimir Sinitsyn[3]
sowsoft@land.ru

Abstract

The paper deals with the technology of automatic text classification under thematic categories. The characteristic feature of it is the controlled dictionary approach, i.e. the use of a preset thematic dictionary. But the technology doesn't involve high-cost manual indexing of terms and categories. It is based on machine learning from a set of documents previously classified by domain experts. Three to five documents representing each category are sufficient for machine training. As a rule, these sample documents are course books or review articles. Another characteristic feature here is the use of multi-word terms. A considerable efficiency of the use of multi-word terms in text classification is due to an indirect limitation of the lexical meaning.

Introduction

Nowadays there is a sharp increase in the number of Internet resources. Therefore, the problem of effective search of documents in Internet becomes more pressing with each day, and the distribution of texts in electronic form stimulates the majority of research in text analysis.

One of the possible ways to solve the problem is to work out a great number of specialized web-catalogs. Such resources deal with a certain thematic domain and are aimed at certain users. It is much easier and faster to find the information in specialized web-catalogs than with the help of universal search engines. The development of a good specialized web-resource involves processing of a huge amount of information. Manual processing is labor- and time-consuming. Taking into account that potential developers of specialized resources are mostly limited in funds, a low-cost high-quality technology of automatic text processing is required.

The paper covers the research and development of automatic text classification technology and the software based on this technology.

1 Description of the Approach Used

The developed classification technology makes use of expert information and a terminological dictionary. An expert organizes several rubrics (categories) where he places a number of sample documents. Then the program gathers general information on each category from these documents and classifies all other documents (Pic. 1).

Pic.1 Task Solution Flowchart

1. Organize a terminological dictionary.
2. Collect text corpus.
3. Compile a directory.
4. Cycle in each category.

4.1. Select sample texts for the category (five documents).
4.2. Generate a micro-dictionary for the category.
4.3. Set a threshold.
4.4. Carry out a sampling classification.
4.5. Correct the micro-dictionary and the threshold.
4.6. Carry out a complete classification under the category.

5. Edit the directory and place it in Internet.

Each text is characterized by terms and their frequency of occurrence. General information on a category also contains a list of terms and their

[1] Research was supported by Russian Foundation of Basic Researches (grant # 02-07-90413)
[2] Moscow State Institute of Steel and Alloys (Technological University), Moscow State Linguistic University
[3] GNIVC MNS RF

frequency. The higher is the coincidence of information on the text and the category, the higher is the probability of its relevance. To measure the degree of relevance of a text to a category, a coefficient K is introduced. It can take a value from 0 to 100 inclusive. The higher is K-value, the higher is the relevance of the text to a category. The coefficient's threshold is set in the program. Only the documents with K-value higher or equal to the threshold are filed under the category.
Terms can consist of one, two, or three words. The research showed that one-word terms insufficiently characterize the text. To account for this feature different weight coefficients are introduced.

2 Formal Task Definition and Solution

According to Sebastiani (1999), a general task of classification is defined as follows. To assign a Boolean value to each pair $(d_i, c_j) \in D \times C$, where D is a domain of documents, and C is a set of pre-defined categories (rubrics). A value of P_{ij} for a document i equaled as **True** means that the document files under a category j, in case of P_{ij} = False – it doesn't.
Let there be a terminological dictionary containing sets L_1, L_2, L_3 where L_1 is a set of one-word terms, L_2 – two-word terms, L_3 – three-word terms
A set of documents: $D = \{d_1, \ldots, d_n\}$
W_{1i}, W_{2i}, W_{3i} are sets of one-, two-,and three-word terms from a document d_i

1. Selection of sample documents for a category c_j.

D_{*j} is a subset of samples $D_{*j} \subset D$

2. Generation of a micro-dictionary for the category (M_{1j}, M_{2j} and M_{3j}).

Find intersection of sets of terms from sample documents and dictionary:
$M_{1j} = W_{11*} \cap W_{12*} \ldots W_{1n*} \cap L_1$ - one-word terms
$M_{2j} = W_{21*} \cap W_{22*} \ldots W_{2n*} \cap L_2$ - two-word terms
$M_{3j} = W_{31*} \cap W_{32*} \ldots W_{3n*} \cap L_3$ - three-word terms

