VILNIUS GEDIMINAS TECHNICAL UNIVERSITY
INSTITUTE OF MATHEMATICS AND INFORMATICS

Vaidas BALYS

MATHEMATICAL MODELS FOR SCIENTIFIC TERMINOLOGY AND THEIR APPLICATIONS IN THE CLASSIFICATION OF PUBLICATIONS

SUMMARY OF DOCTORAL DISSERTATION

PHYSICAL SCIENCES,
MATHEMATICS (01P)

VGTU
LEIDYKLA
VILNIUS TECHNIKA 2009

Doctoral dissertation was prepared at the Institute of Mathematics and Informatics in 2004–2009.

Scientific Supervisor

Prof Dr Habil Rimantas RUDZKIS (Institute of Mathematics and Informatics, Physical Sciences, Mathematics – 01P).

The dissertation is being defended at the Council of Scientific Field of Mathematics at Vilnius Gediminas Technical University:

Chairman

Prof Dr Habil Kęstutis KUBILIUS (Institute of Mathematics and Informatics, Physical Sciences, Mathematics – 01P).

Members:

Prof Dr Habil Remigijus LEIPUS (Vilnius University, Physical Sciences, Mathematics – 01P),

Prof Dr Valentinas PODVEZKO (Vilnius Gediminas Technical University, Social Sciences, Economics – 04S),

Prof Dr Habil Alfredas RAČKAUSKAS (Vilnius University, Physical Sciences, Mathematics – 01P),

Prof Dr Habil Leonas SAULIS (Vilnius Gediminas Technical University, Physical Sciences, Mathematics – 01P).

Opponents:

Prof Dr Kęstutis DUČINSKAS (Klaipėda University, Physical Sciences, Mathematics – 01P),

Assoc Prof Dr Marijus RADAVIČIUS (Institute of Mathematics and Informatics, Physical Sciences, Mathematics – 01P).

The dissertation will be defended at the public meeting of the Council of Scientific Field of Mathematics at the Institute of Mathematics and Informatics, Room 203 at 1 p. m. on 2 October 2009.

Address: Akademijos g. 4, LT-08663 Vilnius, Lithuania.
Tel.: +370 5 274 4952, +370 5 274 4956; fax +370 5 270 0112;
e-mail: doktor@adm.vgtu.lt

The summary of the doctoral dissertation was distributed on 1 September 2009.

A copy of the doctoral dissertation is available for review at the Libraries of Vilnius Gediminas Technical University (Saulėtekio al. 14, LT-10223 Vilnius, Lithuania) and the Institute of Mathematics and Informatics (Akademijos g. 4, LT-08663 Vilnius, Lithuania).

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS INSTITUTAS

Vaidas BALYS

MOKSLINĖS TERMINIJOS MATEMATINIAI MODELIAI IR JŲ TAIKYMAS LEIDINIŲ KLASIFIKAVIME

DAKTARO DISERTACIJOS SANTRAUKA

FIZINIAI MOKSLAI,
MATEMATIKA (01P)

VGTU
LEIDYKLA
VILNIUS TECHNIKA 2009

Disertacija rengta 2004–2009 metais Matematikos ir informatikos institute.
Mokslinis vadovas

prof. habil. dr. Rimanas RUDZKIS (Matematikos ir informatikos institutas, fiziniai mokslai, matematika – 01P).

Disertacija ginama Vilniaus Gedimino technikos universiteto Matematikos mokslo krypties taryboje:
Pirmininkas

prof. habil. dr. Kęstutis KUBILIUS (Matematikos ir informatikos institutas, fiziniai mokslai, matematika – 01P).

Nariai:

prof. habil. dr. Remigijus LEIPUS (Vilniaus universitetas, fiziniai mokslai, matematika – 01P),

prof. dr. Valentinas PODVEZKO (Vilniaus Gedimino technikos universitetas, socialiniai mokslai, ekonomika – 04S),

prof. habil. dr. Alfredas RAČKAUSKAS (Vilniaus universitetas, fiziniai mokslai, matematika – 01P),

prof. habil. dr. Leonas SAULIS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P).

Oponentai:

prof. dr. Kęstutis DUČINSKAS (Klaipėdos universitetas, fiziniai mokslai, matematika – 01P),

doc. dr. Marijus RADAVIČIUS (Matematikos ir informatikos institutas, fiziniai mokslai, matematika – 01P).

Disertacija bus ginama viešame Matematikos mokslo krypties tarybos posėdyje 2009 m. spalio 2 d. 13 val. Matematikos ir informatikos institute, 203 auditorijoje.
Adresas: Akademijos g. 4, LT-08663 Vilnius, Lietuva.
Tel.: (8 5) 274 4952, (8 5) 274 4956; faksas (8 5) 270 0112;
el. paštas doktor@adm.vgtu.lt
Disertacijos santrauka išsiuntinėta 2009 m. rugsėjo 1 d.
Disertaciją galima peržiūrėti Vilniaus Gedimino technikos universiteto (Saulėtekio al. 14, LT-10223 Vilnius, Lietuva) ir Matematikos ir informatikos instituto (Akademijos g. 4, LT-08663 Vilnius, Lietuva) bibliotekose.
VGTU leidyklos „Technika“ 1651-M mokslo literatūros knyga.

