Multimed Tools Appl (2013) 65:63–74
DOI 10.1007/s11042-012-1149-2

Relational large scale multi-label classification method for video categorization

Wojciech Indyk · Tomasz Kajdanowicz · Przemyslaw Kazienko

Published online: 19 June 2012

Abstract The problem of automated video categorization in large datasets is considered in the paper. A new Iterative Multi-label Propagation (IMP) algorithm for relational learning in multi-label data is proposed. Based on the information of the already categorized videos and their relations to other videos, the system assigns suitable categories—multiple labels to the unknown videos. The MapReduce approach to the IMP algorithm described in the paper enables processing of large datasets in parallel computing. The experiments carried out on 5-million videos dataset revealed the good efficiency of the multi-label classification for videos categorization. They have additionally shown that classification of all unknown videos required only several parallel iterations.

Keywords Multi-label classification · Relational learning · MapReduce · Classification in networks · Automated video categorization · Automated video tagging · Cloud computing · Parallel computing

1 Introduction

Over the last few years, multi-label classification methods for multimedia handling have been more and more expected. It was caused by a large number of areas, in

W. Indyk · T. Kajdanowicz · P. Kazienko (✉)
Wroclaw University of Technology, Wybrzeze Wyspianskiego 27,
50-370 Wroclaw, Poland
e-mail: kazienko@pwr.wroc.pl

W. Indyk
e-mail: wojciech.indyk@pwr.wroc.pl

T. Kajdanowicz
e-mail: tomasz.kajdanowicz@pwr.wroc.pl

which technological achievements resulted in explosion of structured data, particularly in multi-label data. Modern applications, such as semantic scene classification, music categorization and many others have had to be treated with new methods adapted accordingly. There can be found several traces of multi-label origins in machine learning literature but the first noticeable multi-label problem formulation appeared in [14]. Since that time a wide range of methods and techniques for multi-label classification has been proposed. In general, multi-label classification methods may be organized into two main categories, according to taxonomy proposed in [23]: problem transformation methods and algorithm adaptation methods. Whereas the former methods transform the multi-label classification problem either into one or more single-label classification, the latter provide specific learning algorithms in order to handle multi-label data directly.

Obviously, there can be proposed some other taxonomies for multi-label classification methods, such as with respect to the application area, the size of the output problem being solved (number of concurrent labels), the size of input space (number of input attributes) or the cost function being optimized. However, the great majority of them is not applicable for relational domains and cannot process really large datasets.

Nowadays, relations between objects are commonly modelled by different kinds of networks. For instance, a video can be linked to several other relevant videos. In such settlement, a network model becomes generic base for further, different types of processing and analyses. One of them is classification of network's nodes. It means that a node has to be assigned to one or more labels. This assignment may be accomplished by one of the classification methods, either by inference based on known profiles of these nodes (regular concept of classification) or using relational information derived from the network model. This second approach utilizes information about connections between nodes (structure of the network) and can be very useful in assigning labels to the nodes being classified. For example, it is very likely that a given video x is related to sport (label *sport*), if x is directly linked by many other videos about sport.

The strongest motivation behind usage of relational model is its ability to reflect relationships between correlated observations (videos). For example, in the network of videos it is possible to propagate information about the known categories of the known film to other unknown films linked from the given one. A new algorithm for video categorization is proposed in this paper. It takes advantage of the above distribution process with respect to the principle of relational influence propagation [2, 16, 20]. The realization of the algorithm stays in accordance to arising trend of data explosion in transactional systems, where enormous amount of data requires sophisticated analytical methods. There is a huge need to process big data in parallel—in clouds, especially in complex analysis like multi-label classification.

Iterative Multi-label Propagation (IMP) algorithm for relational learning in multi-label data, which is proposed and examined in the paper, facilitates processing on huge data. Section 2 covers related work while in Section 3 a proposal of MapReduce approach to relational large scale multi-label classification using label propagation in the network is explained. Section 4 contains the description of the experimental setup and obtained results. The paper is concluded in Section 5.

