N-Gram Feature Selection for Authorship Identification

John Houvardas and Efstathios Stamatatos

Dept. of Information and Communication Systems Eng.
University of the Aegean
83200 – Karlovassi, Greece
stamatatos@aegean.gr

Abstract. Automatic authorship identification offers a valuable tool for supporting crime investigation and security. It can be seen as a multi-class, single-label text categorization task. Character n-grams are a very successful approach to represent text for stylistic purposes since they are able to capture nuances in lexical, syntactical, and structural level. So far, character n-grams of fixed length have been used for authorship identification. In this paper, we propose a variable-length n-gram approach inspired by previous work for selecting variable-length word sequences. Using a subset of the new Reuters corpus, consisting of texts on the same topic by 50 different authors, we show that the proposed approach is at least as effective as information gain for selecting the most significant n-grams although the feature sets produced by the two methods have few common members. Moreover, we explore the significance of digits for distinguishing between authors showing that an increase in performance can be achieved using simple text pre-processing.

1 Introduction

Since early work on 19th century, authorship analysis has been viewed as a tool for answering literary questions on works of disputed or unknown authorship. The first computer-assisted approach aimed at solving the famous *Federalist Papers* case [1] (a collection of essays, a subset of which claimed by both Alexander Hamilton and James Madison). However, in certain cases, the results of authorship attribution studies on literary works were considered controversial [2]. In recent years, researchers have paid increasing attention to authorship analysis in the framework of practical applications, such as verifying the authorship of emails and electronic messages [3,4], plagiarism detection [5], and forensic cases [6].

Authorship identification is the task of predicting the most likely author of a text given a predefined set of candidate authors and a number of text samples per author of undisputed authorship [7, 8]. From a machine learning point of view, this task can be seen as a single-label multi-class text categorization problem [9] where the candidate authors play the role of the classes.

One major subtask of the authorship identification problem is the extraction of the most appropriate features for representing the style of an author, the so-called *stylometry*. Several measures have been proposed, including attempts to quantify vocabulary richness, function word frequencies and part-of-speech frequencies. A good review of stylometric techniques is given by Holmes [10]. The vast majority of

J. Euzenat and J. Domingue (Eds.): AIMSA 2006, LNAI 4183, pp. 77 – 86, 2006.
© Springer-Verlag Berlin Heidelberg 2006

proposed approaches are based on the fact that a text is a sequence of words. A promising alternative text representation technique for stylistic purposes makes use of character n-grams (contiguous characters of fixed length) [11, 12]. Character n-grams are able to capture complicated stylistic information on the lexical, syntactic, or structural level. For example, the most frequent character 3-grams of an English corpus indicate lexical (|the|[1], |_to|, |tha|), word-class (|ing|, |ed_|), or punctuation usage (|._T|, |_"T|) information. Character n-grams have been proved to be quite effective for author identification problems. Keselj et al. [12] tested this approach in various test collections of English, Greek, and Chinese text, improving previously reported results. Moreover, a variation of their method achieved the best results in the ad-hoc authorship attribution contest [13], a competition based on a collection of 13 text corpora in various languages (English, French, Latin, Dutch, and Serbian-Slavonic). The performance of the character n-gram approach was remarkable especially in cases with multiple candidate authors (>5).

Tokenization is not needed when extracting character n-grams, thus making the approach language independent. On the other hand, they considerably increase the dimensionality of the problem in comparison to word-based approaches. Due to this fact, n-grams of fixed length have been used so far (e.g. 3-grams). The selection of an optimal *n* depends on the language. Dimensionality reduction is of crucial importance, especially in case we aim to extract variable-length n-grams. That is, the combination of all 2-grams, 3-grams, 4-grams, etc. is much higher than the word-forms found in a text. Therefore when variable length n-grams are used, an aggressive feature selection method has to be employed to reduce the dimensionality of the feature space. To this end, traditional feature selection methods, such as information gain, chi square, mutual information etc. could be used. In general, these methods consider each feature independent of the others and attempt to measure their individual significance for class discrimination.

In this paper, we propose a feature selection method for variable-length n-grams based on a different view. The original idea is based on previous work for extracting multiword terms (word n-grams of variable length) from texts in the framework of information retrieval applications [14, 15]. According to the proposed approach, each feature is compared with other similar features of the feature set and the most important of them is kept. The factor that affects feature importance is its frequency of occurrence in the texts rather than its ability to distinguish between classes. Therefore, following the proposed method, we produce a feature subset which is quite different with the one produced by a traditional feature selection method. Experiments on a subset of the new Reuters corpus show that our approach is at least as effective as information gain for distinguishing among 50 authors when a large initial feature set is used while it is superior for small feature sets. Moreover, we examine a simple pre-processing procedure for removing redundancy in digits found in texts. It is shown that this procedure improves the performance of the proposed approach.