3. Classifying

$$\left.\begin{aligned} E_{1ij} &= M_{1j} \cap W_{1i} \\ E_{2ij} &= M_2 \cap W_{2i} \\ E_{3ij} &= M_3 \cap W_{3i} \end{aligned}\right\} \quad \text{for } i = 1 \ldots n \quad (1)$$

Find cardinal number:
$N_{1ij} = | E_{1ij} |$, $N_{2ij} = | E_{2ij} |$, $N_{3ij} = | E_{3ij} |$

K_{1ij}, K_{2ij}, K_{3ij} are intermediate coefficients of one-, two-, and three-word terms from a document d_i and category c_j

$$\left.\begin{aligned} K_{1ij} &= (N_{1ij} / | M_{1j} |) \cdot 100\% \\ K_{2ij} &= (N_{2ij} / | M_{2j} |) \cdot 100\% \\ K_{3ij} &= (N_{3ij} / | M_{3j} |) \cdot 100\% \end{aligned}\right\} \quad \text{for } i = 1 \ldots n$$

K_{ij} is a relevance coefficient of a document d_i to a category c_j

$$Kij = \frac{0.2K1ij + 1.3K2ij + 1.5K3ij}{3} \cdot 100\% \quad (2)$$

$$P_{io} = \begin{cases} 1, \text{ if } K_{ij} \geq K^* \\ 0, \text{ if } K_{ij} < K^* \end{cases}$$

K* is a threshold of K

P_{ij} is a conditional probability of filing a document d_i under category c_j

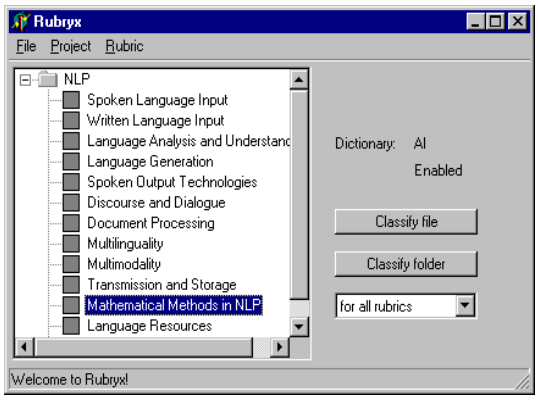
3 Classifier Description

A classifier was developed to solve the defined task. The source data for the classifier are web-resources (sites) previously downloaded from Internet. Each site contains a number of web-sites in html-format. Moreover, each web-site contains a natural-language text. The aim of the classifier is to rank the source corpus of web-documents under categories predefined by the user (classification).
Rubryx classifier is a 32-bit Windows application which consists of the main window and a number of subordinate dialog-windows.[4] The main window is represented in Picture 2.
Before automatic classifying of documents, source data should be fed in and the classifier has to be set. The source data include names of rubrics (categories), several sample documents for each category, a full path to the folder with the documents to be classified.
First, there should be defined the number of categories under which the source documents are to be filed, and the structure of a catalog. The catalog has a tree-structure and includes sections and rubrics (categories). An example of a catalog is represented in Picture 2.

[4] The second version of Rubryx is described.

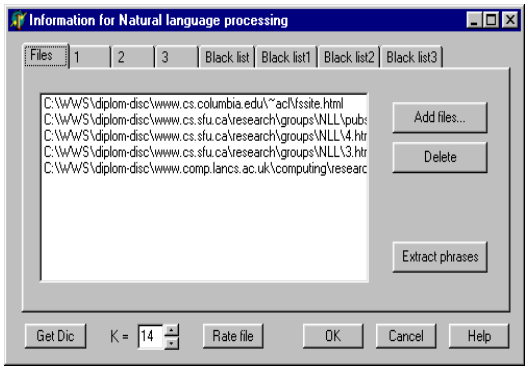
A number of corresponding sample documents should be related to the category. To do that, one should choose the rubric (category), and choose "Properties" in its context menu. Picture 3 shows a dialog window to select files, button "Add files…" adds files to the list and button "Delete" deletes files from the list. When the list is ready one should click on "Extract phrases" to get the list of most significant phrases from the input documents.