Introduction

Scientific problem

The classification of publications is an important scientific text processing activity providing means for accumulating information and knowledge and, which is most important, for retrieving and reusing its fragments whenever needed. The manual completion of this task is tedious, inefficient, and of doubtless sensibility in the face of current technological capabilities. Therefore, this dissertation considers the problem of automatic classification of specific scientific contents.

Topicality of the work

The concept of scientific knowledge management covers a number of common problems important to anyone who has to deal with scientific contents in one way or another. These include a convenient and attractive presentation of data and information, intuitive and flexible search tools, logical and active links between elements, etc. The automatic classification of texts, especially scientific ones, is among the most important problems of scientific knowledge management. The machine can now perform or at least help to perform the task usually accomplished by author, e. g., assign keywords or scientific classifiers. The results of classification do not have to be fixed and limited any longer – they can answer the needs of a certain system. If the classification algorithm is based on the analysis of certain relationships between elements of the text, the results of this analysis may well be used solving a number of related scientific text processing tasks. The explicit relations between elements of a scientific field, e. g., its terminology, are of great value themselves.

Automatic text classification, as it is common with the problems of an applied nature, combines and borrows ideas, methods, and researchers from a list of various fields including the probability theory and mathematical statistics, statistical data analysis, artificial intelligence, machine learning, data mining, information retrieval and natural language processing. However, despite the overwhelming amount of works related to this problem, there is a lack of research that address the specificity of scientific texts. Straightforwardly applying methods of convenient text (e-mail messages, news reports) classification, the specificity of scientific publications is not taken into consideration. Therefore, there is a natural need to create mathematical methods that would allow to make use of this specificity to develop more accurate and more suitable classification algorithms.

Research object

The research object of the dissertation is methods of multivariate discriminant analysis.

The aim and tasks of the dissertation

The main objective of the dissertation is to propose and analyse mathematical classification methods based on the analysis of scientific terminology distribution over texts that could be applied for solving the scientific publication classification problem. A number of tasks are formulated for achieving the main goal:

- develop a probabilistic model for distribution of scientific terminology over texts of publications and propose the procedures for statistical identification of this model;
- propose constructive classification algorithms, based on the model and identification procedures;
- propose mathematical methods to incorporate auxiliary information, related to positioning of terms in the text as well as the context between them;
- analyse and compare the proposed and alternatively chosen algorithms by running experiments on real data;

Research methods

The following research methods are used in the dissertation: analysis of the related scientific material, mathematical modelling (construction of a probabilistic model and procedures of its identification) and experiment (the analysis and comparison of the proposed and alternative algorithms by doing experiments on real publications).

Scientific novelty

The results of the work extend and supplement the results of other authors in the related fields. The research differs from the other ones in the problems considered, solutions proposed, and the results achieved:

- the problem of automatic classification of scientific texts is considered and the proposed methods address the specificity of such texts;
- the probabilistic model is developed for distribution of scientific terminology over texts, and its identification procedures and constructive classification algorithms are proposed;

- the procedure for selecting the most informative terms, based on the theory of statistical hypothesis testing is proposed;
- the methods for incorporating auxiliary contextual information into the classification methods are formulated;
- an exhaustive comparative analysis of the proposed and alternative algorithms was made on the base of real-world data of mathematical publications;
- the following aspects related to the specificity of scientific publications were analysed by running experiments on real data: how choosing certain parts of texts, terms vocabulary, and classifiers as well as using long texts influence the accuracy of classification.

Practical value of the results

The scientific publication classification algorithms presented in the dissertation can be implemented and used in automated systems that collect and present scientific information. This would provide users with convenient means to search and navigate the data. The proposed methodology based on statistical analysis of scientific terminology distribution may as well be used when solving a number of scientific contents processing problems. The explicitness and transparent interpretation of proposed model parameter relations makes inclusion of expert knowledge practical.

Defended propositions

1. The proposed classification methodics based on statistical analysis of scientific terminology distribution over texts.
2. The proposed procedures of the terminology distribution model identification.
3. The proposed methods for incorporating auxiliary information into classification algorithms.

The scope of the scientific work

The scientific work consists of Introduction, three chapters, Conclusions, References, list of authors publications and one appendix. The total scope of the dissertation is 101 page, 99 formulae, 15 figures, 13 tables and 86 items of references.

1. Analytical review of the text classification methods

In this chapter, we present a detailed description of the dissertation topic and review of basic approaches and the most popular methods. The problem of scientific publications classification is considered within the basic framework of multilabel classification by multivariate discriminant analysis with certain modifications and techniques for dealing with texts (representation, reduction of feature space dimensionality, etc.). The learning set of correctly classified documents is considered and algorithms are constructed by analysing these data.