The rest of this paper is organized as follows. Section 2 presents our approach for n-gram feature selection. Section 3 presents the corpus used in the experiments and a baseline method. Section 4 includes the performed experiments while in section 5 the conclusions drawn by this study are summarized and future work directions are given.

[1] We use '|' and '_' to denote n-gram boundaries and a single space character, respectively.

2 N-Gram Feature Selection

The proposed method for variable-length n-gram feature selection is based on an existing approach for extracting multiword terms (i.e., word n-grams of variable length) from texts. The original approach aimed at information retrieval applications (Silva [15]). In this study, we slightly modified this approach in order to apply it to character n-grams for authorship identification. The main idea is to compare each n-gram with similar n-grams (either longer or shorter) and keep the dominant n-grams. Therefore, we need a function able to express the "glue" that sticks the characters together within an n-gram. For example, the "glue" of the n-gram |the_| will be higher than the "glue" of the n-gram |thea|.

2.1 Selecting the Dominant N-Grams

To extract the dominant character n-grams in a corpus we modified the algorithm *LocalMaxs* introduced in [15]. It is an algorithm that computes local maxima comparing each n-gram with similar n-grams. Given that:

- $g(C)$ is the *glue* of n-gram C, that is the power holding its characters together.
- *ant*(C) is an *antecedent* of an n-gram C, that is a shorter string having size n-1.
- *succ*(C) is a *successsor* of C, that is, a longer string of size n+1, i.e., having one extra character either on the left or right side of C.

Then, the dominant n-grams are selected according to the following rules:

$$\begin{aligned} &if\left(C.length > 3\right) \\ &g(C) \geq g(ant(C)) \wedge g(C) > g(succ(C)), \forall ant(C), succ(C) \\ &if\left(C.length = 3\right) \\ &g(C) > g(succ(C)), \forall succ(C) \end{aligned} \tag{1}$$

In the framework of authorship identification task, we only consider 3-grams, 4-grams, and 5-grams as candidate n-grams, since previous studies have shown they provide the best results [12]. As an alternative, we also consider words longer than 5 characters as candidate n-grams. Note that 3-grams are only compared with successor n-grams. Moreover, in case no words are used, 5-grams are only compared with antecedent n-grams. So, it is expected that the proposed algorithm will favor 3-grams and 5-grams against 4-grams.

2.2 Representing the Glue

To measure the glue holding the characters of a n-gram together various measures have been proposed, including specific mutual information [16], the φ^2 measure [17], etc. In this study, we adopt the *Symmetrical Conditional Probability* (SCP) proposed in [14]. The SCP of a bigram |xy| is the product of the conditional probabilities of each given the other:

$$SCP(x, y) = p(x \mid y) \cdot p(y \mid x) = \frac{p(x, y)}{p(x)} \cdot \frac{p(x, y)}{p(y)} = \frac{p(x, y)^2}{p(x) \cdot p(y)} \quad (2)$$

Given a character n-gram $|c_1 \ldots c_n|$, a *dispersion point* defines two subparts of the n-gram. A n-gram of length *n* contains *n*-1 possible dispersion points (e.g., if * denote a dispersion point, then the 3-gram |the| has two dispersion points: |t*he| and |th*e|). Then, the SCP of the n-gram $|c_1 \ldots c_n|$ given the dispersion point $|c_1 \ldots c_{n-1} * c_n|$ is:

$$SCP((c_1 \ldots c_{n-1}), c_n) = \frac{p(c_1 \ldots c_n)^2}{p(c_1 \ldots c_{n-1}) \cdot p(c_n)} \quad (3)$$

The SCP measure can be easily extended so that to account for any possible dispersion point (since this measure is based on fair dispersion point normalization, will be called *fairSCP*). Hence the *fairSCP* of the n-gram $|c_1 \ldots c_n|$ is as follows:

$$fairSCP(c_1 \ldots c_n) = \frac{p(c_1 \ldots c_n)^2}{\frac{1}{n-1} \sum_{i=1}^{i=n-1} p(c_1 \ldots c_i) \cdot p(c_{i+1} \ldots c_n)} \quad (4)$$