Pic.2 The main window[5]



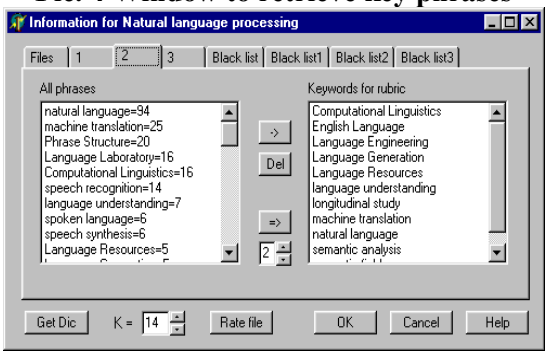
Then, one should retrieve keywords for the category from the list of phrases. Picture 4 provides an example of the list of significant phrases. In this example, the general thematic dictionary for the category is not used. While using the dictionary, the list "All phrases" is generated only of the phrases to be found in both the sample documents and the general dictionary.

Pic.3 Window for the selection of sample documents



One can retrieve the key phrases either manually with "–>" button, or automatically with "=>" button. In case of automatic retrieval, the threshold of significance must be set – the number of sample documents where the phrase can be found.

Pic. 4 Window to retrieve key phrases



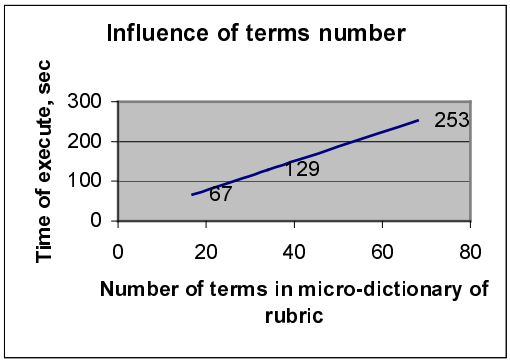
Once the key words for the category have been extracted, one should click on "Get Dic" to generate a micro-dictionary for the category and set the value for the threshold coefficient K. The documents with the coefficient higher or equal to K will be filed under the category. The classifier offers a default value depending on the samples, but manual correction is advised. To serve the purpose, there is a button "Rate file" to measure the K-value for a definite file.

On clicking "OK" the category is ready. All other categories from the catalog should be compiled likewise. After that, a folder with documents can be classified on clicking "Classify folder".

4 Experimental Evaluation

The dependence of classification time from two parameters, namely the number of terms and the size of the documents, was analyzed. The experiments showed the linear dependence in both cases.

Pic. 5. Dependence of classification time from the number of terms.

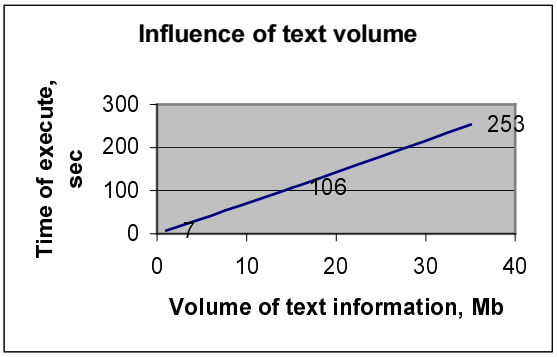


[5] The list of NLP categories in this example are formed on the base of contents of the book (Cole, Ronald, et al 1998).

Let T be the time necessary to carry out the algorithm, A be the sum of terms in micro-dictionaries for the categories, B be the size of the documents in bites and C be the time necessary for the classifier to process one bite. Then T is calculated according to the formula:

T = A *B * C

Pic. 6. Dependence of classification time from the size of documents.



