Text Representation in Multi-label Classification: Two New Input Representations

Rodrigo Alfaro[1,2] and Héctor Allende[1,3]

[1] Universidad Técnica Federico Santa María, Chile.
[2] Pontificia Universidad Católica de Valparaíso, Chile.
[3] Universidad Adolfo Ibáñez, Chile.
rodrigo.alfaro@ucv.cl ; hallende@inf.utfsm.cl

Abstract. Automatic text classification is the task of assigning unseen documents to a predefined set of classes. Text representation for classification purposes has been traditionally approached using a vector space model due to its simplicity and good performance. On the other hand, multi-label automatic text classification has been typically addressed either by transforming the problem under study to apply binary techniques or by adapting binary algorithms to work with multiple labels. In this paper we present two new representations for text documents based on label-dependent term-weighting for multi-label classification. We focus on modifying the input. Performance was tested with a well-known dataset and compared to alternative techniques. Experimental results based on Hamming loss analysis show an improvement against alternative approaches.

Keywords: Multi-label text classification, text modelling, problem transformation.

1 Introduction

Large amounts of text document available on digital format on the web contain useful information for a wide variety of purposes. The amount of digital text is expected to increase significantly in the near future; thus, the need for the development of data analysis solutions becomes urgent. Text classification (or categorisation) is defined as the assignment of a Boolean value to each pair $\langle d_j, c_i \rangle \in \mathcal{D} \times \mathcal{C}$, where $\mathcal{D}$ is the domain of documents and $\mathcal{C} = \{c_1, ..., c_{|\mathcal{C}|}\}$ is the set of predefined labels [12].

Binary classification (BC) is the simplest and most widely studied case. In BC, a document is classified into one of two mutually exclusive classes. BC can be extended to solve multi-class problems. Moreover, if a document is categorised with either one label or multiple labels at once, it is called a single-label or multi-label problem, respectively [12].

Tsoumakas and Katakis [14] presents a formal description of multi-label methods. In [14], $L = \{\lambda_j : j = 1 \dots l\}$, where λ_j corresponds to the j-th label,

is the finite set of labels in a multi-label learning task, and $D = \{f(\mathbf{x_i}; Y_i); i = 1...m\}$ denotes a set of multi-label training data, where $\mathbf{x_i}$ is the feature vector and $Y_i \subseteq L$ is the set of labels of the $i-$th example. Methods for solving this problem are grouped into two types, namely, problem transformation and algorithm adaptation. The first type of methods is algorithm-independent; it transforms the multi-label learning task into one or more single-label classification tasks. Thus, this type of method can be implemented using efficient binary algorithms. The most common problem transformation method (PT4) learns $|L|$ binary classifiers $H_l : X \to \{l, \neg l\}$, one for each different label l in L. PT4 transforms the original data set into $|L|$ data sets $D_{l:l=1...|L|}$. Each D_l labels every example in D with l if l is contained in the example or $\neg l$, otherwise. PT4 yields the same solution for both the single-label and multi-class problems using a binary classifier. For the classification of a new instance x, this method generates a set of labels as the union of the labels generated by the $|L|$ classifiers $H_{PT4}(x) = \bigcup_{l \in L}\{l\} : H_l(x) = l$. The second type of method extends specific learning algorithms for handling multi-label data directly. This extensions are achieved by adjustments such as modifications to classical formulations from statistics or information theory. The pre-processing of documents for better representation can also be grouped in this type.

Multi-label classification is an important problem for real applications, as can be observed in many domains, such as functional genomics, text categorisation, music mining and image classification.

The purpose of this paper is to present a new representation for documents based on label-dependent term-weighting. Lan et al. [6] propose $tf-rf$ representation for two classes of single-label classification problems. Our representation is a generalisation of the $tf-rf$ applied to multilabel classification problems.

This paper is organised as follows. In section 2, we briefly introduce multi-label text classification. In section 3, we analyse text representation. Our proposal for two new methods of representation is illustrated in section 4. In section 5, we compare the performance of our proposal with other algorithms. The last section is devoted to concluding remarks.

