AUSTRIAN JOURNAL OF STATISTICS
Volume 37 (2008), Number 1, 109–118

Classification of Publications Based on Statistical Analysis of Scientific Terms Distributions

Vaidas Balys and Rimantas Rudzkis
Institute of Mathematics and Informatics, Vilnius

Abstract: The problem of classification of scientific texts is considered. Models and methods based on probabilistic distributions of scientific terms in text are discussed. The comparative study of proposed and a few of popular alternative algorithms was performed. The results of experimental study over real-world data are reported.

Keywords: Probability Distribution, Statistical Classification.

1 Introduction

The researches in the fields of text processing, knowledge discovery, and management face increasing attention in the light of ongoing changes of information finding and acquisition models. Classification of a chunk of text is arguably one of the most often addressed problems in the field. Currently, there is a reasonable list of plain text classification (actually, text categorization is the more popular name for the problem) algorithms that gained high popularity over years. See Sebastiani (2002) for a nice introduction into some of them.

Scientific texts are rather different from everyday-language texts in a number of aspects therefore there is a need for studies that would address these differences and eventually develop methods and algorithms suitable to deal with this specific content. Scientific texts usually come in form of publications that have a strict format – they have titles, abstracts, and then full texts that are highly structured. Titles and abstracts are usually freely available and they present a concise summary of the content of publications therefore it is natural to use them for classification. On the other hand, full texts provide information that is not present anywhere else and there is a question if and to what extent this could be used. The language of (mathematical) publications is rather specific filled with a lot of scientific terms together with such multidimensional elements like formulae. The natural idea that scientific terms should be the most informative for correct classification is one of the key assumptions of our research.

There are two popular ways to classify a scientific publication. This can be done by assigning keywords or by assigning labels from standard classification scheme, like MSC (Mathematical Subject Classification) for mathematics. MSC elements are classifiers by their nature, therefore it is natural to test algorithms on them. Keywords are different as they have no strictly limited vocabulary and authors tend to assign them basing on their own likings. Even though keywords could be treated the same as MSC classifiers, the more desirable would be to develop models covering the aspects related to the specificity of keywords. However, this question is not covered in this study.

The paper deals with classification of scientific papers, or more precisely – with classification of mathematical papers from the field of probability theory and mathematical statistics.

2 The Model

2.1 Definitions

Here, a rather brief introduction of stochastic terms distribution models (that mathematically define the concept of identification cloud, Hazewinkel, 2004) based approach to the problem of classification is presented. For more comprehensive review see Rudzkis, Balys, and Hazewinkel (2006).

Let K denote some classification system of scientific texts which is identified with a set of all possible labels of the classes in that system.

Let V be a vocabulary, i.e., set of scientific terms of a certain scientific field that are relevant to the classification of texts. The chronologically numerated vector of article's a elements $(a_1, \dots, a_d)$, $d = d(a)$, where $a_i \in V$ and not necessarily $a_i \neq a_j$, is called the projection of the article a. Sometimes it is convenient to identify the projection of an article a with an infinite sequence $(a_1, a_2, \dots)$, where $a_i = 0$ for all $i > d(a)$. Here $0 \in V$ denotes an additional zero term which does not exist in reality. Let A be a set of projections of all articles from a certain scientific field. In what follows the word "projection" is omitted and $a \in A$ is called just an article.

From the point of view of classification an article is not necessarily a homogenous piece of text – in the general case, it consists of $q = q(a) \geq 1$ continuous homogenous parts which are classified as different in system K. Non-intersecting intervals of indices $I_j(a) \subset \{1, \dots, d(a)\}$ and class labels $w_j(a) \in K, j = \overline{1,q}$ correspond to these parts. Here $\bigcup_{j=1}^{q} I_j(a) = N(a)$ and $w_j \neq w_{j-1}, j = \overline{2,q}$: if two adjacent parts of the text are attributed to the same class they are joined into one.

Let N be the set of natural numbers. Let an article $a \in A$ and a set of indices $I \subset N$ be chosen randomly so that the part of an article $\{(a_\tau, \tau), \tau \in I\}$ is homogenous: $I \subset I_\nu(a), \nu \in \{1, \dots, q\}$. This part is attributed to the class $\eta = w_\nu(a)$ in the system K. A common problem of classification is to determine the unknown class η using the observed vector $a_I = (a_\tau, \tau \in I)$.