2 Related work

The most basic classification task—single-label classification—aims to assign an object (e.g. video) to exactly one class out of two or more possible classes. For example, a video can be categorized to exactly one of three classes: it is either (i) fully, (ii) partly or (iii) not at all about *sport*. The more sophisticated, multi-label classification, assigns an object to multiple classes simultaneously. it means that a video is classified to several categories, e.g. simultaneously to *sport*, *news and politics*, *gaming*, and *science*. Such set of four labels is an element of power set, i.e. all possible subsets of the label-set.

In order to accomplish the multi-label classification task, algorithms of two types have been introduced: problem transformation methods and algorithm adaptation methods. Among others the representatives of the first group are: Learning by Pairwise Comparison [7], Calibrated label ranking [8], Pruned sets [19], or RAkEL [22]. The second group of methods is represented by Bayesian multi-label classification [15], The Collective Multi-Label classifier (CML) and Collective Multi-Label with Features classifier (CMLF) [9], Ranking Support Vector Machine [6], Multi-label C4.5 decision tree [4] or Multi-label *k*-Nearest Neighbours [24].

The above mentioned methods either learn independent binary classifiers denoting the relevance of each class (especially problem transformation methods) or try to capture strong co-occurrence patterns and dependencies among the classes by modelling joint modes of labels or applying distinct cost functions. However, the most common approach assumes learning independent binary classifier for each class, and then infers the class labels irrespectively for each test instance. Some experiments have shown that such binary relevance classifiers are able to successfully handle multi-label data [12], especially with the simple label coding using Error Correcting Output Codes (ECOC).

Nevertheless, the mentioned above traditional machine learning techniques concentrate on identically and independently distributed data. This is not a case in real-world problems where data is relational in its nature and the important source of information is provided by the correlations reflected by the objects network structure. The recent research has focused on making use of the relational structure [17] or extended feature space [13] in order to improve the quality of prediction. The idea of multi-label classification based on the MapReduce concept was preliminary proposed in [10].

3 Relational large scale multi-label classification using MapReduce

The proposed Iterative Multi-label Propagation (IMP) algorithm for relational learning in multi-label data uses Markov random walk approach to process the information of labelled and unlabelled data represented as a graph. Recently, this idea has been applied to solve many problems, such as classification of partially labelled text [21], binary digits recognition [25], image annotation [1] or derivation of lexical relatedness between terms [18]. In general, it considers label probability distribution over the known nodes in the graph and propagates it to the unknown ones using connections between them.

In the paper, we adapt the general method proposed in [16] and introduce a new Iterative Multi-label Propagation algorithm. The algorithm assumes the accomplishment of multi-label inference by implementation of binary relevance approach. This means that each label is modelled individually in the Markov random walk. Therefore, each label from the set of possible labels (label-set) is modelled by the separate probability distribution over the known nodes. The solution of the algorithm is based on physical modelling of *harmonic energy minimization* introduced in [25]. The modelled function of relational influence propagation relay on the minimization of energy function depicted in (1).

Let $G(V, E, W)$ denote a graph with vertices-nodes V (a node is a video), arcs-edges $(i, j) \in E$ between pairs of nodes i, j, $i \neq j$, and an $n \times n$ arcs weight matrix W containing weights w_{ij} for each edge (i, j). Then, in such a graph, we have the energy ε for a given potential function f:

$$\varepsilon(f) = \frac{1}{2} \sum_{(i,j)\in E} w_{ij}(f(i) - f(j))^2 \tag{1}$$

where $f(\cdot)$—the potential of a node.

In the energy function (1), it is assumed that it converges when the labels probabilities are balances in the graph. The potential $f(\cdot)$ may be interpreted as the label probability, which is disseminated according to the distribution of edge weights in the graph structure.

According to [25] in a weighted graph G(V,E,W) with $n = |V|$ vertices,the label propagation may be solved by linear (2) and (3).

$$\forall i, j \in V \sum_{(i,j)\in E} w_{ij} P_i = \sum_{(i,j)\in E} w_{ij} P_j \tag{2}$$

$$\forall i \in V \sum_{c\in classes(i)} P_i = 1 \tag{3}$$

where P_i denotes the probability density of classes for node i.