3 Experimental Settings

3.1 Corpus

In 2000, a large corpus for the English language, the Reuters Corpus Volume 1 (RCV1) including over 800,000 newswire stories, become available for research purposes. A natural application of this corpus is to be used as test bed for topic-based text categorization tasks [18] since each document has been manually classified into a series of topic codes (together with industry codes and region codes). There are four main topic classes: CCAT (corporate/industrial), ECAT (economics), GCAT (government/social), and MCAT (markets). Each of these main topics has many subtopics and a document may belong to a subset of these subtopics. Although, not particularly designed for evaluating author identification approaches, the RCV1 corpus contains 'by-lines' in many documents indicating authorship. In particular, there are 109,433 texts with indicated authorship and 2,361 different authors in total.

RCV1 texts are short (approximately 2KBytes – 8KBytes), so they resemble a real-world author identification task where only short text samples per author may be available. Moreover, all the texts belong to the same text genre (newswire stories), so the genre factor is reduced in distinguishing among the texts. On the other hand, there are many duplicates (exactly the same or plagiarized texts). The application of *R*-measure to the RCV1 text samples has revealed a list of 27,754 duplicates [19].

The RCV1 corpus has already been used in author identification experiments. In [19] the top 50 authors (with respect to total size of articles) were selected. Moreover, in the framework of the *AuthorID* project, the top 114 authors of RCV1 with at least 200 available text samples were selected [20]. In contrast to these approaches, in this study, the criterion for selecting the authors was the topic of the available text samples. Hence, after removing all duplicate texts found using the *R*-measure, the top

50 authors of texts labeled with at least one subtopic of the class CCAT (corporate/industrial) were selected. That way, it is attempted to minimize the topic factor in distinguishing among the texts. Therefore, since steps to reduce the impact of genre have been taken, it is to be hoped that authorship differences will be a more significant factor in differentiating the texts. Consequently, it is more difficult to distinguish among authors when all the text samples deal with similar topics rather than when some authors deal mainly with economics, others with foreign affairs etc. The training corpus consists of 2,500 texts (50 per author) and the test corpus includes other 2,500 texts (50 per author) non-overlapping with the training texts.

3.2 Information Gain as Baseline

Most traditional feature selection methods are information-theoretic functions attempting to measure the significance of each feature in distinguishing between the classes. In [21] the main feature selection methods are extensively tested in the framework of (topic-based) text categorization experiments. Among document frequency thresholding, information gain, mutual information, chi square, and term strength, the most effective methods were proved to be information gain and mutual information.

Information gain represents the entropy reduction given a certain feature, that is, the number of bits of information gained about the category by knowing the presence or absence of a term in a document:

$$IG(t_k, c_i) = \sum_{c \in \{c_i, \overline{c_i}\}} \sum_{t \in \{t_k, \overline{t_k}\}} P(t,c) \cdot \log \frac{P(t,c)}{P(t) \cdot P(c)} \tag{5}$$

Since information gain considers each feature independent of the others, it is not able to detect multiple redundant features that have the same ability to distinguish between classes. On the other hand, it offers a ranking of the features according to

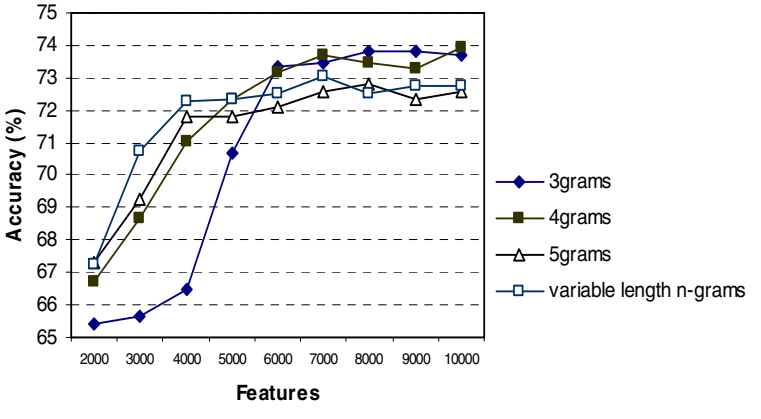

Fig. 1. Authorship identification results using information gain for feature selection

their information gain score, so a certain number of features can be easily selected. In this study, information gain was used as the baseline feature selection method. Any proposed method should have performance at least equal with the performance of information gain.