The following algorithms were chosen for experiments and comparison to our ones and are presented in more detail: naive Bayes with Laplace smoothing (*nB*) that uses the Bayes theorem and a naive assumption on term independence in estimating posterior probabilities of classes having the text of a document observed; k-Nearest Neighbours (*kNN*) that defers the learning phase until the document to be classified arrives and then consults the classification decisions of the most similar documents; Linear Least Squares Fit (*LLSF*) that assumes the existence of linear dependence between weights of classes and weights of terms of the document; Support Vector Machines (*SVM*) that interprets documents as vectors in a multidimensional Euclidean space and looks for hyperplanes that correctly separate (belonging to a class – on one side, the remaining – on the other one) the learning data for each class so that the separation margins are as wide as possible.

Even though there is a long list of efficient algorithms, some of which are presented in more detail in this chapter, there is a lack for research addressing the specificity of scientific publications such as the mathematical language, long non-homogeneous texts, certain structure, etc.

2. Models for scientific terminology and their applications in the classification

In this chapter, the method for classifying scientific publications is proposed. It is based on the probabilistic model of scientific term distributions over texts and its identification procedures. Let us now present a brief overview of our approach.

Definitions and notation

Let K denote a classification system of scientific texts which is identified with a set of all possible labels of the classes in that system. In the case of scientific (mathematical) publications, it usually consists of classifiers from the standard schemes like MSC or keyphrases from some controlled vocabulary.

Let V be a vocabulary of scientific terms (not only single words but also phrases) of a certain field that are relevant to the classification of texts. The chronologically numerated vector of the article a elements $(a_1, \ldots, a_d)$, $d = d(a)$, where $a_i \in V$ and not necessarily $a_i \neq a_j$, is called the projection of the article a. 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 homogeneous piece of text – in the general case, it consists of $q = q(a) \geq 1$ continuous homogeneous parts which are classified as different in system K. However, in the dissertation, we restrict ourselves to the case when the whole considered text is homogeneous.

Let N be a set of natural numbers. Let an article $a \in A$ and a set of indices $I \subset N$ be chosen randomly. The article is attributed to the class η in the system K which we need to estimate using the observed vector $a_I = (a_\tau, \tau \in I)$.

Probability distributions and a Bayes classifier

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

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

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

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

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

where $d(y) = dim(y)$ denotes the dimension of vector y and $|I| = card\ I$ denotes the cardinality of set I.

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

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

Using the distributions, described in equations (1) and (2), the Bayes classifier which minimises mean classification losses can be defined. In the usual case of a trivial loss function, it is the maximum a posterior 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 $P(a_I|\cdot)$:

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

Model identification

In order to use classification method (5), the distribution Q and the functional ψ must be estimated.

Let us have the learning sample of the observed parts of texts and their classification results $X_n = (y(1), \eta(1)), \ldots, (y(n), \eta(n))$, where $\eta(i) \in K, y(i) \in Y$, $Y = \{y = (y_1, \ldots, y_d) : y_i \in V, d \in N\}$. The empirical analogue of $Q(w)$ is determined by $\widehat{Q}(w) = \sum_{j=1}^{n} 1_{\{\eta(j)=w\}}/n$. The functional $\psi_w(y)$ can be estimated by using a common k-Nearest Neighbours method with specifically designed distance measure. Let us now consider the parametric estimation. 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 following assumptions substantially simplify the procedures of estimation.

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

$$P(y|w) = \prod_{i=1}^{d(y)} P(y_i|w) \tag{6}$$

hold.

Now the definition of (3) can be replaced by

$$\psi_w(v) = P(v|w)/P(v), \quad v \in V, w \in K \tag{7}$$

and the Bayes classification rule for classifying the observed a_I is determined by

$$\widehat{\eta} = \arg\max_{w \in K} \left[Q(w) \prod_{\tau \in I} \psi_w(a_\tau) \right]. \tag{8}$$

The definition of (7) based on assumption (6) 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 $w \in K$

$$P(y|w) = P(y_1|w) \prod_{i=1}^{d-1} \left[P(y_i, y_{i+1}|w)/P(y_i|w) \right] \tag{9}$$

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

In this case, (3) can be replaced by two functionals: $\psi_w(v)$ defined in (7) and

$$\psi_w(v,u) = P(v,u|w)/P(v,u), \quad v,u \in V, \tag{10}$$

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 equation (4) according to equations (7), (9), and (10):

$$\widehat{\eta} = \arg\max_{w\in K}\left[Q(w)\psi_w(a_r)\prod_{i=r}^{m-1}\left[\psi_w(a_i,a_{i+1})/\psi_w(a_i)\right]\right]. \tag{11}$$

In order to use algorithms (8) and (11), the functionals $\psi_w(\cdot)$ and $\psi_w(\cdot,\cdot)$ have to be estimated. First the empirical estimates of the probabilities $P(\cdot)$, $P(\cdot,\cdot)$, $P(\cdot|\cdot)$ and $P(\cdot,\cdot|\cdot)$ are calculated by counting the corresponding frequencies and these estimates are used in (7) and (10) yielding empirical estimates $\widetilde{\psi}_w(v)$ and $\widetilde{\psi}_w(v,u)$. The smoothing is then performed – the unreliable estimates, i. e., the ones based on too few observations, are modified.

Selecting the most informative terms

Only a part of the most informative terms is actually used for classification. Let us now present the method for selecting them in the case of assumption (6).