2 Multi-label text classification

The automatic classification of multi-label text has not been thoroughly addressed in the existing literature. Although many multi-label datasets are available, most of the techniques for automatic text classification consider them only as single-label dataset. One of the first approaches developed was Boostexter, an algorithm based on Boosting for the multi-label case [11]. This algorithm adjust the weights of training examples and their labels in the training phase; labels that are hard (easy) to predict correctly get incrementally higher (lower) weights. Among the proposal presented in [14], problem transformation is the most widely used. However, the automatic classification of multi-label text has been addressed by algorithms that directly capture the characteristics of the multi-label problem. Zhang and Zhou, for example, solved the multi-label problem using Backpropagation for Multilabel Learning (Bp-MLL), using artificial

neural networks with multiple outputs. Bp-MLL is derived from Backpropagation by employing a novel error function capturing the characteristics of multilabel learning [16].

Regardless of the solution approaches to the Multi-label problem and the algorithms that solve it, according to Joachims [4], any text classification task has complexities due to the high-dimensional feature space, a heterogeneous use of terms, and a high level of redundancy. Multi-label problems have additional complexities, including a large number of tags per document. These characteristics of a multi-label problem require different methods of evaluation than those used in traditional single-label problems.

3 Problem representation

The performance of a reasoning system depends heavily on problem representation. The same task may be easy or difficult, depending on the way it is described [3]. The explicit representation of relevant information enhances machine performance. Also, a more complex representation may work better with simpler algorithms.

Document representation has high impact on the task of classification [5]. Some elements used for representing documents include N-grams, single-word, phrases, or logical terms and statements. The vector space model is one of the most widely used models for ad-hoc information retrieval, mainly because of its conceptual simplicity and the appeal of its underlying metaphor of using spatial proximity for semantic proximity [9].

Space representation can be conceived has a kernel representation. Kernel methods are an approach for solving machine learning problems. Joachims was among the first author to use kernel-based methods to categorise text [4]. Cristianini et al. utilised the kernel-based approach for representing the vector space model and latent semantic indexing [2]. Similarly, Tsivtsivadze et al. established a mapping of input data into a feature space by means of a kernel function and then used learning algorithms to discover relationships in that space [13].

In the vector space model (VSM), the contents of a document are represented by a vector in the term space $d = \{w_1; \ldots; w_k\}$, where k is the size of the term (or feature) set. Terms may be measured at several levels, such as syllables, words, phrases, or any other semantic and/or syntactic unit used to identify the content of a text. Different terms have different importance within a text, and thus, the relevance indicator w_i (usually between 0 and 1) represents how much the term t_i contributes to the semantics of the document d.

For weight terms in the vector space model, word frequency of occurrence in the document can be used as term weight for term-weighting. However, there are more effective methods for term-weighting. The basic information used to derive term-weighting is term frequency, document frequency, or sometimes collection frequency.

There are different mappings of text to input space across different text classifications. Leopold and Kindermann, for example, combines mappings with different kernel functions in support vector machines [8]. According to Lan et al.

[7], two important decisions for choosing a representation based on VSM are the following. First, what should constitute a term? For example, should it be a sub-word, word, multi-word or meaning? Second, how should a term be weighted? Term-weighting can be a binary function or term frequency-inverse document frequency ($tf-idf$) developed by Salton and Buckley [10], using feature selection metrics such as χ^2, information gain (IG), or gain ratio (GR). Term-weighting methods improve the effectiveness of text classification by assigning appropriate weights to terms. Although text classification has been studied for several decades, term-weighting methods for text classification are usually borrowed from the traditional information retrieval (IR) field, including, for example, the Boolean model, $tf-idf$, and its variants.

Table 1 shows the variables that we will consider in a term-weighting method for multi-label problems.

Table 1. Variables utilized in a term-weighting in multi-label problem for a term t with $|L|$ labels

	t	$\overline{t}$
$label_1$	a_{t,λ_1}	d_{t,λ_1}
$label_{\lambda_j}$	a_{t,λ_j}	d_{t,λ_j}
$label_{\|L\|}$	$a_{t,\|L\|}$	$d_{t,\|L\|}$

where a_{t,λ_j} is the number of documents in the class λ_j containing the term t and d_{t,λ_j} is the number of documents in the class λ_j that do not contain the term t.

3.1 Bag-of-Words representation ($tf-idf$)

The most widely used document representation for text classification is $tf-idf$ [12], where for a two classes problem (where, $label_1$ is $class^+$ and $label_2$ is $class^-$) each component of the vector is computed as:

$$tf-idf_{td} = f_{t,d} \log_{10}\Big(\frac{N}{N_t}\Big), \tag{1}$$

where $f_{t,d}$ is the frequency of term t in the document d, $N = (a_{t,\lambda_1} + d_{t,\lambda_1} + a_{t,\lambda_2} + d_{t,\lambda_2})$ is the number of documents, and $N_t = (a_{t,\lambda_1} + a_{t,\lambda_2})$ is the number of documents containing the term t.