3.3 Author Identification Experiments

In each performed experiment, the following procedure was followed:

- An initial large feature set consisting of n-grams of variable length is extracted from the training corpus. This feature set includes the L most frequent n-grams for certain values of *n*. That is, for L=5,000, the 5,000 most frequent 3-grams, the 5,000 most frequent 4-grams, and the 5,000 most frequent 5-grams compose the initial feature set. In some cases, the most frequent long words (length>5) are also added to the initial feature set.
- A feature selection method is applied to this large feature set.
- A Support Vector Machine (SVM) is trained using the reduced feature set. In all experiments, linear kernels are used with C=1.
- The SVM model is applied to the test set and the microaverage accuracy is calculated.

4 Results

The first experiment was based on information gain measure to select the most significant features. Given an initial feature set of 15,000 features (including the 5,000 most frequent 3-grams, the 5,000 most frequent 4-grams, and the 5,000 most frequent 5-grams) information gain was used to extract the best 2,000 to 10,000 features with a step of 1,000. For comparative purposes, we also used information gain to select fixed-length n-grams. Hence, using as initial feature set the 15,000 most frequent 3-grams, information gain was used to select the best 2,000 to 10,000 features with a step of 1,000. The same approach was followed for 4-grams and 5-grams. The results are shown in Figure 1. As can be seen, the variable-length n-grams outperform fixed-length n-grams for relatively low dimensionality (when less than 5,000 features are selected). However, when the dimensionality arises, the variable-length n-grams selected by information gain fail to compete with fixed-length n-grams (especially, 3-grams and 4-grams). Moreover, in all cases the performance of the model is not significantly improved beyond a certain amount of selected features.

In the second experiment, we applied the proposed method to the same problem. Recall that our method is not able to select a predefined number of features since it does not provide feature ranking. However, the number of selected features depends on the size of the initial feature set. So, different initial feature sets were used, including 6,000 to 24,000 variable-length n-grams, equally distributed among 3-grams, 4-grams, and 5-grams. Moreover, we also performed experiments using words longer than 5 characters as candidate n-grams. The results of these experiments are depicted in Figure 2. Note that the results of this figure are not directly comparable with the results of figure 1 since different initial feature sets were used. When using exactly the same initial feature set with information gain, the accuracy based on our

method reaches 73.08%. It can be seen that the proposed method can produce much more accurate classifiers in comparison with information gain when using a low number of features. In addition, these reduced feature sets were selected from a more restricted initial feature set. For example, when the initial feature set comprises 6,000 variable-length n-grams, our method selects 2,314 features producing an accuracy of 72%. Recall that a feature set of 4,000 variable length n-grams (selected out of 15,000 n-grams) produced by information gain reaches accuracy of 72%. On the other hand, the addition of long words to the feature set does not seem to significantly contribute to the classification accuracy.

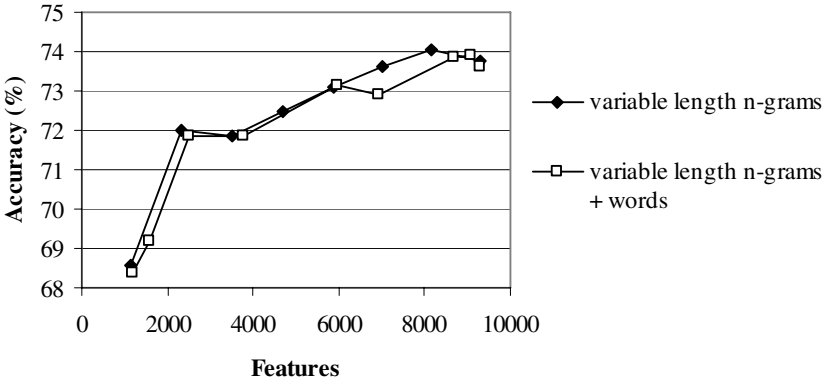

Fig. 2. Results of the proposed method using only variable-length n-grams and variable-length n-grams plus words longer than 5 characters

Table 1. Comparison of the feature sets produced by information gain (IG) and the proposed method (PM) in terms of common members (CM) for three cases

	IG	**PM**	**CM**	**IG**	**PM**	**CM**	**IG**	**PM**	**CM**
3-grams	647	1337	127	647	2,938	317	851	5,510	530
4-grams	909	423	161	2,228	705	462	2,327	1,012	510
5-grams	758	554	131	1,816	1,048	315	5,000	1,656	1,257
Total	2,314	2,314	419	4,691	4,691	1094	8,178	8,178	2,297
Accuracy	69.4%	72.00%		72.16%	72.48%		72.56%	74.04%	