Let $h = |V|$. The functional $\psi_w(\cdot)$ determines the arrangements of set V for each $w \in K$:

$$\psi_w(v_1) \geq \psi_w(v_2) \geq \ldots \geq \psi_w(v_h), \quad v_{(\cdot)} \in V. \tag{12}$$

Firstly, we arrange the set V as (12) by using the estimate $\widetilde{\psi}_w(\cdot)$ instead of unknown $\psi_w(\cdot)$. Then we choose only a part of the most informative terms:

$$\widehat{\psi}_w(v_k) = \begin{cases} \widetilde{\psi}_w(v_k), \text{ if } k \in L, \\ \quad 1, \text{ otherwise.}\end{cases}$$

The set $L = L(w) \subset \{1, \ldots, h\}$ consists of indices of the terms that have $\widetilde{\psi}_w(v_k)$ significantly differing from 1. In order to get L, let us consider the hypothesis

$$H_0 : \ \psi_w(v) = 1 \tag{13}$$

with an alternative

$$H_1 : \psi_w(v) > 1$$

and let $\overline{\alpha}(v)$ denote a p–value.

Analogously, let us consider the same hypothesis (13), but with another alternative

$$H_1 : \psi_w(v) < 1$$

and let $\underline{\alpha}(v)$ denote the corresponding p–value.

Choosing the level of significance α, the set of the most informative terms may be defined as follows: $L = \overline{L} \cup \underline{L}$, where $\overline{L} = \{k : \widetilde{\psi}_w(v_k) > 1, \overline{\alpha}(v_k) < \alpha\}$ and $\underline{L} = \{k : \widetilde{\psi}_w(v_k) < 1, \underline{\alpha}(v_k) < \alpha\}$. Few other methods for selecting the most informative terms are presented and analysed in the dissertation. In the dissertation, formulae for determining the values of $\overline{\alpha}$ and $\underline{\alpha}$ are presented.

The parametric model for the remaining informative terms, depending only on the positions of terms in the arrangement (12), is proposed.

Auxiliary information

By choosing only a part of scientific terms from the text and using stationary distributions $P(\cdot)$, $P(\cdot|\cdot)$, $P(\cdot,\cdot)$, $P(\cdot,\cdot|\cdot)$, we ignore the information of the context between the terms as well as the positioning of these terms in the text. When encountering long and non-homogeneous texts which is the case with scientific publications, it is natural to consider at least a part of this information as it is obvious that a term has different discriminative weight depending on which part of an article text it is observed in.

Let us redefine the projection of an article by $a = ((a_1, \lambda_1), \ldots, (a_d, \lambda_d))$, $d = d(a)$, where λ_i is the auxiliary information for a term a_i related to the positioning of this term in the text. We define $\lambda_i = (\lambda_i^{(t)}, \lambda_i^{(w)}, \lambda_i^{(s)}, \lambda_i^{(p)})$ where the components denote the sequential number of term, word, sentence, and paragraph of the term a_i. It would be natural to use some higher-level position indicators like "in abstract", "in introduction", etc. While this could be done rather easily, we do not do that in the dissertation because these elements were unidentifiable in the data we possessed for the experiments.

Let us introduce functionals $\sigma = \sigma(\lambda_{(\cdot)})$ and $\delta = \delta(\lambda_{(\cdot)}, \lambda_{(\cdot)})$ taking values in $[0, 1]$ where σ defines the weight of a term depending on its position and δ defines the rate of dependence of two terms depending on their positions. Having the learning sample $X_n = (y(1), \lambda(1), \eta(1)), \ldots, (y(n), \lambda(n), \eta(n))$, $\eta(i) \in K, y(i) \in Y$, enriched with auxiliary information $\lambda(j) = (\lambda_1(j), \ldots, \lambda_d(j))$,

$d = d(y(j))$, we derive the weighted estimates for $P(\cdot)$ and $P(\cdot|\cdot)$ as follows:

$$S(j) = \sum_{k=1}^{d(j)} \sigma(\lambda_k(j)),$$

$$\widetilde{P}^*(v) = \sum_{j=1}^{n} \sum_{k=1}^{d(j)} \sigma(\lambda_k(j)) \cdot \mathbb{1}_{\{y_k(j)=v\}} / \sum_{j=1}^{n} S(j),$$

$$\widetilde{P}^*(v|w) = \sum_{j=1}^{n} \sum_{k=1}^{d(j)} \sigma(\lambda_k(j)) \mathbb{1}_{\{y_k(j)=v, \eta(j)=w\}} / \sum_{j=1}^{n} \mathbb{1}_{\{\eta(j)=w\}} \cdot S(j).$$

Estimates for $P(\cdot,\cdot)$ and $P(\cdot,\cdot|\cdot)$ that are also presented in the dissertation. The new estimates are substituted in (7) and (10) for the unknown true values and then the same procedures for smoothing and selecting the most informative terms are applied with these new estimates of $\psi_w(\cdot)$ and $\psi_w(\cdot,\cdot)$.