2.2 Probability Distributions

Since (a, I, η) is the result of a random experiment, the probability distribution in the set K is defined by

$$Q(w) = \mathbb{P}\{\eta = w\}, \qquad w \in K. \tag{1}$$

Let Y be a set of all possible values of a_I. In the set Y the following conditional probability distributions are defined:

$$P(y) = \mathbb{P}\big\{a_I = y \big| \, |I| = d(y)\big\},$$

$$P(y|w) = \mathbb{P}\{a_I = y \big| \, |I| = d(y), \eta = w\}, \qquad w \in K, \tag{2}$$

where $d(y) = \dim(y)$, $|I| = \mathrm{card}(I)$.

If η and $|I|$ are independent, after observing a_I, the posteriori probability of the random event $\{\eta = w\}$ is determined by $Q(w|a_I) = Q(w) \cdot \psi_w(a_I)$, where

$$\psi_w(y) = P(y|w)/P(y), \qquad y \in Y. \tag{3}$$

The concept of identification cloud may be defined by the functional ψ_w that reflects how the probability to observe some text changes if this text appears to be classified under certain class.

Using the distributions, described in Equations (1) and (2), Bayes classifier which minimizes mean classification losses can be defined. If the loss function is trivial, i.e., it equals to some constant in case of misclassification, it is simply the maximum posteriori classifier:

$$\widehat{\eta} = \arg\max_{w \in K} P(a_I|w)Q(w) \tag{4}$$

in which $\psi_{(\cdot)}(a_I)$ can be substituted for $P(a_I|\cdot)$:

$$\widehat{\eta} = \arg\max_{w \in K} \psi_w(a_I)Q(w). \tag{5}$$

2.3 Inference

In order to use classification method (5) the distribution Q and the functional ψ must be estimated. Below the statistical estimation methods are presented.

Let us have the learning sample of observed parts of texts and their classification results $X = (y(1), \eta(1)), \dots, (y(n), \eta(n))$, where $\eta(i) \in K$, $y(i) \in Y$, $Y = \{y = (y_1, \dots, y_d) : y_i \in V, d \in N\}$.

Nonparametric Estimation. The empirical analogue of $Q(w)$ is d etermined by $\widehat{Q}(w) = \sum_{j=1}^{n} 1_{\{\eta(j)=w\}}/n$.

The functional $\psi_w(y)$ is estimated by using a common k-nearest neighbors method. Let for all $y, z \in Y$, $\rho(y, z)$ be a non-negative functional which is called a pseudo-distance from element z to element y. For a fixed $y \in Y$, one can choose k “nearest neighbours” from the sample. Let $J(y) \subset \{1, \dots, n\}$ be a set of k indices of observations $y(1), \dots, y(n)$, for which the pseudo-distance to y are the smallest ones. The estimate of $\psi_w(y)$ is then determined by $\widehat{\psi}_w(y) = \sum_{j \in J(y)} 1_{\{\eta(j)=w\}}/\widehat{Q}(w)k$. Here $0/0 = 1$. The variable $k = k(n)$ depends on the size of the sample and conditions $k \to \infty, k/n \to 0$, as $n \to \infty$ hold.

There is a long list of pseudo-distance functions that could be used, the most popular being Euclidian distance and cosine similarity functions. In Rudzkis et al. (2006) one which takes into consideration both the frequencies and positions of scientific terms in text is proposed.

Parametric Estimation. For parametric estimation additional definitions are needed. Let the index $\tau \in I$ be a random variable. The distribution on set V is defined by $P(v) = \mathbb{P}\{a_\tau = v\}$ and the corresponding conditional distribution is given by $P(v|w) = \mathbb{P}\{a_\tau = v|\eta = w\}$, $w \in K$.

The two following assumptions substantially simplify the procedures of estimation.

Assumption 1 (conditional stationarity and independence). Let for all $y \in Y$ and $w \in K$ hold

$$P(y|w) = \prod_{i=1}^{d} P(y_i|w),$$

where $d = d(y)$ as before is the dimension of the vector y.

Now the definition of the identification cloud (3) can be changed to

$$\psi_w(v) = P(v|w)/P(v), \qquad v \in V, w \in K, \tag{6}$$

while the Bayes classification rule for classifying the observed a_I is determined now by

$$\widehat{\eta} = \arg\max_{w \in K} \left[Q(w) \prod_{\tau \in I} \psi_w(a_\tau) \right] . \tag{7}$$

The definition of the identification cloud (6) based on this assumption ignores information that can be derived from the order of the terms in the text. Thus, we introduce a weaker assumption.