Let assume the set of nodes V is partitioned into labelled V_L and unlabelled V_U vertices, $V = V_L \cup V_U$. Let P_u denote the probability distribution over the labels associated with vertex $u \in V$. For each node $v \in V_L$, for which P_v is known, a dummy node v' is inserted such that $w_{v'v} = 1$ and $P_{v'} = P_v$. This operation is equivalent to 'clamping' discussed in [25]. Let V_D be the set of dummy nodes. Then the solution of (2) and (3) can be performed according to Iterative Multi-Label Propagation, separately for each label, see Algorithm 1.

Algorithm 1 The pseudo code of Iterative Multi-label Propagation algorithm

1: **repeat**
2: **for all** $v \in (V \cup V_D)$ **do**
3: **for all** $\lambda \in lab\,el{-}set$ **do**
4: $P_v(\lambda) = \frac{\sum_{(u,v)\in E} w_{uv} P_u(\lambda)}{\sum_{(u,v)\in E} w_{uv}}$
5: **end for**
6: **end for**
7: **until** convergence

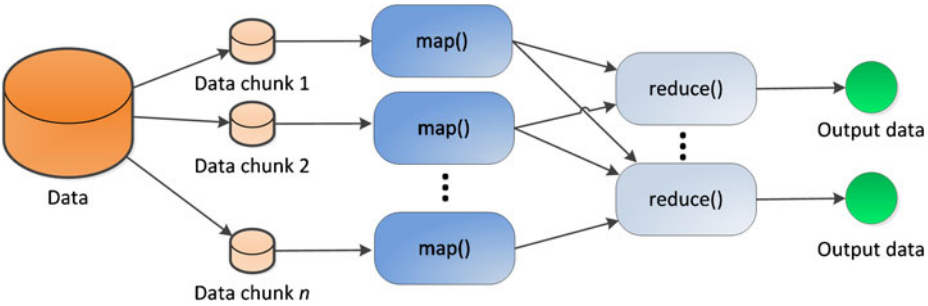

Fig. 1 The MapReduce programming model. This figure is excerpted and modified from [10]

It can be noticed, the appropriate label probability for each node is calculated in the loop (line 3 of the Algorithm 1). In this step, only the local information is only (more precisely—the neighbours u of the node v). Therefore, the calculation for the whole network can be performed in parallel using the MapReduce concept [5], as depicted in Algorithm 2. The single MapReduce iteration (the whole parallel Algorithm 2) replaces lines 2–6 in Algorithm 1. The general idea of MapReduce parallel computing is shown in Fig. 1.

Algorithm 2 The pseudo code of the single iteration using the MapReduce approach to Iterative Multi-label Propagation algorithm

1: $map < node; adjacencyList >$
2: **for all** $n \in adjacencyList$ **do**
3: propagate$< n; node.label(1), node.label(2), \ldots, node.label(|label\text{–}set|), n.weight >$
4: **end for**

1: $reduce < n, list(list(node.label), weight) >$
2: **for all** $\lambda \in label\text{–}set$ **do**
3: propagate$< n, \frac{\sum node.label(\lambda) \cdot weigth}{\sum weight} >$
4: **end for**

The MapReduce approach to Iterative Multi-label Propagation algorithm consists of two consecutive phases. The Map phase takes the graph structure: all labelled and dummy nodes, then propagates their labels according to adjacency list (the nearest neighbours) and with respect to the weights of edges. The Reduce phase collects labels and their edges' weights due to the key (here—a node) and calculates new labels. The output of the reduce phase and original adjacency list is the input for the map phase of the next iteration.

4 Experimental results

4.1 Dataset

In order to evaluate and demonstrate the proposed Iterative Multi-label Propagation algorithm the Youtube dataset [3] was utilized. The dataset was crawled using YoutubeAPI in 2008 and was partitioned into 58 chunks. There were used only these attributes from the original data that were required to create a graph structure and

the multi-label categorization: *video_id, age, category, related_IDs*. Using *related_IDs* the weighted graph structure was created. The weights were distributed equally among all adjacent videos, i.e. if there were 20 related videos each of them was linked by an edge with the weight of 0.05. The set was partitioned into training set and test set using the age of each video. All objects older than 950 days were assigned to training set, the rest to the test set. The basic features of utilized data set are presented in Table 1.