Analogously, the classification procedures (8) ((11) as well) can be redefined as follows to take into account the auxiliary information:

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

In order to use the enhanced estimation procedures and classification methods one has to define the functionals $\sigma(\lambda)$ and $\delta(\lambda, \lambda)$ that describe the relative weight of the scientific terms and their pairs depending on the distance from the beginning of the text as well as the context between terms. In the dissertation, certain variants are proposed and analysed.

The constructive classification algorithms based on the proposed model, its variants under assumptions, and identification procedures are formulated.

3. Experimental evaluation of classification methods

In this chapter, the results of experiments on the data base of around 15000 publications from the field of probability theory and mathematical statistics, kindly provided by the Institute of Mathematical Statistics, USA, are reported.

The k–fold cross-validation with $k = 5$ performance evaluation method is used. Both MSC classifiers and keywords (keyphrases) are tested as class labels. A fixed number of classes (from 1 to 5) is assigned to the documents and the common precision, recall, and F_1 together with the proposed *average precision* (averaged at the points of recall increase) metrics are measured. Three different term vocabularies are used: the first one consists of all the keywords and keyphrases from the articles in the data base, the second one consists only of keywords and the third one consists of all the English words found in the texts.

The proposed methods are analysed and compared to the chosen naive Bayes, k-Nearest Neighbours, Linear Least Squares Fit (*LLSF*), and Support Vector Machines (*SVM*) algorithms. The methods are analysed and compared under different settings of terms dictionary, class labels set, parts of texts for learning and testing, specific parameters of algorithms, etc.

General conclusions

1. There is a lack of algorithms designed specifically for the classification of scientific texts. The natural language classification algorithms are usually used both in research and practical applications.
2. Employing methods for transforming documents into quantitative vectors and having large data sets of pre-classified publications, the scientific text classification methods may be solved by methods of multivariate discriminant analysis.
3. In the dissertation, a probabilistic model for scientific terminology distribution over texts of publications and the model's identification procedures are formulated. Mathematical methods for incorporating auxiliary information related to the positions of terms in the text and the context between them are proposed. The constructive classification algorithms that may be used for solving the problem of classification of scientific publications are constructed.
4. The experimental study showed that choosing the set of textual elements (words, phrases) used to represent documents strongly influences the results of algorithms. The use of a scientific terms vocabulary (both single-word and phrases) constructed from the publications of the data base instead of the usual vocabulary of all the language words, found in the texts, leads to a 5–10% higher accuracy of algorithms.
5. Employment of full texts (not only abstracts) in both the learning and classification (testing) phase yields up to a 10–12% accuracy increase for the algorithms. An increase in the accuracy of 5–7% is also observed when using full texts only in the learning phase. The used of full texts only in the testing phase yields an increase in the accuracy of up to 3–5%.

6. The procedures proposed for estimating the weights of terms and selecting the most informative ones reduce the effective size of vocabularies by another 20–30% after common reduction by the *Document Frequency* method while the classification accuracy is not reduced.

7. The assumptions of more complicated forms (other than conditional independence and stationarity) of relations between distributions of terms lead to increase of computational efforts without any gain in accuracy. In order to analyse the generalisation of methods that allow to take into consideration distant pairs of terms and the context between them a much bigger learning sets are necessary.

8. The experimental analysis of methods of using auxiliary information related to the positions of terms in text showed that, in case of very long texts, usage of stepwise-reducing weight systems yields an up to 2% increase in accuracy of the proposed algorithm as compared to the variant of algorithm that does not take the auxiliary information into consideration. In this case, the accuracy of the proposed algorithm is slightly (up to 1%) higher than that of the best performing alternative methods.

9. The proposed algorithm has important practical advantages related to the transparent interpretation of results and model parameter relations. This enables one to improve algorithms by using expert knowledge. Also, the algorithms and the results of analysis of terminology relations may well be used for solving related scientific texts processing problems.

List of Published Works on the Topic of the Dissertation

In the reviewed scientific periodical publications

Balys, V.; Rudzkis, R. 2008. Classification of Publications Based on Statistical Analysis of Scientific Terms Distributions, *Austrian Journal of Statistics* 37(1): 109–118.

Rudzkis, R.; Balys, V.; Hazewinkel, M. 2006. Stochastic Modelling of Scientific Terms Distribution in Publications, *Proceedings of 5th International Conference on Mathematical Knowledge Management*, *Lecture Notes in Artificial Intelligence* 4108: 152–164.

Balys, V.; Rudzkis, R. 2005. Stochastinių mokslo terminų aplinkų modelių tyrimas (Analysis of Stochastic Models for Context of Scientific Terminology), *Lietuvos matematikos rinkinys, special issue* 45: 329–334.

Balys, V.; Rudzkis, R. 2004. Mokslo terminų aplinkų modelių taikymas straipsnių klasifikavime (The Applications of Models for Context of Scientific Termi-

nology in the Classification of Articles), *Lietuvos matematikos rinkinys, special issue* 44: 537–541.
Balys, V.; Rudzkis, R. 2003. Mokslinių terminų statistinio pasiskirstymo taikymas straipsnių klasifikavime, (The Applications of Statistical Distribution of Scientific Terminology in the Classification of Articles), *Lietuvos matematikos rinkinys, special issue* 43: 463–467.