The result of the program is a classifier which sets links between a definite category and resource addresses. The output data of resource (texts) processing are stored in a file (Picture 7) which contains the information about the micro-dictionary, the number of terms and values of the coefficients.

Рис.7. Output data[6]

http://www.cs.columbia.edu\~acl\FsBackIssues\finstring22_1.html
computational=57(25)
grammar=0(21)
grammatical=1(5)
information=39(18)
lexical=16(6)
linguistic=9(7)
parsing=1(8)
processing=12(15)
Semantic=7(17)
translation=12(10)
computational linguistics=20(7)
English language=0(3)
language engineering=0(4)
language generation=2(4)
language laboratory=0(10)
language resources=1(3)
language understanding=2(3)
longitudinal study=0(2)
machine translation=7(9)
natural language=16(42)
phrase structure=0(10)
semantic analysis=0(2)
semantic field=0(2)
semantic representation=0(1)
speech recognition=1(8)
text analysis=0(2)
natural language interface=0(9)
natural language processing=4(3)
phrase structure grammar=0(10)
66 32 14 25

K_1=66 K_2=32 K_3=14 K=25

5 Analysis of the Obtained Results

The review given by Sebastiani (1999) discusses the basic principles of text classification. According to this paradigm, the text classification method by Rubryx can be characterized as follows:
- is based on a controlled dictionary;
- uses phrases in ranking texts;
- uses machine learning technology;
- uses hard-classification;
- uses multi-label text categorization;
- uses both category-pivoted and document-pivoted[7] text categorization.

Moreover, another characteristic feature of the program can be added to the list, which hasn't been widely used, yet is highly perspective, namely lexical meaning based approach.

Let's dwell upon advantages and limitations of Rubryx technology. The drawback of traditional methods based on dictionaries and thesauruses is considered to be the high price of manual training, where each keyword or phrase is to be manually assigned to a corresponding category. In our approach, there is no need to do so, as our approach provides a technique of machine learning from a set of previously classified documents. Tentative experiments showed a high efficiency of training used in Rubryx. Another significant result is the proof of a positive effect of key phrases in text. Unlike the two approaches described in Sebastiani. (1999), the first one being based on syntax and the second on statistics, we have chosen a dictionary-based approach which uses phrases. It means that

[6] Each term is followed by two numbers: occurrence of the term in the text and occurrence of the term in the micro-dictionary of the category (in parenthesis).

[7] In the second version of Rubryx.

the basic criteria to include the phrase into the dictionary is it being a lexical collocation, i.e. when it is impossible to insert another word between its members, or it is possible in very rare cases. Basically, this approach can be interpreted as statistic in the sense that the co-occurrence of keywords is analyzed and the words with the most frequent co-occurrence are selected, but we believe that the basic principle here is the terminological approach whereas the statistic regularity is the result of terminological factors.

A characteristic feature of the use of multi-word terms in Rubryx is that a high weight is assigned a priori to the terms consisting of two or three words (N=2 and N=3), which can be seen from the formula (2). Thus, for one-word terms the weight is w_1=0.06(6) (0.2/3), for two-word terms it is w_2= 0.43(3) (1.3/3), for three-word terms it is w_3=0.5 (1.5/3). These weights were estimated empirically during tentative testing of Rubryx. In choosing the relation of weights, the following normalizing condition was met: $w_1+w_2+w_3 = 1$

The choice of a higher weight for phrases with N > 1 was made due to the fact that terminological phrases contribute more to the rank of a document rather than single-word terms. It is explained by means of a fundamental linguistic attribute of terminological phrases to limit the sense of its members. For example, separate words *"machine"* and *"learning"* are less representative for classification purposes than the phrase *"machine learning"*. Terminological phrases naturally limit the lexical meaning of its members; therefore we regard our approach to be based on lexical meaning. But it should not be confused with the task to use classification for word sense disambiguation (WSD). In this case, an indirect disambiguation, or rather limited ambiguity, increases the quality of classification.