3.2 Relevance frequency representation ($tf-rf$)

Lan et al. [7] proposed recently $tf-rf$ as an improved VSM representation based on two classes and single-label problems (where, $label_1$ is $class^+$ and $label_2$ is $class^-$):

$$tf-rf_{td} = f_{t,d} \log_2\Big(2 + \frac{a_{t,\lambda_1}}{\max\big(1, a_{t,\lambda_2}\big)}\Big), \tag{2}$$

where $f_{t,d}$ is the frequency of term t in the document d, a_{t,λ_1} is the number of documents in the positive class containing the term t, and a_{t,λ_2} is the number of documents in the negative class containing the term t. The function $\max\left(1, a_{t,\lambda_2}\right)$ in the denominator allows that the term $tf{-}rf_{td}$ be not indefinite even if a_{t,λ_2} is zero.

According to [7], using this representation in different single-label data sets improves the performance of two-class based classifiers. For multi-class problems, [7] used a one-versus-all method.

Note that $tf{-}rf$ representation is for single-label problems and does not consider the frequency information of the term evaluated in other classes. That is, it only considers the relationship of the appearance of the term in the class under evaluation (that is, positive) versus all the other classes (that is, negative).

4 Our proposal for a new representation of multi-label datasets

On the one hand, $tf{-}idf$ as a representation of documents considers only the frequency of terms in the document (tf) and the frequency of terms in all documents (idf), disregarding the class or label to which the documents belong. On the other hand, $tf{-}rf$ also considers the frequency of terms in the document (tf) and the frequency of terms in all documents of the class under evaluation (rf). That is, in $tf{-}rf$, each document is represented by a different vector when assessing if it belongs to a particular class. From a theoretical point of view, this extension of the $tf{-}rf$ representation of text changes the representation of a document according to the label under evaluation, thereby achieving larger differences between documents belonging to different labels and thus harnessing the performance of binary classifiers. Thus, important information about the frequency in other classes is used, specially when frequency of the term shoes sharp variations as example in Table 2 shows.

Table 2. Example of frequency of a term for each label

	Label 1	Label 2	Label 3	Label 4	Label 5	Label 6	Label 7	Label 8	Label 9
Frequency	53	76	87	66	62	27	25	28	26

We propose the use of a centrality function $\mu{-}$Relevance Frequency of a Label, $tf{-}\mu rfl$, over the frequency of a term for each label, is derived from the term frequency and relevance frequency of a given label; as such, it constitutes a new representation based on $tf{-}rf$ for a multi-label problem.

$$tf{-}\mu rfl_{tdl} = f_{t,d} \log_2 \left(2 + \frac{a_{t,l}}{\mu(a_{t,\lambda_{j/l}})}\right), \tag{3}$$

where $\mu(a_{t,\lambda_{j/l}})$ is a function over the set $a_{t,\lambda_{j/l}} = \{a_{t,\lambda_1}, ..., a_{t,\lambda_{l-1}}, a_{t,\lambda_{l+1}}, ..., a_{t,|L|}\}$.

We will consider $\mu(a_{t,\lambda_{j/l}}) = \max\left(1, mean(a_{t,\lambda_{j/l}})\right)$ for $tf{-}rfl$ representation and $\mu(a_{t,\lambda_{j/l}}) = \max\left(1, median(a_{t,\lambda_{j/l}})\right)$ for $tf{-}rrfl$ representation. Such

functions give centrality measures, the mean is a classical metric and the median is a robust metric.

4.1 Relevance frequency of a label

Relevance frequency of a label, $tf{-}rfl$, is derived from the μ−Relevance Frequency of a Label, $tf{-}\mu rfl$; as such, it constitutes a new representation for a multi-label problem.

$$tf{-}rfl_{tdl} = f_{t,d} \log_2 \left(2 + \frac{a_{t,l}}{\max\left(1, mean(a_{t,\lambda_{j/l}})\right)}\right) \tag{4}$$

In equation 5, the term $mean(a_{t,\lambda_{j/l}})$ is the average number of documents containing the term t for each document labelled other than l.