Depicted in Table 1 *AvgCard* measures the average number of labels associated to nodes (videos) in a given set, see (4).

$$AvgCard(D) = \frac{1}{|D|}\sum_{i=1}^{|D|}|Y_i| \tag{4}$$

where D denotes the video dataset and Y_i the label-set associated with ith node. The *density* measure is calculated according to (5):

$$density(D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i|}{|label-set|} \tag{5}$$

As it can be observed in (5) the density measure returns average fraction of the number of labels used to describe each of videos.

4.2 Results and discussion

Having 66% of nodes with the labels assigned and the graph with relations between nodes extracted, we can apply parallel MapReduce Algorithm 2 to the 34% of unknown nodes—videos. The processing was performed in the 30-nodes cloud from the WrUT Supercomputing Center—The PLATON Science Services Platform. One iteration in such environment took approximately 19.6 min.

As we can see in Fig. 2, the Hamming Loss measure is the smallest and the best after the fifth iteration. Simultaneously, the classification accuracy (Fig. 3) reaches its highest value after the third iteration. This reduction of classification quality after several first iterations is caused by the general idea of Algorithm 2. It propagates knowledge stored in the known nodes (learning set) and passes it over unknown nodes. However, after the third iteration 98.7% of nodes (the continuous line in Fig. 3) are already classified and accessing the rest of nodes takes many following iterations. It is caused by the structure of the data. Some nodes (videos) are linked very sparsely with the rest of the graph. It take many iterations to reach them starting from the known nodes. It should be noticed that at one iteration of Algorithm 2 only nearest neighbours of the known (already categorized) videos can be classified.

Simultaneously, after the third iteration more and more assignments for the known nodes are being changed by the algorithm decreasing the total accuracy. It

Table 1 Description of the basic features of the Youtube data set

Data	No. of videos	No. categories	AvgCard	Density	Distinct label sets
Training set	3,368,184 (66%)	15	1.1	0.06	35
Test set	1,733,756 (34%)	15	1.1	0.07	35

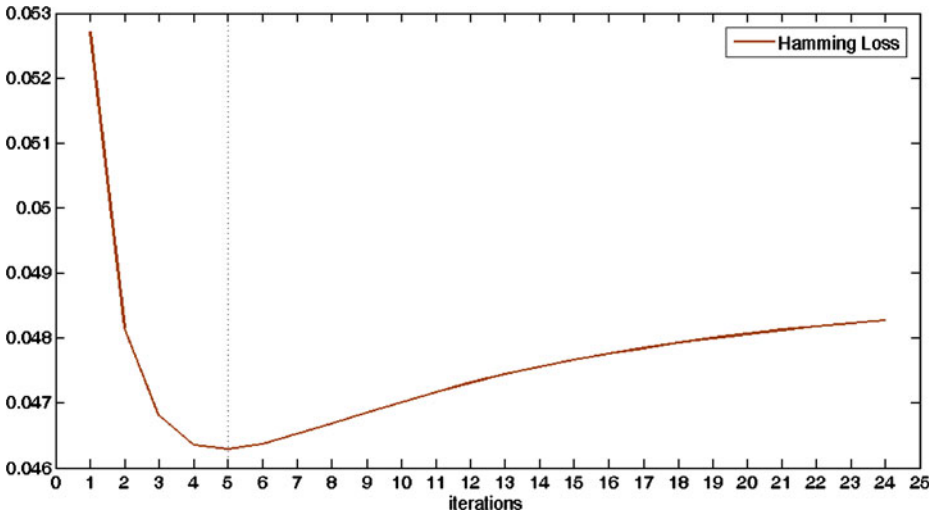

Fig. 2 Hamming Loss measure in consecutive iterations of the algorithm

means that the algorithm classifies the unknown nodes at a given iteration but these newly categorized nodes become known nodes for the following iterations. After the first iteration the contribution of categorized videos increased from 66% (Table 1) to 88.7% (Fig. 3). Note that already after the first three iterations the algorithm reaches most of its achievements: Hamming Loss = 0.04673 (Fig. 2), Accuracy = 47.5% (Fig. 3), the percentage of classified videos = 98.7% (Fig. 3). The multi