We declined explicit use of phrases with words N > 3 in classifying due to two reasons. The first one is a relatively low occurrence of such phrases. The second is the fact that the use of such phrases leads to an excessive number of searches. Yet these phrases can be split to three-word phrases, e.g., the phrase *American National Standard Institute* can be represented as a couple of three-word phrases: *American National Standard* and *National Standard Institute.*

There might seem to be some limitations to the use of Rubryx technology due to a number of circumstances. We have tried to minimize the influence of such limitations.

One is a well-known limitation due to the use of a thematic dictionary. The change to another thematic domain would involve the generation of a new dictionary.

To solve the problem, the following measures are proposed.

There are a great number of thematic dictionaries which can be loaded into the program at low cost and effort. Besides, glossaries which can be found in most course books and manuals on definite scientific domains can be used as dictionaries.

A special program DictTools to deal with dictionaries was developed to provide its compatibility with Rubryx[8].

For new thematic domains which have no definite lexicographic bases yet, a technology of retrieval of set expressions in texts described in Pavlov and Polyakov (2001), can be implemented.

While classifying texts of multi-thematic domain, a merging of dictionaries for different thematic domains into one multi-thematic dictionary can be applied. As an example, there is a lexicon from a polytechnic dictionary (1997).

Another limitation is caused by the necessity to employ a number of experts in each domain for which a classifier is developed. To resolve the problem, the following can be advised. In fact, this limitation is intrinsic to all methods based on machine learning from a set of previously classified documents. Nowadays a huge number of texts are made available in Internet and there are catalogs of classified texts, so the selection of sample documents doesn't seem to be a great problem. The only requirement for an expert is a minor operational experience in Rubryx, because the selection of representative documents and the setting of relevance threshold of a text to a category require some skills in the program. As a rule, we used chapters from course books devoted to a definite topic or reviews to train Rubryx classifier. Another effective source for Rubryx training is programs and Call for papers of strictly specialized scientific conferences which contain a high percentage of terminology.

Traditional applications based on a machine training paradigm involve afterlearning from a set of texts. Furthermore, more sample texts usually provide better classification. In our method, afterlearning is not implemented so far because an increased number of sample documents for training

[8] It will be sited together with the second version at www.rubryx.narod.ru and www.sowsoft.com/rubryx .

may impair rather than improve the result. It is due to the peculiarity of the generation of micro-dictionaries which are created from the intersection of sets: sets of terms from a thematic dictionary and sets of terms from sample documents selected for training (see formula (1) in Task Definition). Yet special research is being carried out to remove the limitation.

One of the main limitations of our approach seems to be the thematic processing in Rubryx, as topics normally have a set of terms as attributes. To classify under genres or styles, the effectiveness of the method can be low. Basically, it can be regarded a drawback but for some genres including specific lexicon. For example, to classify texts of dialects, thieves' or terrorists' cants, and taboos, the use of the classifier is possible. So this limitation cannot be viewed as absolute either.

Conclusion

The outcome of the research is the development of methods and formal task definition for automatic categorization of natural language texts using the controlled dictionary approach. The characteristic feature of this approach is machine learning from pre-defined sample documents. It allows a considerable time saving for classifying alongside effective output. Another feature is the use of terminological phrases with an indirect limitation of the lexical meaning of keywords which leads to an improved classification.

This approach can be applied in libraries and in development of Internet catalogs of thematic resources.

References

F. Sebastiani. (1999) *Machine learning in automated text categorisation: a survey. Technical Report IEI-B4-31-1999*, Istituto di Elaborazione dell'Informazione.

Pavlov O.A., Polyakov V.N. (2001), *Frequency method of stable phrase collocation revelation.* Proc. of Kazan school in computational and cognitive linguistics. TEL-2001. Issue 6. Kazan. October 22-28. 2001 г. (In Russian)

English Russian Large Polytechnic Dictionary. «RUSSO», 1997. 200 thousands articles.

Survey of the State of the Art in Human Language Technology / Cole, Ronald, et al (eds.) Studies in Natural Language Processing . Cambridge University Press 1998. 533 pp.