elektrorevue
ISSN 1213 - 1539

Advanced Text Categorization Methods with Statistical Approach

Ján Tóth, Anna Kondelová, Gregor Rozinaj

Faculty of Electrical Engineering and Information Technology, STU in Bratislava
Email: {jan.toth, anna.kondelova, gregor.rozinaj}@stuba.sk

Abstract *– Text categorization is a basic and important operation in text processing, therefore automatization and efficiency are key attributes for this operation. In this article, more methods for text categorization which are based on statistical approach and their adaptation to the Slovak language have described. Furthermore, the article deals with the issue of profile creation and at the end some results are shown.*

1 Introduction

Text categorization is an important task in text processing which allow the automatic handling of enormous numbers of documents in electronic form. The motivation for creating a text categorization system is its usage in speech synthesis systems. Within the text categorization can be given additional synthesis information to increase naturalness of synthesized speech.

One problem in finding domain profile of documents are different kinds of a textual errors, such as spelling and grammatical errors in text, and character recognition errors in documents that come from OCR. Last but not least it is the flective character of Slovak language which is represented by different prefixes and suffixes generated by inflection.

In this article is described an N-gram-based approach to text categorization that is partly tolerant to textual difficulties which are described above. This system works well for domain classification, achieving up to 80% correct classification rate in tests on Slovak National Corpus articles written in Slovak language.

The system process is based on calculating and comparing domain profiles of N-gram frequencies. In first step, are created profiles for every domain from training data. In testing phase system also creates profile for testing text which is to be classified. In final step, the system computes a distance measure between tested document's profile and each of the domain category profiles. The smallest distance indicates the domain category of the tested document.

Following features are characteristic for this implementation of described algorithm:

- Categorization to the domain is robust against language inflection, spelling and grammatical errors
- Domain's profiles are storage friendly and consuming little storage
- Domain's profiles size are independent on database testing tests

In this paper text categorization building process is summarized. The rest of this paper is structured as followed: Section 2 deals with N-Grams and state of the art of text categorization in general. In section 3 domain profile training process is described. In section 4 the testing process on the texts from Slovak National Corpus is discussed. In this part differences between profiles comparison method based on normalized and non-normalized N-gram frequencies are described. In section 5 achieved results and comparisons are presented. Section 6 concludes this paper with future work.

2 N-Grams

Text categorization is the process of automatically determining text categories according to text content within a given taxonomy system. There are several different methods for text categorization including statistical-based algorithms, Bayesian classification, distance-based algorithms, k-nearest neighbors, decision tree-based methods. Many different evaluation functions have been proposed, such as Term Frequency (TF), Document Frequency (DF), Term Entropy (TE), Mutual Information (MI) etc. Many researchers have applied and compared different evaluation functions. Their results show that MI method has some advantage compared with others [1]. The most classic and widely used method is *tf.idf* (term frequency & inverse document frequency). There are two main factors that should be considered in the *tf.idf* method: 1) term frequency (*tf*), namely the number of a feature in the context; 2) inverse document frequency (*idf*), which is a quantitative representation of the feature's distribution in the text set [2]. In [3] authors describe simple classification distance algorithm based on N-grams frequencies and their positions. This system work well for subject classification and achieve 80% correct classification rate.

In this paper some improvements are performed and decision criterion of used categorization method based on tests is determined. In this work two methods are compared: 1) categorization based on N-gram non-normalized frequency 2) categorization based on N-gram normalized frequency (*tf*). In next parts of this paper these two methods are described.

In our definition, N-grams are sequences of n consecutive characters which are generated from words. In some applications N-grams can be formed by bigger texts parts e.g. words etc. Since number of possible strings of length N is a lot smaller than number of possible single words in a language, therefore character N-gram has smaller dimensionality [3]. System that uses N-grams has to be resistant to some noisy characters such as dashes, points, comas, quotes and many others witch carried no information. The key benefit of N-gram-based matching derives from its very nature: since every string is decomposed into small parts, any errors that are present tend to affect only a limited number of those parts. Other

N-grams are correct however some are incorrect. If we count N-grams that are common to two texts, we get a measure of their similarity that is robust to a wide variety of textual errors.

If word "football" is considered, correct generated tri-grams in Figure 1-a are shown. Misspelled word "footboll" will generate tri-grams shown in Figure 1-b.

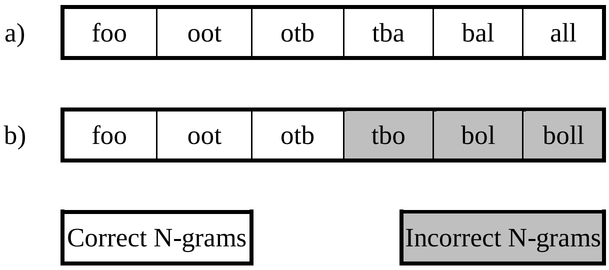

Figure 1: Correct and incorrect generated N-grams.