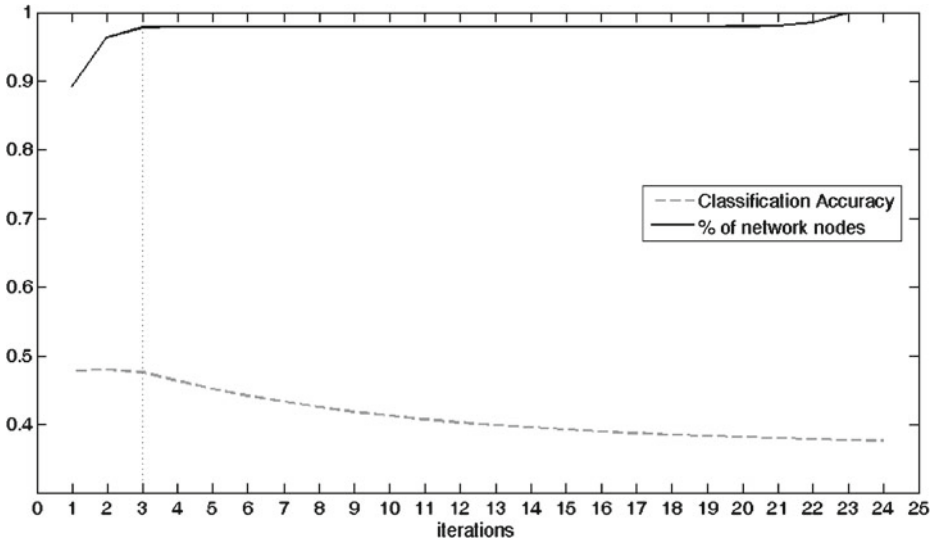

Fig. 3 Classification Accuracy and percentage of nodes reached in consecutive iterations of the algorithm

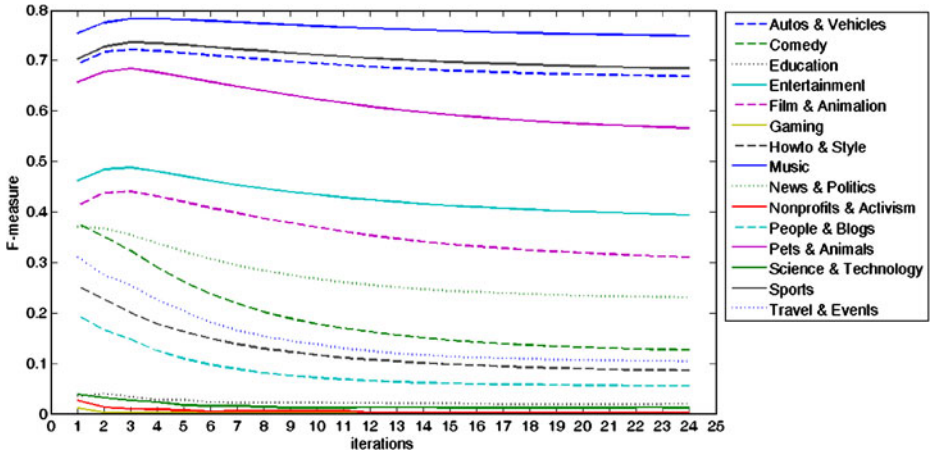

Fig. 4 F-measure results for distinct labels in consecutive iterations of the algorithm

categorization at the total level of almost 50% (accuracy) is not a bad result and not rarely achievements of 30% may be treated as good ones.

Analysis of results on individual labels separately (Fig. 4), provides interesting observations. On some labels Algorithm 2 provides very good results (like *Music*, *Autos & Vehicles*, *Sports*, *Pets & Animals*)—over 0.65 of F-measure. It means that relations between videos reflected by the *related_IDs* attribute and utilized by the algorithm are matched well by energy model from (1), (2) and (3). Additionally, these kinds of categories are easier to be precisely recognized by humans creating the *related_IDs* attributes.