4.2 Robust relevance frequency of a label

Robust relevance frequency of a label, $tf{-}rrfl$, also is derived from the μ−Relevance Frequency of a Label, $tf{-}\mu rfl$; as such, this is the second new representation for a multi-label problem.

$$tf{-}rrfl_{tdl} = f_{t,d} \log_2 \left(2 + \frac{a_{t,l}}{\max\left(1, median(a_{t,\lambda_{j/l}})\right)}\right) \tag{5}$$

The use of the *median* should yield more robust results in datasets containing large differences between the frequency of the occurrence of a term in a given set of labels versus other labels sets under evaluation.

4.3 Classification method

The proposed term-weighting methods includes information on the frequency of the occurrence of a term t in each set of documents labelled other than the label under evaluation. It is expected that $mean(a_{t,\lambda_{j/l}})$ and $median(a_{t,\lambda_{j/l}})$ will be higher if the term t appears more frequently in documents with label $\lambda_j = l$ than in documents with other labels $\lambda_{j/l}$, and they will be lower, in contrast, if the term t is more frequent in documents with labels other than l.

Our proposal is based on the $tf{-}rfl$ and $tf{-}rrfl$ representations and the SVM binary ensemble. It transforms the multi-label problem into a PT4 form [14], and then for each document d, the $tf{-}rfl$ and $tf{-}rrfl$ representations are derived for each label λ_j and classified using $|L|$ binary classifiers.

5 Experiments

The evaluation of the proposed $tf{-}rfl$ and $tf{-}rrfl$ representations was carried out using the Reuters-21578 Distribution 1.09. The Reuters-21578 data set consists of 21,578 Reuters newswire documents that appeared in 1987, less than half of which have human-assigned topic labels. The data set and the validation

mechanism used are the same as in [16], that is, the subsets of the k classes with the largest number of articles are selected for $k = 3, \ldots, 9$, resulting in seven different data sets denoted as First3, First4, ..., First9. Also, in this test 3-fold cross-validation is run ten times on each data set. Our classification method reports the average values among ten runs. Table 3 shows the data set characteristics.

Table 3. Characteristics of the pre-processed data set. Note that PMC denotes the percentage of documents belonging to more than one class and ANL denotes the average number of labels for each document

Data Set	Number of Classes	Number of Documents	Vocabulary Size	PMC	ANL
First3	3	7,258	529	0.74%	1.0074
First4	4	8,078	598	1.39%	1.0140
First5	5	8,655	651	1.98%	1.0207
First6	6	8,817	663	3.43%	1.0352
First7	7	9,021	677	3.62%	1.0375
First8	8	9,158	683	3.81%	1.0396
First9	9	9,190	686	4.49%	1.0480

First, we must transform the problem into a PT4 form, dividing the data into k input data sets for $k = 3, \ldots, 9$ binary classifiers, whereby each machine classifies one-against-others labels. Four representations were constructed from the data set, namely, the classical $tf{-}idf$ and $tf{-}rf$ representations and our proposed $tf{-}rfl$ and $tf{-}rrfl$ representations. An ensemble of binary SVM classifiers was used. Each machine employed a linear kernel; the parameters were optimised by maximising the classification margin between each pair of classes. The ensemble was implemented with LibSVM [1], where each machine worked with random sampling. Two-thirds of the examples were used for training, and one-third was used for testing. Note that all $tf{-}ifd$ representations are the same, regardless the label under evaluation, while the $tf{-}rf$, $tf{-}rfl$ and $tf{-}rrfl$ representations are different for each label.

Multi-label classification methods require different performance metrics than those used in traditional single-label classification methods. These measures can be grouped into bipartitions and rankings [15]. Since our method is not based on ranking, as in [11] and [16], the evaluation of the results in this research was performed using Hamming loss by considering bipartitions to evaluate how many times an instance-label pair was misclassified. This measure of error is defined as:

$$hloss(h) = \frac{1}{d}\sum_{i=1}^{d}\frac{1}{|L|}|h(\mathbf{x_i})\Delta Y_i|, \tag{6}$$

where $h(\mathbf{x_i})$ is the set of labels defined by the classifier for the documents, Y_i is the original labels of the documents and Δ is the difference between both. Performance is better when $hloss(h)$ is near 0.

Table 4 shows the different representations and their performance in term of Hamming loss. In this metric, for data set with fewer classes, Boostexter is better than $tf{-}rfl$ and $tf{-}rrfl$ for 0.00356 and 0.00218 respectively. For data set with more classes (namely, First5, First6, First7, First8 and First9), $tf{-}rfl$ is better than the other algorithms. Table 4 also shows that $tf{-}rrfl$ is better than the other algorithms for data sets with more classes (namely, the First4, First5, First6, First7, First8 and First9).