As we can see, if from some reason we have error in processed word, only three N-grams are incorrect.

3 Generating N-Gram domain profiles

Human languages have some words which occur more frequently than others. One of the most common ways of expressing this idea is known as Zipf's Law [4], which is define as:

> *The nth most common word in a human language text occurs with a frequency inversely proportional to n.*

That said there is always a set of words which dominates most of the other words of the language in terms of frequency of use. Figure 2 shows the distribution of N-grams frequencies from sport domain profile. As we can see this dependency is inversely proportional and therefore it comply with Zipf's law.

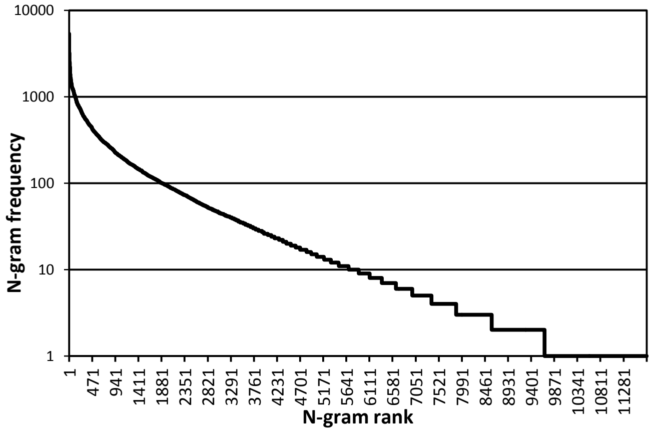

Figure 2: Example of a Zipfian distribution of N-gram frequencies from sport domain profile.

Zipf's Law implies that classifying documents with N-gram frequency statistics will not be very sensitive to cutting off the distributions at a particular rank. It also implies that if we are comparing documents from the same category they should have similar N-gram frequency distributions. On this idea we have built our categorization system. Figure 3 illustrates training process our system.

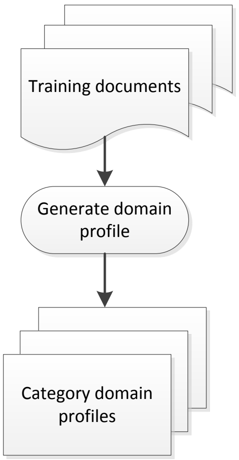

Figure 3: Domain profiles training process.

As training documents Slovak National Corpus database system is used [5]. In this database manually annotated corpus r-mark-3.0 is included. In this corpus every text has information about its domain. In the current work three domain category are trained: culture (clt), technics (com) and sport (spo). From the database of the documents we have created samples with the same reasonable size. From these, we would generate a set of N-gram frequency profiles to represent each of the categories. In this operation this following steps are done:

- Split text into separate tokens formed by letters. Digits, punctuation, dashes and other noisy characters are discarded. We also removed repetitive white spaces.
- Generating N-grams from each token.
- Fill the table with N-grams and their frequencies. Each N-gram has own counter.
- Sorting this table. The table is ordered by the N-grams frequencies.
- Export table as domain profile file.

The resulting files are N-gram frequency domain profiles for each domain category. When these profiles (N-gram frequencies) are plotted we get Zipfian distribution shown in Figure 2.

4 Testing N-Gram domain profiles

As we mentioned, testing text have similar N-gram frequency distribution as the N-gram frequency distribution of its domain profile category. Therefore in this step we need to repeat the same process as when we trained domain profile. In result we get N-gram frequency profile for tested text. Now we have to calculate and compare N-gram distribution of tested text with each domain profiles; in our case: culture, technics and sport. The whole process is shown in Figure 4 below.

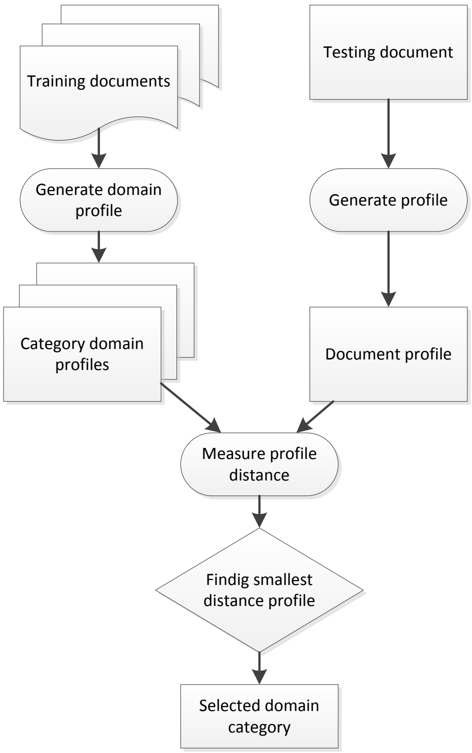

Figure 4: Whole process of text categorization system.