On the other hand, there are some labels like *Nonprofits & Activism*, *Gaming*, and *Science & Technology* which tend to occur in the isolated way—the movies categorized with these labels pretty often do not have neighbours with the same label so if they are unknown they cannot inherit the proper labels from their neighbours. It is additionally enhanced by the relatively small number of labels (categories) assigned to a single video—only 1.1 in average (see *AvgCard* column in Table 1).

Classification accuracy at the level of nearly 50% in such environment where relations (the *related_IDs* attribute) not necessarily link videos with the same label should be treated as a very good result.

5 Conclusions

A new method for multi-label categorization of videos for large-scale datasets performed by means of the MapReduce paradigm is proposed in the paper. Using parallel computing enables processing large-scale datasets in the efficient way. The idea of multi-label categorization consists in iterative propagation of known label-sets over the relations linking videos. No other information except relations and known multi-labels (multiple categories) of some videos is necessary to categorize the rest of films. The other video profiles (attributes) were not used for the purpose of classification.

The experiments carried out on over 5 million of videos crawled from the YouTube service have revealed that MapReduce parallel processing may be very efficient. Besides, only few iterations (about 3) are needed to reach the best accuracy at the level of almost 50%. Additionally, categorization of some labels is more accurate, while for the others it is hard to achieve good results. It comes mostly from the nature of the relations between videos existing in the data.

The diverse classification accuracy results obtained for individual labels could be improved by modified video crawling process, according to concept presented in [11] using labels' distribution.

Acknowledgements The method presented in the paper is an extended description of algorithm proposed in [10] presented at The 4th International Workshop on Engineering Knowledge and Semantic Systems, IWEKSS 2012.

This work was partially supported by The Polish National Center of Science the research project 2011–2012, 2011–2014 and Fellowship co-financed by The European Union within The European Social Fund.

The authors are grateful to Wroclaw Networking and Supercomputing Center for granting access to the computing infrastructure built in the project No. POIG.02.03.00-00-028/08 "PLATON—Science Services Platform".

References

1. Azran A (2007) The rendezvous algorithm: multiclass semi-supervised learning with markov random walks. In: Proceedings of the International Conference on Machine Learning (ICML). ACM, pp 49–56
2. Chakrabarti S, Dom B, Indyk P (1998) Enhanced hypertext categorization using hyperlinks. In: Proceedings of SIGMOD-98, ACM international conference on management of data, pp 307–318
3. Cheng X, Dale C, Liu J (2008) Statistics and social network of youtube videos. In: 16th International Workshop on Quality of Service, IWQoS 2008, pp 229–238
4. Clare A, King R (2001) Knowledge discovery in multi-label phenotype data. In: PKDD 2001, Lecture Notes in Computer Science, vol 2168. Springer, pp 42–53
5. Dean J, Ghemawat S (2004) Mapreduce: simplified data processing on large clusters. In: Proceedings of the 6th conference on symposium on opearting systems design & implementation. USENIX Association, Berkeley, pp 10–24
6. Elisseeff A, Weston J (2001) A kernel method for multi-labelled classification. In: NIPS. MIT Press, pp 681–687
7. Furnkranz J (2002) Round robin classification. J Mach Learn Res 2:721–747
8. Furnkranz J, Hullermeier E, Loza-Mencia E, Brinker K (2008) Multilabel classification via calibrated label ranking. Mach Learn 73(2):133–153
9. Ghamrawi N, McCallum A (2005) Collective multi-label classification. In: CIKM05. ACM, pp 195–200
10. Indyk W, Kajdanowicz T, Kazienko P, Plamowski S (2012) Mapreduce approach to collective classification. In: ICAISC 2012. Lecture Notes in Computer Science, vol 7267, pp 656–663
11. Jung JJ (2012) Evolutionary approach for semantic-based query sampling in large-scale information sources. Inf Sci 182(1):30–39
12. Kajdanowicz T, Kazienko P (2012) Multi-label classification using error correcting output codes. Int J Appl Math Comput Sci (in press)
13. Kazienko P, Kajdanowicz T (2012) Label-dependent node classification in the network. Neurocomputing 75(1):199–209