A closer look at the feature sets produced by information gain and the proposed method will reveal their properties. To this end, table 1 presents the distribution in 3-grams, 4-grams, and 5-grams of the best features produced by information gain and the proposed method, respectively, as well as the amount of common members of the two sets. Three cases are shown corresponding to 2,314, 4,691, and 8,178 best features selected by the two methods. As can be seen, information gain favors 4-grams and especially 5-grams for large feature sets while for small feature sets the selected features are (roughly) equally distributed. On the other hand, the proposed method mainly favors 3-grams in all cases, followed by 5-grams. Interestingly, the common members of the two datasets are only a few. For example, in case of 2,314

best features, the proposed method selected 1,337 3-grams and the information gain selected 647 3-grams. However, the intersection of the two sets consists of 127 3-grams only. This indicates that the examined methods focus on different kinds of information when selecting the features. Indeed, information gain will select all the n-grams |and|, |and_|, |_and|, |_and_| given that the use of word 'and' is important for distinguishing between the authors. Thus, the reduced feature set will contain redundant features. On the other hand, the proposed method will select at least one of these n-grams. Hence, when equal number of features is selected by the two methods, the feature set of the proposed method will be richer in different n-grams corresponding to different kind of stylistic information.

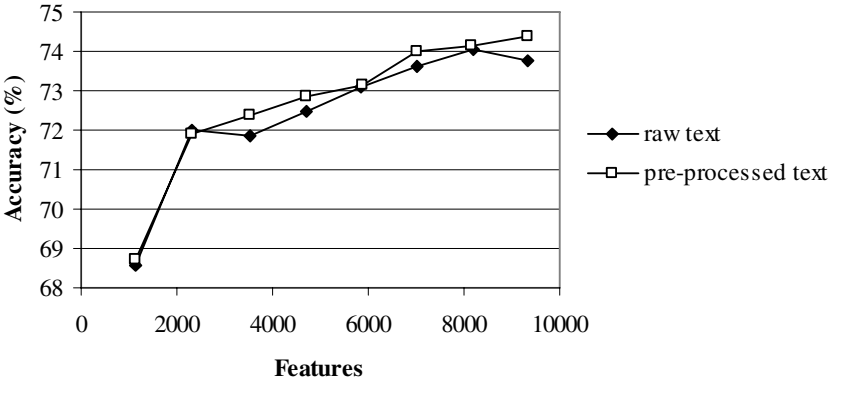

Fig. 3. Performance results using raw text and pre-processed text where all digit characters were replaced by the same symbol

4.1 Text Pre-processing

The experiments we have presented so far were conducted on raw text. No pre-processing of the text was performed apart from removing XML tags irrelevant to the text itself. However, simple text pre-processing may have a considerable impact in the framework of text categorization tasks [22]. In this study, we emphasize on pre-processing texts for removing redundancy of digit characters.

The information represented by digits may correspond to dates, values, telephone numbers etc. The use of digits is mainly associated with text-genre (press reportage, press editorial, official documents, etc.) rather than authorship. Given a character n-gram text representation, multiple digit-based n-grams will be extracted from a text. In many cases, the important stylistic information is the use of digits rather than the exact combinations of digits. Hence, if all digits are replaced with a special symbol (e.g., '@'), the redundancy in character n-grams would be much lower. For example, all |1999|, |2000|, |2001|, and |2002| 4-grams would be replaced by |@@@@|. Frequent use of this transformed 4-gram could be due to frequent reference to dates.

We examine the effect of this simple pre-processing procedure on the authorship identification task. Figure 3 depicts the classification accuracy results using the proposed feature selection method on variable-length n-grams extracted from raw text

(as previously) and pre-processed text (with digit characters replaced by a symbol). The amount of features selected based on the pre-processed text is slightly smaller. More importantly, the performance of the model based on pre-processed text is better especially when using more than 2,000 features. This indicates that simple text transformations can yield considerable improvement in accuracy.

5 Discussion

We presented a new approach for feature selection aimed at authorship identification based on character n-gram text representation. The proposed method is able to select variable length n-grams based on a technique originally applied for extracting multiword expressions from text. The key difference with traditional feature selection methods is that the significance of a feature is measured in comparison with other similar features rather than its individual ability to discriminate between the classes. Therefore, the produced feature set is stylistically richer since it contains the dominant character n-grams and is less likely to be biased by some powerful n-grams that essentially represent the same stylistic information.