Assumption 2 (conditional stationarity and Markovian property). Let for all $y \in Y$ and $w \in K$ hold

$$P(y|w) = P(y_1|w) \prod_{i=1}^{d-1} \left[P(y_i, y_{i+1}|w)/P(y_i|w) \right] , \tag{8}$$

where $P(v,u|w) = \mathbb{P}\{a_\tau = v, a_{\tau+1} = u | \eta = w\}$.

In this case, the identification cloud is described by two functionals: $\psi_w(v)$ defined in (6) and

$$\psi_w(v,u) = P(v,u|w)/P(v,u), \qquad v, u \in V, \tag{9}$$

where $P(v,u) = \mathbb{P}\{a_\tau = v, a_{\tau+1} = u\}$.

Let $I = \{r, r+1, \ldots, m\}$. Then the Bayes rule of classification is obtained by modifying the Equation (4) according to the Equations (6), (8) and (9):

$$\widehat{\eta} = \arg\max_{w \in K} \left[Q(w) \psi_w(a_r) \prod_{i=r}^{m-1} \left[\psi_w(a_i, a_{i+1})/\psi_w(a_i) \right] \right] . \tag{10}$$

In order to use the algorithms (7) and (10) the functionals $\psi_w(v)$ and $\psi_w(v,u)$ have to be estimated. Here we propose one of the simpliest ways to do that (see Rudzkis et al., 2006 for full description). First the empirical estimates of the probabilities $P(\cdot)$, $P(\cdot,\cdot)$, $P(\cdot|\cdot)$ and $P(\cdot,\cdot|\cdot)$ are calculated by substituting them with corresponding frequencies. Then these estimates are used in (6) and (9) thus yielding empirical identification clouds $\widetilde{\psi}_w(v)$ and $\widetilde{\psi}_w(v,u)$. The smoothing is performed – the unreliable estimates, i.e., those that are based on too few observations, are modified.

Let $h = |V|$. The functionals $\widetilde{\psi}_w(\cdot)$ and $\widetilde{\psi}_w(\cdot,\cdot)$ determine the arrangements of set V for every $w \in K$ and every pair $(w, v),\ v \in V$:

$$\widetilde{\psi}_w(v_1) \geq \widetilde{\psi}_w(v_2) \geq \ldots \geq \widetilde{\psi}_w(v_h), \qquad v_{(\cdot)} \in V,$$

$$\widetilde{\psi}_w(v,u_1) \geq \widetilde{\psi}_w(v,u_2) \geq \ldots \geq \widetilde{\psi}_w(v,u_h), \qquad u_{(\cdot)} \in V.$$

These arrangements are used as following: a fixed number of highest and a fixed number of lowest $\widetilde{\psi}_w(v)$ values for each w are left unchanged while others are declared "uninformative":

$$\widehat{\psi}_w(v_k) = \begin{cases} 1, & \text{if } s < k < h - l, \\ \widetilde{\psi}_w(v_k), & \text{otherwise.} \end{cases}$$

For $\widetilde{\psi}_w(u,v)$ the same procedure is applied for each fixed pair w and v separately.

In Rudzkis et al. (2006) a procedure for selecting s and l which is based on hypothesis testing theory is proposed. There is also a way to choose optimal values by running experiments on data.

The remaining "informative" parts of identification clouds then can be fitted to some parametric model (see Rudzkis et al., 2006 for one of them) thus yielding the final identification clouds that can be used when applying classification algorithms (7) and (10). As of now the identification cloud of a class $w \in K$ is understood as a list of scientific terms (and pairs of terms) having $\psi_w(\cdot)$ and $\psi_w(\cdot,\cdot)$ values different from 1.

3 Experimental Evaluation

3.1 The Data

The experiments were conducted on basis of almost 15000 articles from the field of probability theory and mathematical statistics kindly provided by the Institute of Mathematical Statistics, USA. 44 MSC classifiers (24 from 60XXX subtree and 20 from 62XXX subtree) were chosen for the experiments each having a learning set of at least 100 articles, thus resulting in a total of 5337 articles. The statistics of number of classifiers (they are also called *categories* in the following) assigned to the articles is presented in Table 1.

Table 1: Statistics of number of MSC classifiers assigned.

Number of Classifiers	1	2	3	4	5	6	7
Number of Articles	2617	1905	638	161	15	0	1
Percentage of Articles	49.04	35.69	11.95	3.02	0.28	0	0.02

The dictionary of scientific terms was constructed by extracting all the keywords from the articles in the database. The single words that build up keywords-phrases were also added. This resulted in a list of 17632 unique terms. 9587 of them are found in the full texts, 4506 in the abstracts and 2770 in the titles of the chosen subset of articles. These terms cover approximately one fourth of all the words found in texts and it is yet to be verified what impact would make using a larger dictionary of terms or even all the words from the texts.