Table 4. Experimental results of SVM Ensembles with $tf{-}idf$, $tf{-}rf$, $tf{-}rfl$ and $tf{-}rrfl$ compared with others learning algorithms in terms of Hamming loss. Bp-MLL* and BoosTexter* as reported by [16]

Data set	First3	First4	First5	First6	First7	First8	First9	Average
SVM tf-idf	0.02797	0.02641	0.02590	0.02477	0.02246	0.02083	0.01981	0.02402
SVM tf-rf	0.02814	0.02687	0.02611	0.02522	0.02287	0.02118	0.02012	0.02436
SVM tf-rfl	0.02716	0.02590	0.02526	0.02412	0.02186	0.02026	0.01930	0.02341
SVM tf.rrfl	0.02578	**0.02478**	**0.02427**	**0.02321**	**0.02110**	**0.01958**	**0.01870**	**0.02249**
Bp-MLL*	0.0368	0.0256	0.0257	0.0271	0.0252	0.0230	0.0231	0.02664
BoosTexter*	**0.0236**	0.0250	0.0260	0.0262	0.0249	0.0229	0.0226	0.02446

To evaluate the results, as in [16] a test based on the two-tailed paired t-test at the 5 percent significance level was implemented. According to these results, SVM Ens $tf{-}rfl$ performs better than SVM Ens $tf{-}idf$ (4.2595×10^{-6}), SVM Ens $tf{-}rf$ (2.0376×10^{-7}) and Bp-MLL (3.74×10^{-2}). In addition, SVM Ens $tf{-}rrfl$ performs better than SVM Ens $tf{-}idf$ (2.5368×10^{-5}), SVM Ens $tf{-}rf$ (4.2013×10^{-6}) and Bp-MLL (1.63×10^{-2}). The p-value shown in parentheses provides a further quantification of the significance level. The results shown in Table 5 show the level of statistic significance as compared to alternative approaches with respect to Hamming loss. We can see that diferences between Boostexter have not statistical significance for data sets with fewer labels (First3, First4, First5), but for data sets with more labels (First6, First7, First8 and First9), Boostexter has the worst performance among all algorithms.

Table 5. Statistical analysis of results in terms of p-value on t-student test. NSS mean "Is Not Statistically Significant"

	SVM tf-rfl	SVM tf-rf	SVM tf-idf	Bp-MLL	BoosT.
SVM tf.rrfl	1.0754×10^{-4}	4.2013×10^{-6}	2.5368×10^{-5}	1.63×10^{-2}	NSS
SVM tf-rfl	-	2.0376×10^{-7}	4.2013×10^{-6}	3.74×10^{-2}	NSS
SVM tf-rf		-	4.2595×10^{-6}	NSS	NSS
SVM tf-idf			-	NSS	NSS
Bp-MLL				-	NSS

Finally, in Figure 1, we show how the different weighting methods discriminate when a term is important for a classifier or not. In this case, using $rrfl$ and rfl the term is weighted to high for labels 1, 2, 3, 4 and 5, and lower for

labels 6, 7, 8 and 9. Note that *idf* does not discriminate when evaluating each label and rf slightly discriminates.

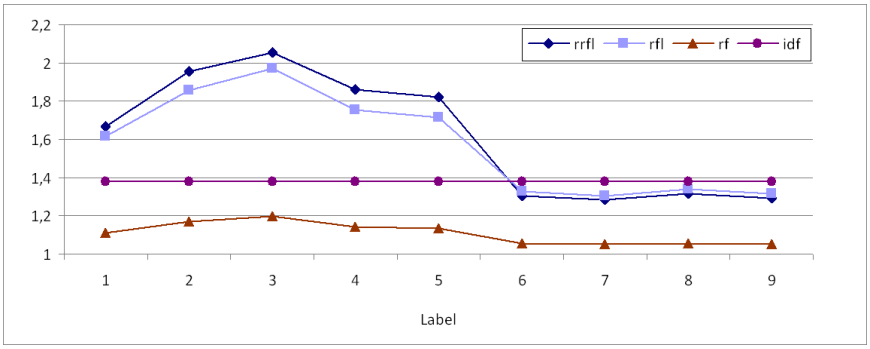

Fig. 1. Term-weights assigned by different representations for each label

6 Remarks and conclusions

Multi-label classification is an important topic in information retrieval and machine learning. Text representation and classification have been traditionally addressed using $tf{-}idf$ due to its simplicity and good performance. Changes in input representation can employ knowledge about the problem, a particular label, or the class to which the document belongs. Other representations can be developed for overcoming a particular problem directly, without transformation. New benchmarks should be used to validate the results; however, the preprocessing of multi-labelled texts must be standardised.