In the other editions

Rudzkis, R.; Balys, V. 2007. On Statistical Classification of Scientific Texts, in *Proceedings of 8th International Conference on Computer Data Analysis and Modeling* 1: 100–103.
Balys, V.; Rudzkis, R. 2004. Stochastic Models for Keyphrase Assignment, in *Proceedings of 7th International Conference on Computer Data Analysis and Modeling* 1: 118–122.

About the author

Vaidas Balys was born in Panevėžys, on 21 of March, 1980.

He graduated from the J. Balčikonis gymnasium in 1998. Vaidas won the first place in two consecutive Lithuanian Mathematics Olympiads in 1997 and 1998, and participated in the International Mathematics Olympiads. He entered Vilnius University, Faculty of Mathematics and Informatics in 1998. First degree in Informatics, 2002. Master of Science in Informatics with magna cum laude diploma in 2004. In 2004–2008 – PhD student of Institute of Mathematics and Informatics. At present works in UAB "VTEX", LaTeX-based technical typesetter and data supplier for science publishers.

MOKSLINĖS TERMINIJOS MATEMATINIAI MODELIAI IR JŲ TAIKYMAS LEIDINIŲ KLASIFIKAVIME

Tiriamoji problema

Publikacijų klasifikavimas yra viena iš svarbių mokslo tekstų tvarkymo veiklų, sudarančių galimybes kaupti, ir, kas svarbiausia, esant poreikiui surasti bei panaudoti mokslinės informacijos ir žinių fragmentus. Rankinis šio darbo atlikimas ne tik neefektyvus ir sudėtingas, bet ir neprasmingas dabartinių techninių galimybių kontekste. Todėl šioje disertacijoje nagrinėjama specifinio mokslinio turinio automatinio klasifikavimo problema.

Darbo aktualumas

Mokslo žinių valdymo sąvoka apima kelis pagrindinius uždavinius, aktualius kiekvienam vienaip ar kitaip susiduriančiam su moksliniu turiniu: patogus ir patrauklus informacijos ir duomenų pateikimas, lanksti ir intuityvi paieška, logiškos ir aktyvios sąsajos tarp elementų ir kt. Mokslo žinių valdymo problemų rate automatinis tekstų, ypač – mokslinių, klasifikavimas yra vienas aktualiausių uždavinių. Dabar tai, ką paprastai atlikdavo pats teksto autorius, pvz., nurodydamas raktinius žodžius ar išvardindamas mokslo sritį apibūdinančius standartinius klasifikatorių kodus, gali atlikti arba bent jau gali padėti atlikti automatinė sistema. Rezultatai nebeprivalo būti apriboti ir fiksuoti – jie gali kisti, priklausomai nuo konkrečių sistemos poreikių. Jei klasifikavimo algoritmai remiasi tam tikrų teksto elementų ryšių analize, jos rezultatus galima naudoti sprendžiant kitus mokslo tekstų apdorojimo uždavinius. O patys sąryšiai tarp mokslo srities elementų, pavyzdžiui, jos terminijos, savo ruožtu yra vertingas rezultatas, suteikiantis papildomų žinių apie nagrinėjamąją sritį.

Automatinis tekstų klasifikavimas, kaip ir didžioji dalis taikomojo pobūdžio uždavinių, apjungia bei skolinasi idėjas, metodus ir tyrėjus iš daugelio skirtingų mokslo krypčių bei sričių, tokių kaip tikimybių teorija ir matematinė statistika, statistinė duomenų analizė, dirbtinis intelektas, automatinis mokymasis (*machine learning*), duomenų gavyba (*data mining*), informacijos ištraukimas (*information retrieval*), kalbos apdorojimas (*natural language processing*) ir kt. Tačiau, nors bendras su tekstų klasifikavimu susijusių atliktų tyrimų skaičius yra labai didelis, darbų, nagrinėjančių būtent mokslo tekstams pritaikytus metodus, beveik nėra. Tiesiogiai taikant įprastų tekstų (elektroninio pašto žinutės, naujienų pranešimai) klasifikavimo algoritmus neatsižvelgiama į mokslo publikacijų specifiką: ilgą ir nehomogenišką tekstą, griežtą struktūrą, specifinę kalbą ir kt. Todėl iškyla natūralus poreikis sukurti matematinius metodus, kurie leistų į šią specifiką atsižvelgti ir ją panaudoti konstruojant tikslesnius ir geriau pritaikytus klasifikavimo algoritmus.

Tyrimo objektas

Darbo tyrimų objektas yra daugiamatės diskriminantinės analizės metodai.

Darbo tikslas ir uždaviniai

Pagrindinis darbo tikslas yra pasiūlyti ir ištirti matematinius klasifikavimo metodus, paremtus mokslo terminijos pasiskirstymo tekstuose analize, kuriuos būtų galima taikyti taikomajam mokslo publikacijų klasifikavimo uždaviniui spręsti.