The terms are not only single words – see Table 2 for statistics of terms' length (measured in words). The effect of adding phrases to the dictionary was estimated.

Three types of texts were available: titles, abstracts and full texts. In Table 3, basic statistical information on the length (measured in number of scientific terms) of these parts is presented.

3.2 The Algorithms

All the considered algorithms implement the so-called supervised learning approach, i.e., learning over pre-labelled corpus. The algorithms (with exception of *kNN*) perform by

Table 2: Statistics of scientific terms' length.

Term's Length	1	2	3	4	5
Number of Terms	4231	4818	525	12	1

Table 3: Statistics of text parts' length (measured in number of scientific terms.

	max	average	median
Title	10	3	3
Abstract	83	19	17
Full Text	3441	513	433

analyzing the positive and the negative examples of classes and building discriminative rules so that they classify learning data as correct as possible. The algorithms support the ranking procedure: for each document a list of categories that could be assigned to the document with corresponding weights is delivered. Then, depending on some thresholding strategy, a subset of categories with the highest weights is chosen.

The considered algorithms include: *IDC* and *IDCM* – identification clouds algorithms that use independence assumption (see Equation (7)) and Markovian assumption (see Equation (10)); *nB (naive Bayes with additive smoothing)* – a simple algorithm that builds on the assumption of word independence over the text and makes use of Bayes rule to compute scores for categories (see Mitchell, 1996); *kNN* – a common instance-based k nearest neighbors algorithm (see Yang, 1994) that skips the phase of learning and makes decisions by analyzing true decisions of documents closest to the one to be classified; *SVM* – a popular Support Vector Machines (see Vapnik, 1995, Joachims, 1998) method that tries to separate documents of different classes by the widest margin. The specific implementation LIBSVM (Chang and Lin, 2001) was used.

3.3 Performance Measures

The common k-fold cross validation procedure with $k = 5$ was used to evaluate the algorithms.

The common measures *precision* (P) and *recall* (R) see e.g. Yang (1999)) were used to compare the true and the guessed classification. The precision is the proportion of guessed categories that are truly assigned to the document while the recall is the proportion of truly assigned categories that are matched by the guessed ones.

To evaluate the nature of tradeoff between precision and recall a number of measures could be used (e.g. precision-recall graph). In this research a single-valued measure similar to 11-point average precision (Yang, 1999) was used. For each document having m categories precision values at points of recall increases are calculated. Recall increases when subsequently added category from the suggested ranked list coincides with one of truly assigned. At this point the precision is equal to a number of matched categories divided by a number of steps taken. The average precision for a document is equal to the average of these precision values: $P_{avg}(d) = (1/n_1 + 2/n_2 + \cdots + m/n_m)/m$ where d is text considered (part of some article) and n_i is the number of steps taken until i true categories were matched. The average over all documents gives the measure of algorithm's efficiency, denoted by P_{avg} and called *average precision*.

3.4 Results

The table 4 presents the estimated efficiency of algorithms (measured by averaged precision P_{avg}) over various combinations of learning and testing sets. Each algorithm achieved it's best result with some optimal parameters that were determined by running experiments and using k-fold cross validation.

DF (document frequency) method was used to exclude non-informative terms as according to Yang and Pedersen (1997) it is among the best for non-aggressive feature space dimensionality reduction. DF values ranging from 3 to 6 were found to be optimal – the dimension was reduced by a factor of 2–3 while efficiency of algorithms increased.

Adding phrases to the dictionary of single-worded scientific terms made a positive influence on results as one could naturally expect. The increase of about 10% of P_{avg} was observed for all methods.

kNN algorithm ($k \approx 30$) performed the worst while *SVM* came out as a winner by a slight margin (see Table 4). *IDC* algorithm performed the best when identification clouds included only those terms that make positive influence to the identification of class, i.e. those having $\psi_w(\cdot) > 1$. *IDCM* algorithm performed a bit better than *IDC* but the difference generally was too small to compensate for the added complexity. The main reason seems to be that the overwhelmingly big part of the pairs of terms that appear in the texts of articles are observed too rare to analyze their distributional patterns. The most obvious solution to this problem would be to use much bigger datasets. The another way is to substitute strict Markovian condition "one term next to the other term" with more loose one like "one term within some distance from the other term". Preliminary analysis shows that this modification improves results but exhaustive experiments are needed to refine this heuristic idea and estimate the gain.