4.1 Non-normalized N-gram frequency comparison method

In part “Measure profile distance” process following simple statistical measure is used. This measure determines how far out of place is N-gram in one profile from its place in the other profile [6]. For each N-gram in the testing text its counterpart in the domain profiles is found, and then calculate how far out of place it is. This calculation is based on non-normalized N-gram frequency. Simple example of this calculation is shown in Figure 5 below.

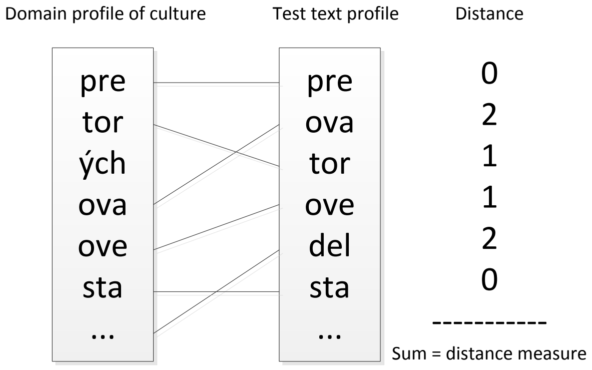

Figure 5: Example of the N-Gram distance measure calculation.

N-gram distribution shown in Figure 5 is not real and is created only for example purposes. For test text distances are calculated between its profile and each domain profile. Domain profile that has the smallest distance is marked as most similar to the test text profile. Based on this distance we declare also domain category of tested text.

4.2 Normalized N-gram frequency comparison method

The second method to calculate distance between domain profiles and profile of testing text is based on normalized N-gram frequency with sum of all N-gram frequencies in given domain profile shown in formula (1) below [7]:

$$tf_i = \frac{n_i}{\sum_k n_k} \qquad (1)$$

where $\mathbf{n_i}$ is the number of occurrences of N-gram in a given profile. The denominator represents the sum of the frequencies of all N-grams occurred in a given domain profile.

In this case distance between normalized N-gram frequency of domain profiles and testing text same as in previous, non-normalized method are described.

4.3 CCR dependence on the length of text

By testing booth used methods of text categorization we were interested in determining correct categorization rate of testing texts and also the dependence this CCR by length of testing texts. CCR was calculated using the formula (2) shown below [8]:

$$CCR = \frac{\#\, correctly\ categorizated\ texts}{\#\, categorizated\ texts} \times 100\ [\%] \qquad (2)$$

In the next part of this article achieved results from our tests as well as a comparison dependence of CCR from testing texts length are mentioned. Achieved results show that booth of tested methods for profiles comparison (non-normalized and normalized) achieve different values depending on the length of tested texts. Measured results also show that there is a threshold which determines the appropriateness of using one of the methods to achieve highest value of CCR.

5 Achieved results

As we mentioned we have trained three domain categories: culture, technics and sport from training documents. The result was three domain category profiles. Each profile has about 11000 numbers of N-grams.

5.1 Non-normalized N-gram frequency comparison method

Test sample from group of 100 testing texts with length of 100 characters is consisted. In this test correct categorization rate (CCR) is calculated. This test several times to achieve more accurate results was performed. In Table 1, Table 2 and Table 3 results for each trained domain category are presented.

Table 1: Culture testing texts.

test #	1	2	3	4	5	6	7	8	9	10
CCR[%]	75	77	83	80	79	85	79	82	81	80

Table 2: Technic testing texts.

test #	1	2	3	4	5	6	7	8	9	10
CCR[%]	76	79	81	81	83	85	75	81	82	79

Table 3: Sport testing texts.

test #	1	2	3	4	5	6	7	8	9	10
CCR[%]	82	79	83	80	75	81	81	83	77	79

As we can see we achieved Correct Categorization Rate from 75 to 85 percent for each domain.

In the next test we were interested in how Categorization Correct Rate depends on the length of the tested text. These tests on each categorization domain were performed. In Figure 6, Figure 7 and Figure 8 CCR dependencies on text length are shown. In Table 4, Table 5 and Table 6 our achieved results for each category are shown.

Table 4: Sport testing texts.