Pagrindiniam darbo tikslui pasiekti suformuluoti šie uždaviniai:

- sudaryti mokslo terminijos pasiskirstymo publikacijų tekste tikimybinį modelį ir sukurti modelio identifikavimo procedūras;
- sukurti šiuo modeliu ir jo identifikavimo procedūromis pagrįstus konstruktyvius klasifikavimo algoritmus;

- pasiūlyti matematinius metodus, kaip klasifikavimo algoritmuose atsižvelgti į papildomą informaciją, susijusią su mokslo terminų pozicijomis ir kontekstu tarp jų;
- realių duomenų pagrindu ištirti pasiūlytus sprendimus, atlikti sukurtų bei alternatyvių klasifikavimo algoritmų palyginamąją analizę.

Tyrimų metodai

Darbe taikomi šie tyrimų metodai: mokslinės literatūros disertacijos tema analizė, matematinis modeliavimas (tikimybinio modelio ir jo identifikavimo procedūrų formulavimas) ir eksperimentas (pasiūlytųjų sprendimų analizė ir palyginimas su alternatyviais metodais, atliekant bandymus su realių publikacijų duomenų baze).

Darbo mokslinis naujumas

Atlikto darbo rezultatai papildo ir praplečia kitų šioje bei giminiškose srityse atliktų tyrimų rezultatus. Nuo kitų autorių darbų skiriasi šiais nagrinėjamais klausimais, siūlomais sprendimais bei pasiektais rezultatais:

- nagrinėtas mokslo publikacijų tekstų klasifikavimo uždavinys, siūlomuose metoduose atsižvelgta į šių tekstų specifiką;
- sudarytas tikimybinis mokslo terminijos pasiskirstymo tekste modelis, sukurtos jo identifikavimo procedūromis, suformuluoti originalūs klasifikavimo algoritmai;
- pasiūlytas informatyviausių mokslo terminų nustatymo būdas, paremtas statistinių hipotezių tikrinimo teorijos metodų taikymu;
- sukurti papildomos informacijos, susijusios su terminų pozicijomis tekste bei kontekstu tarp jų, naudojimo klasifikavime metodai;
- atlikta išsami palyginamoji pasiūlytųjų bei keleto populiarių kitų autorių klasifikavimo metodų analizė realios mokslo publikacijų bazės pagrindu;
- tos pačios bazės pagrindu ištirti šie su mokslo publikacijų specifika susiję aspektai: atskirų teksto dalių naudojimo, mokslo terminijos žodyno parinkimo, pačios klasifikavimo sistemos (klasių rinkinio) parinkimo, ilgų tekstų naudojimo įtaka klasifikavimo tikslumui.

Darbo rezultatų praktinė vertė

Darbe suformuluoti konstruktyvūs mokslo publikacijų klasifikavimo algoritmai gali būti tiesiogiai realizuoti mokslo informaciją kaupiančiose ir pateikiančiose automatizuotose sistemose, o tai leistų naudotojams pasiūlyti patogias ir lanksčias duomenų paieškos ir navigacijos priemones. Siūloma mokslo terminijos pasiskirstymo statistine analize paremta klasifikavimo metodika gali būti pritaikyta sprendžiant ir kitus mokslo tekstų apdorojimo uždavinius. O suformuluoto modelio parametrų sąryšių išreikštinumas ir aiški interpretacija sudaro praktines galimybes į metodus įtraukti ekspertines žinias.

Ginamieji teiginiai

1. Pasiūlyta klasifikavimo metodika, pagrįsta statistine mokslo terminijos pasiskirstymo publikacijų tekstuose analize.

2. Sukurtos pasiūlyto terminijos pasiskirstymo modelio identifikavimo procedūros.

3. Pasiūlyti papildomos kontekstinės informacijos panaudojimo klasifikavime metodai.

Darbo apimtis

Disertaciją sudaro įvadas, trys pagrindiniai skyriai, išvados, naudotos literatūros sąrašas, autoriaus publikacijų disertacijos tema sąrašas ir vienas priedas. Pirmasis skyrius skirtas analitinei mokslinės literatūros disertacijos tema apžvalgai. Antrajame skyriuje pateikiama autoriaus siūloma klasifikavimo metodika. Trečiajame skyriuje pateikiami eksperimentinio pasiūlytų ir alternatyvių metodų tyrimo rezultatai.

Darbo apimtis yra 101 puslapis, tekste panaudotos 99 numeruotos formulės, 15 paveikslų ir 13 lentelių. Rašant disertaciją buvo pasinaudota 86 literatūros šaltiniais.

Bendrosios išvados

1. Moksliniams tekstams klasifikuoti skirtų specialių algoritmų nėra, paprastai tyrimuose ir praktiniuose taikymuose naudojami įprasti bendrinės kalbos tekstų klasifikavimo metodai.

2. Pritaikius dokumentų reprezentavimo skaitiniais požymių vektoriais metodus ir turint dideles teisingai suklasifikuotų dokumentų aibes, taikomąjį mokslo tekstų klasifikavimo uždavinį galima spręsti naudojant daugiamatės diskriminantinės analizės metodus.