Table 4: P_{avg} of algorithms for various learn / test set combinations.

	nB	IDC	IDCM	kNN	SVM
Title / Title	0.513	0.420	0.420	0.490	0.506
Abstract / Abstract	0.589	0.575	0.574	0.544	0.594
Abstract / Text	0.603	0.607	0.607	0.561	0.601
Text / Abstract	0.620	0.606	0.610	0.568	0.629
Text / Text	0.659	0.630	0.637	0.589	0.667

From the Table 4 it is evident that using full texts yields improvement in algorithms' performance. What is not evident from this table is that not all the text is useful. The Figure 1 shows how the averaged precision P_{avg} changes when more and more of text (the length of text is measured in number of scientific terms) is used for both learning and testing. The similar picture could be also seen in abstract / text setting.

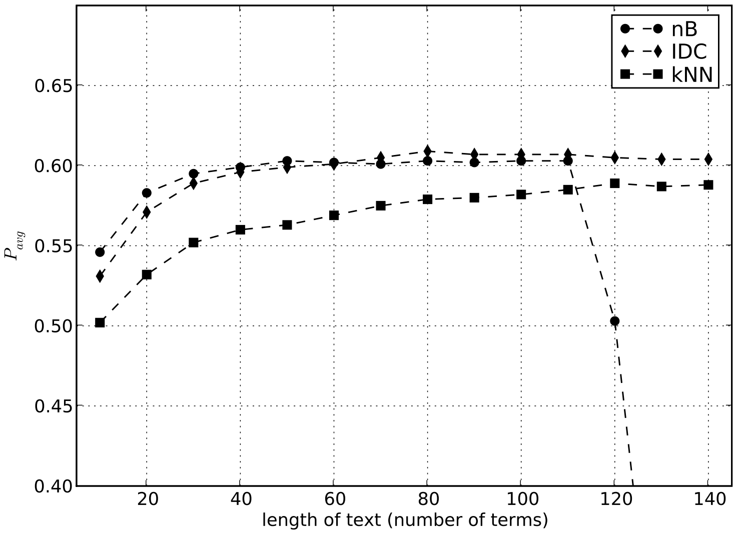

Figure 1: Influence of text length on the performance.

The first 100 – 110 terms of text are useful as the average precision increases. 60 of them which is three times the average length of abstracts are enough to reach performance values near to the maximum. Starting from approximately 120 terms the performance of *nB* drastically decrease while others stay at a stable level. That decrease could probably be explained by the nature of scientific publications: at some point the introductory parts are finished and real mathematical content takes the place which is specific, filled with a lot of irrelevant and unseen terms.

The *IDC* algorithm could be seen as a modification of *nB*. For *nB* all the terms are important while *IDC* picks only a fraction of them (the identification cloud) for each class The Figure 2 shows the tradeoff between size of identification cloud and the algorithm's performance. The size of identification cloud was fixed at various values ranging from 10 to 1000. The ratio of averaged precision P_{avg} for *IDC* and *nB* is depicted on the graph ('text100' stands for the subset of full text containing first 100 scientific terms).

It is evident that size of identification cloud of 400 – 600 (which is less than 10% of the size of the list of terms found in texts) is enough for the performance value to reach 95% of the performance value of *nB*.

Figure 3 shows the values of precision and recall for all the algorithms for the setting when learning and testing is performed on the full texts including first 100 scientific terms. The simple thresholding strategy where a fixed number of highest ranked categories (from 1 to 5) is assigned to the document was implemented.

4 Conclusions

The conducted experiments confirmed a natural assumption that using full texts of articles instead of only abstracts improve performance of classification algorithms. However, only a limited portion of the text from the beginning is useful as starting from some point