Text length[chars]	10	20	40	60	80	100	150
CCR [%]	56	70	78	78	86	88	90

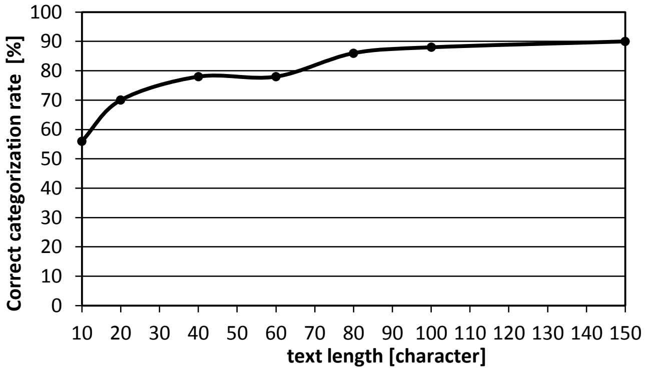

Figure 6: Correct categorization rate depend function on length of sport testing texts.

Table 5: Technic testing texts.

Text length[chars]	20	40	60	80	100	150
CCR [%]	61	70	76	80	82	86

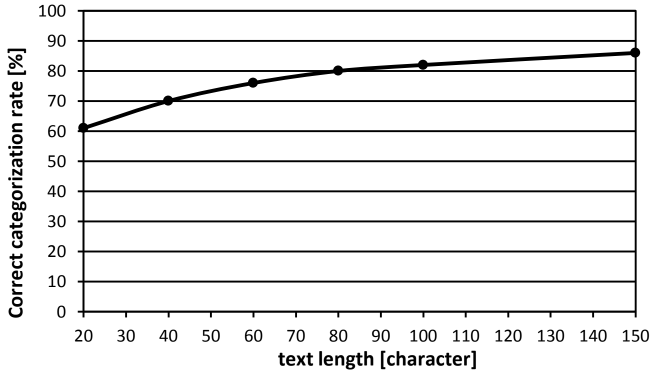

Figure 7: Correct categorization rate depend function on length of technic testing texts.

Table 6: Culture testing texts.

Text length[chars]	20	40	60	80	100	150
CCR [%]	37	47	54	56	60	63

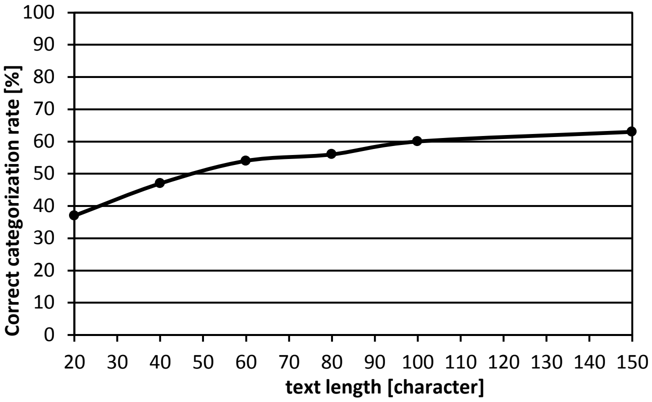

Figure 8: Correct categorization rate depend function on length of culture testing texts.

5.2 Normalized N-gram frequency comparison method

As we mentioned before, we compare method for profiles comparison based on non-normalized N-gram frequency with method based on normalized N-gram frequency in profile. We compare CCR dependence on length of tested texts. Achieved results in Figure 9, Figure 10 and Figure 11 for each category are shown.

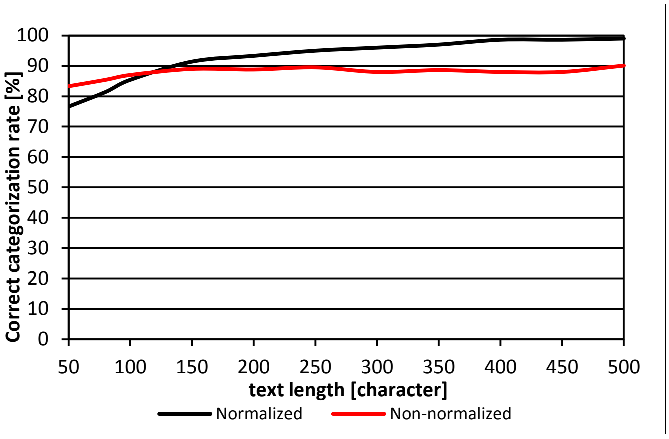

Figure 9: Achieved CCR by normalized and non-normalized N-gram frequency in technic testing texts.