3. Darbe suformuluotas tikimybinis mokslo terminijos pasiskirstymo publikacijų tekstuose modelis bei jo identifikavimo procedūros. Sukurti matematiniai metodai, kaip į klasifikavimą įtraukti papildomą kontekstinę informaciją, susijusią su terminų pozicijomis tekste ir kontekstu tarp jų. Sukurti konstruktyvūs klasifikavimo algoritmai, kurie gali būti taikomi mokslo publikacijų klasifikavimo uždaviniui spręsti.

4. Atliktas eksperimentinis tyrimas parodė, kad teksto elementų (žodžių, frazių), kuriuos atitinkantys skaitiniai požymiai reprezentuoja dokumentus, aibės parinkimas stipriai įtakoja algoritmų rezultatus. Naudojant iš tyrimams naudotos publikacijų duomenų bazės sudarytą mokslo terminijos (pavienių žodžių ir frazių) žodyną gautas apie 5–10 % didesnis tirtų klasifikavimo algoritmų tikslumas, nei naudojant įprastą visų kalbos žodžių, sutinkamų straipsnių tekstuose, žodyną.

5. Pilnų tekstų (ne tik santraukų) naudojimas ir mokymo, ir klasifikavimo fazėse iki 10–12 % padidina tirtųjų algoritmų tikslumą. Apie 5–7 % tikslumo padidėjimas stebimas naudojant pilnus tekstus vien tik mokymo fazėje. Pilnų tekstų naudojimas tik klasifikavimo fazėje duoda iki 3–5 % tikslumo padidėjimą.

6. Pasiūlytos terminų svorių nustatymo bei informatyvių terminų atrinkimo procedūros leidžia apie 20–30 % sumažinti realiai naudojamo žodyno, jau sumažinto įprastu *DF* metodu, apimtis, neprarandant algoritmo tikslumo.

7. Prielaidos apie sudėtingesnes terminijos skirstinių tarpusavio sąryšio formas (negu sąlyginis nepriklausomumas ir stacionarumas) kelis kartus padidina tikimybinių algoritmų skaičiavimo sąnaudas, tačiau nepadidina tikslumo. Pasiūlytiems metodų apibendrinimams, leidžiantiems atsižvelgti į nutolusias terminų poras bei kontekstą tarp jų, ištirti reikalingos didesnės mokymo imtys.

8. Papildomos kontekstinės informacijos, susijusios su terminų pozicijomis tekste, panaudojimo metodų eksperimentinis tyrimas parodė, kad labai ilgų tekstų atveju naudojant žingsneliais mažėjančių svorių sistemas pasiekiamas iki 2 % didesnis pasiūlytojo algoritmo tikslumas, lyginant su algoritmu, nenaudojančiu papildomos informacijos. Šiuo atveju pasiūlytasis algoritmas nenusileidžia ir net nežymiai (iki 1 %) lenkia alternatyvius tiksliausius algoritmus.

9. Pasiūlytasis algoritmas turi svarbių aiškios rezultatų ir modelio parametrų sąryšių interpretacijos lemiamų taikomojo pobūdžio privalumų. Metodus galima papildyti ekspertinėmis žiniomis, o sukurti algoritmai ir terminijos sąryšių analizės rezultatai gali būti panaudoti sprendžiant kitus mokslo tekstų apdorojimo uždavinius.

Trumpos žinios apie autorių

Vaidas Balys gimė 1980 m. kovo 21 d. Panevėžyje.

1998 m. baigė Panevėžio J. Balčikonio gimnaziją. 1997 ir 1998 m. respublikinėse moksleivių matematikos olimpiadose užėmė pirmąją vietą ir dalyvavo tarptautinėse matematikų olimpiadose Lietuvos komandos sudėtyje. 1998 m. įstojo į Vilniaus universiteto Matematikos ir informatikos fakultetą. 2002 m. baigė pagrindinių studijų informatikos programą ir įgijo informatikos bakalauro kvalifikacinį laipsnį. 2004 m. su pagyrimu (magna cum laude) baigė Vilniaus universiteto magistrantūros studijų kompiuterinio modeliavimo programą ir įgijo informatikos magistro kvalifikacinį laipsnį. 2004–2008 m. – Matematikos ir informatikos instituto doktorantas. Nuo 2007 m. dirba UAB „VTEX“.

Vaidas BALYS

MATHEMATICAL MODELS FOR SCIENTIFIC TERMINOLOGY
AND THEIR APPLICATIONS IN THE CLASSIFICATION OF PUBLICATIONS

Summary of Doctoral Dissertation
Physical Sciences, Mathematics (01P)

Vaidas BALYS

MOKSLINĖS TERMINIJOS MATEMATINIAI MODELIAI
IR JŲ TAIKYMAS LEIDINIŲ KLASIFIKAVIME

Daktaro disertacijos santrauka
Fiziniai mokslai, matematika (01P)

2009 07 17. 1,5 sp. l. Tiražas 70 egz.
Vilniaus Gedimino technikos universiteto
leidykla „Technika",
Saulėtekio al. 11, 10223 Vilnius,
http://leidykla.vgtu.lt
Spausdino UAB „Biznio mašinų kompanija"
J. Jasinskio g. 16A, 01112 Vilnius