In this paper, we have presented the $tf{-}\mu rfl$ as a novel text representations for the multi-label classification approach. This proposal was assessed with two new input representation $tf{-}rfl$ and $tf{-}rrfl$. This representation considers the label to which the document belongs. Combining, this problem transformation with algorithm adaptation. The performance of this representation was tested in combination with an SVM ensemble using a known dataset. The results show statistically significant improvement as compared to alternative approaches with respect to Hamming loss. We believe that the contribution of the proposed multi-label representation is due to a better understanding of the problem under consideration. In future studies, we plan to compare our method to other $tf{-}idf$ representations and to investigate other label-dependent representations and procedures in order to reduce the dimension of the feature space depending on the relevance of each label.

Acknowledgement

This work has been partially funded by the Research Grants: Fondecyt 1110854 and Research Grant Basal FB0821 "Centro Científico Tecnólogico de Valparaíso"

References

[1] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM: a library for support vector machines*, 2001, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[2] Nello Cristianini, John Shawe-Taylor, and Huma Lodhi, *Latent semantic kernels*, Journal of Intelligent Information Systems **18** (2002), no. 2-3, 127–152.

[3] Eugene Fink, *Automatic evaluation and selection of problem-solving methods: Theory and experiments*, Journal of Experimental and Theoretical Artificial Intelligence **16(2)** (2004), 73–105.

[4] T. Joachims, *Learning to classify text using support vector machines – methods, theory, and algorithms*, Kluwer-Springer, 2002.

[5] M. Keikha, N.Sh. Razavian, F. Oroumchian, and H. S. Razi, *Document representation and quality of text: An analysis.*, In Survey of Text Mining II: Clustering, Classifcation, and Retrieval, Springer-Verlag, London, 2008, pp. 135–168.

[6] Man Lan, Chew-Lim Tan, and Hwee-Boon Low, *Proposing a new term weighting scheme for text categorization*, AAAI'06: Proceedings of the 21st national conference on Artificial intelligence, AAAI Press, 2006, pp. 763–768.

[7] Man Lan, Chew Lim Tan, Jian Su, and Yue Lu, *Supervised and traditional term weighting methods for automatic text categorization*, IEEE Transactions on Pattern Analysis and Machine Intelligence **31** (2009), 721–735.

[8] Edda Leopold and Jörg Kindermann, *Text categorization with support vector machines. How to represent texts in input space?*, Machine Learning **46** (2002), no. 1-3, 423–444.

[9] Christopher Manning and Hinrich Schutze, *Foundations of statistical natural language processing*, The MIT Press, 1999.

[10] Gerard Salton and Christopher Buckley, *Term-weighting approaches in automatic text retrieval*, Information Processing and Management: an International Journal **24** (1988), no. 5, 513–523.

[11] Robert E. Schapire and Yoram Singer, *Boostexter: A boosting-based system for text categorization*, Machine Learning, 2000, pp. 135–168.

[12] Fabrizio Sebastiani, *Machine learning in automated text categorization*, ACM Comput. Surv. **34** (2002), no. 1, 1–47.

[13] Evgeni Tsivtsivadze, Tapio Pahikkala, Jorma Boberg, and Tapio Salakoski, *Kernels for text analysis*, Chapter in Advances of Computational Intelligence in Industrial Systems **116** (2008), 81–97.

[14] G. Tsoumakas and I. Katakis, *Multi label classification: An overview*, International Journal of Data Warehouse and Mining **3** (2007), no. 3, 1–13.

[15] G. Tsoumakas, I. Katakis, and I. Vlahavas, *Mining multi-label data*, Data Mining and Knowledge Discovery Handbook, 2nd edition, O. Maimon, L. Rokach (Ed.), Springer, 2010.

[16] Min-Ling Zhang and Zhi-Hua Zhou, *Multilabel neural networks with applications to functional genomics and text categorization*, IEEE Transactions on Knowledge Data Engineering. **18** (2006), no. 10, 1338–1351.