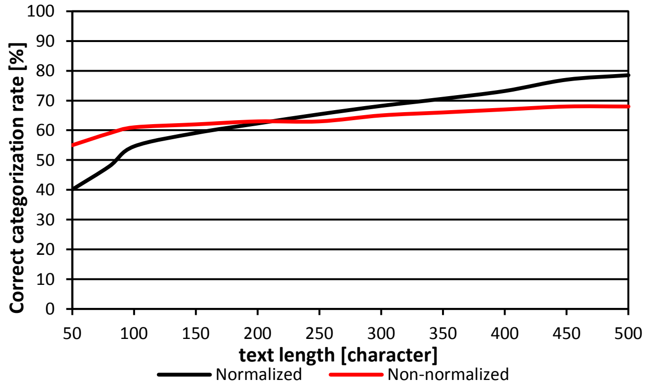

Figure 10: Achieved CCR by normalized and non-normalized N-gram frequency in culture testing texts.

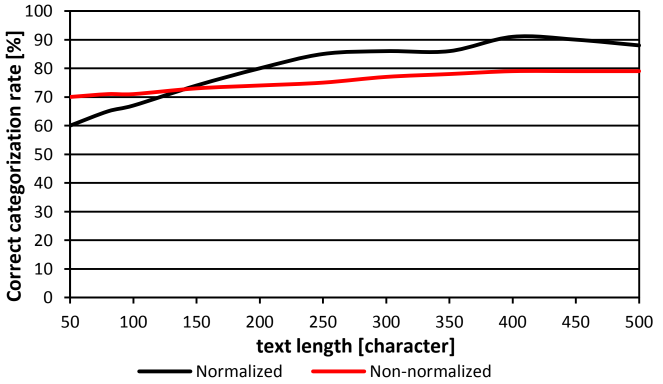

Figure 11: Achieved CCR by normalized and non-normalized N-gram frequency in sport testing texts.

Our shown figures results that each of used methods for different sizes of tested texts is suitable. An interval of the measured thresholds of our tested texts is from 117 to 210 characters.

The length ranges of tested texts and appropriate methods for profiles comparison according to them are shown in Table 7.

Table 7: Length thresholds for each trained domains.

	Non-normalized method	Normalized method
Technic	length ≤ 117 chars	length > 117 chars
Culture	length ≤ 210 chars	length > 210 chars
Sport	length ≤ 140 chars	length > 140 chars

6 Conclusion

In this paper our text categorization system with statistical approach based on N-grams is presented. Two methods for N-gram profile comparison based on normalized and non-normalized N-gram frequency are compared. CCR dependence on length of tested texts using both methods is presented. From our tested data results each method for another length of tested texts is suitable.

This work is not dealing with text preprocessing functions like stemming or any other N-gram weighting. It could be fulfilled in another work and future research.

Acknowledgement

This work has been supported by HBB-Next - Next-Generation Hybrid Broadcast Broadband FP7-ICT-2011-7 - 287848-06, VEGA 1/0961/11, VEGA 1/0708/13 and under the program to support young researchers on STU.

References

[1] PEI, Z., SHI, X., MARCHESE, M., LIANG, Y. *Text Categorization Method Based on Improved Mutual Information and Characteristic Weights Evaluation Algorithm.* In Fourth International Conference on Fuzzy Systems and Knowledge Discovery, pp. 87-91, 24-27 Aug. 2007

[2] SALTON, G., BUCKLEY, B. *Term weighting approaches in automatic text retrieval. Information Processing and Management.* In Information Processing & Management, 1988, 24.5: pp. 513-523.

[3] CAVNAR, W.B. *Using an N-Gram-Based Document Representation with a Vector Processing Retrieval Model.* In Proceedings of the Third Text Retrieval Conference, NIST Special Publication 500-225, 1995, p. 269-277

[4] ZIPFS, G. K. *Human Behavior and the Principle of Least Effort, an Introduction to Human Ecology*.

[5] Slovenský národný korpus. *r-mak-3.0.* Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2009. WWW: http://korpus.juls.savba.sk.

[6] CAVNAR, W. B., TRENKLE, J. M. *N-Gram Based Text Categorization*. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.

[7] SUZUKI, M., YAMAGISHI, N., YI-CHING, T., ISHIDA, T., GOTO, M., *English and Taiwanese text categorization using N-gram based on Vector Space Model*. Information Theory and its Applications (ISITA), pp. 106,111, 17-20 Oct. 2010

[8] SUZUKI, M., YAMAGISHI, N., YI-CHING, T., HIRASAWA, S., *Multilingual text categorization using Character N-gram*. Soft Computing in Industrial Applications, 2008. SMCia '08. pp. 49,54, 25-27 June 2008