Another difference with traditional feature selection approaches is that there is no ranking of the features according to their significance. Essentially, it is not possible to select a predefined number of features. Although this fact complicates the experimental comparison with other approaches, it is not of crucial importance for the practical application of the proposed method to real-world cases.

We also presented experiments about the significance of digits in the framework of author identification tasks. The removal of redundancy in digit characters improves classification accuracy when a character n-gram text representation is used. Furthermore, the cost of this procedure is trivial. It remains to be tested whether alternative text transformations are useful as well.

In this study, we restricted our method to certain n-gram types (3-grams, 4-grams, and 5-grams). To keep dimensionality on low level, we used words longer than 5 characters as an alternative for longer n-grams. However, the results when using the additional words were not encouraging. It would be interesting for one to explore the full use of long n-grams as well as the distribution of selected n-grams into different n-gram lengths especially when texts from different natural languages are tested.

References

1. Mosteller, F., Wallace, D.: Inference in an Authorship Problem. Journal of the American Statistical Association, 58:302 (1963) 275-30.
2. Labbé, C., Labbé, D.: Inter-textual distance and authorship attribution: Corneille and Molière. Journal of Quantitative Linguistics, 8 (2001) 213-31.
3. de Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining E-mail Content for Author Identification Forensics. *SIGMOD Record*, 30:4 (2001) 55-64.
4. Abbasi, A., Chen, H.: Applying Authorship Analysis to Extremist-Group Web Forum Messages. IEEE Intelligent Systems, 20:5 (2005) 67-75.
5. van Halteren, H.: Linguistic Profiling for Author Recognition and Verification. In Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (2004) 199-206.

6. Chaski, C.: Empirical Evaluations of Language-based Author Identification Techniques. Forensic Linguistics, 8:1 (2001) 1-65.
7. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic Text Categorization in Terms of Genre and Author. Computational Linguistics 26:4 (2000) 471-495.
8. Peng, F., Shuurmans, F., Keselj, V., Wang, S.: Language Independent Authorship Attribution Using Character Level Language Models. In Proc. of the 10th European Association for Computational Linguistics (2003).
9. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34:1 (2002) 1-47.
10. Holmes, D.: The Evolution of Stylometry in Humanities Scholarship. Literary and Linguistic Computing, 13:3 (1998) 111-117.
11. Kjell, B., Addison Woods, W, Frieder O.: Discrimination of authorship using visualization. Information Processing and Management 30:1 (1994).
12. Keselj, V., Peng, F., Cercone, N. Thomas, C.: N-gram-based Author Profiles for Authorship Attribution. In Proc. of the Conference Pacific Association for Computational Linguistics (2003).
13. Juola, P.: Ad-hoc Authorship Attribution Competition. In Proc. of the Joint ALLC/ACH2004 Conf. (2004) 175-176.
14. Silva, J., Dias, G., Guilloré S., Lopes, G.: Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. LNAI, 1695 (1999) 113-132.
15. Silva, J., Lopes, G.: A local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units. In Proc. of the 6th Meeting on the Mathematics of Language (1999) 369-381.
16. Church K., Hanks K.: Word Association Norms, Mutual Information and Lexicography. Computational Linguistics, 16:1 (1990) 22-29.
17. Gale W., Church K.: Concordance for parallel texts. In Proc. of the 7th Annual Conference for the new OED and Text Research, Oxford (1991) 40-62.
18. Lewis, D., Yang, Y., Rose, T., Li, F.: RCV1: A New Benchmark Collection for Text Categorization Research. Journal of Machine Learning Research, 5 (2004) 361-397.
19. Khmelev, D. Teahan, W.: A Repetition Based Measure for Verification of Text Collections and for Text Categorization. In Proc. of the 26th ACM SIGIR (2003) 104–110.
20. Madigan, D., Genkin, A., Lewis, D., Argamon, S., Fradkin, D., Ye, L.: Author Identification on the Large Scale. In Proc. of CSNA (2005).
21. Yang, Y., Pedersen J.: A Comparative Study on Feature Selection in Text Categorization. In Proc. of the 14th Int. Conf. on Machine Learning (1997).
22. Marton, Y., Wu, N., Hellerstein, L.: On Compression-Based Text Classification. Advances in Information Retrieval: 27th European Conference on IR Research, Springer LNCS – 3408, pp. 300-314 (2005).