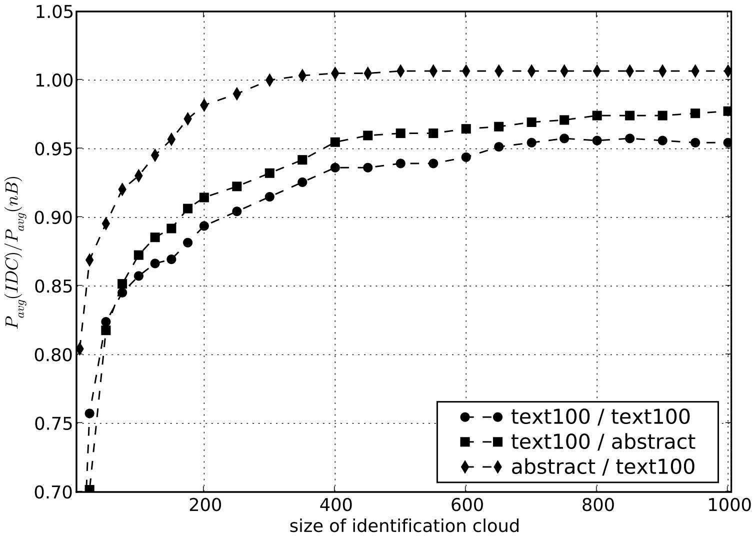

Figure 2: $p_{AVG}(IDC)/p_{AVG}(nB)$ for various settings.

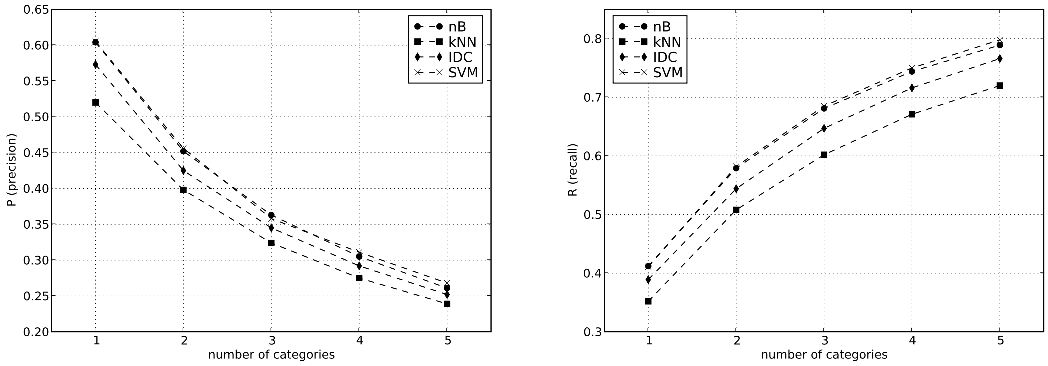

Figure 3: Precision and recall for text100 / text100 setting.

no significant improvement is registered or even drop of performance can be observed. Adding phrases to the list of single-worded terms improves performance by about 10%.

All the considered algorithms show comparable results with *SVM* outperforming others by a slight margin, however it is hampered by it's computational overhead. *Naive Bayes* is much simpler and it demonstrates results similar to that of *SVM*. *k nearest neighbors* ended being the worst of the considered methods. Identification clouds based algorithms with a limited size of clouds reach the adequate performance as compared to the *naive Bayes*. Only insignificant improvement of using Markovian assumption instead of independence assumptions was observed. However, certain modifications to this assumption promise better results. These modifications together with combined methods are to be studied over the richer data bases in the nearest future.

References

Chang, C., and Lin, C. (2001). LIBSVM: a library for support vector machines. (Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm)

Hazewinkel, M. (2004). Dynamic stochastic models for indexes and thesauri, identification clouds, and information retrieval and storage. In R. Baeza-Yates (Ed.), *Recent Advances in Applied Probability* (p. 181-204).

Joachims, T. (1998). Text categorization with suport vector machines: Learning with many relevant features. In *ECML '98: Proceedings of the 10th European Conference on Machine Learning* (pp. 137–142). New York: Springer-Verlag.

Mitchell, T. M. (1996). *Machine Learning*. McGraw-Hill.

Rudzkis, R., Balys, V., and Hazewinkel, M. (2006). Stochastic modelling of scientific terms distribution in publications. In J. M. Borwin and W. M. Farmer (Eds.), *Mathematical Knowledge Management* (Vol. 4108, p. 152-164). Springer Berlin / Heidelberg.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, *34*, 1-47.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Yang, Y. (1994). Expert network: effective and efficient learning from human decisions in text categorization and retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and Development in information retrieval* (pp. 13–22). New York: Springer-Verlag.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval*, *1*, 69-90.

Yang, Y., and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *ICML '97: Proceedings of the 14th International Conference on Machine Learning* (pp. 412–420). Morgan Kaufmann Publishers Inc.

Corresponding Author's Address:

Rimantas Rudzkis
Probability Theory and Statistics Department
Institute of Mathematics and Informatics
Akademijos str. 4
LT-08663 Vilnius
Lithuania

E-mail: `rudzkis@ktl.mii.lt`