14. Lin J, Ligomenides P, Lo S, Freedman M, Mun S (1994) Hybrid neural-digital computer-aided diagnosis system for lung nodule detection on digitized chest radiographs. In: Proceedings of the IEEE symposium on computer-based medical systems. IEEE, pp 207–212
15. McCallum A (1999) Multi-label text classification with a mixture model trained by em. In: Proceedings of the AAAI' 99 workshop on text learning
16. Neville J, Jensen D (2000) Iterative classification in relational data. In: Proc. AAAI-2000 workshop on learning statistical models from relational data. AAAI Press, pp 13–20
17. Peters S, Jacob Y, Denoyer L, Gallinari P (2012) Iterative multi-label multi-relational classification algorithm for complex social networks. Soc Netw Anal Min 2:17–29
18. Rao D, Yarowsky D (2009) Ranking and semi-supervised classification on large scale graphs using map-reduce. In: Proceedings of the 2009 workshop on graph-based methods for natural language processing, association for computational linguistics, TextGraphs-4, pp 58–65
19. Read J (2008) A pruned problem transformation method for multi-label classification. In: Proceedings of the New Zealand computer science research student conference, Christchurch, New Zealand, pp 143–150
20. Slattery S, Mitchell T (2000) Discovering test set regularities in relational domains. In: Proceedings of the International Conference on Machine Learning (ICML). Morgan Kaufmann, pp 895–902
21. Szummer M, Jaakkola T (2001) Clustering and efficient use of unlabeled examples. In: Proceedings of Neural Information Processing Systems (NIPS), vol 14
22. Tsoumakas G, Vlahavas I (2007) Random k-labelsets: An ensemble method for multilabel classification. Lecture Notes in Artificial Intelligence, vol LNAI 4701. Springer, pp 406–417
23. Tsoumakas G, Katakis I, Vlahavas I (2010) Mining multi-label data. In: Maimon O, Rokach L (eds) Data mining and knowledge discovery handbook. Springer, pp 667–685
24. Zhang M, Zhou Z (2005) A k-nearest neighbor based algorithm for multi-label classification. In: IEEE International conference on granular computing, vol 2. The IEEE Computational Intelligence Society, pp 718–721
25. Zhu X, Ghahramani Z, Lafferty J (2003) Semi-supervised learning using gaussian fields and harmonic functions. In: ICML, pp 912–919

Wojciech Indyk is a student at Wroclaw University of Technology, Poland. He has been engaged in multiple research projects, including research project on multi-label classification for the debt portfolio valution. His research interest include multi-label classification, relational machine learning and network analysis. He co-authored four scientific papers and participated in the 7th International Summer School on Pattern Recognition (ISSPR), Plymouth, UK, in 2011. Recently, he serves as the president of "Data and exploration" student's scientific group at Wroclaw University of Technology.

Tomasz Kajdanowicz received his M.Sc. degree from Wroclaw University of Technology, Poland in 2008. Currently he is finalizing PhD studies at Wroclaw University of Technology with expected PhD defence in 2012. He serves as and researcher in various research projects and simultaneously acts as an independent consultant working with IT companies in Poland. He was an organizing-chair of international workshops MMAML'11 and MMAML'12. He has already served as a member of international programme committees and the reviewer for international journals and scientific conferences. His research areas focus on social network analysis, machine learning and hybrid artificial inteligence systems as well as their applications, especially in the industry. While participating in multiple research and development projects, he collaborates with leading financial enterprises in Poland. He authored more than thirty scientific papers and articles.

Przemyslaw Kazienko received his MSc and PhD degrees in computer science with honours, both from Wroclaw University of Technology, Poland, in 1991 and 2000, respectively. He obtained his habilitation degree from Silesian University of Technology, Poland, in 2009. Recently, he serves as a professor of Wroclaw University of Technology at the Institute of Informatics, Poland. He was also a Research Fellow at Intelligent Systems Research Centre, British Telecom, UK in 2008. For several years, he has held the position of the deputy director for development at Institute of Applied Informatics. He was a co-chair of international conference WSKS'11 and workshops RAAWS'05, RAAWS'06, MMAML'10, MMAML'11, MMAML'12, SNAA'11, SNAA'12, CSNA'12 and a Guest Editor of New Generation Computing, International Journal of Applied Mathematics and Computer

Science, Journal of Universal Computer Science and International Journal of Computer Science & Applications. He regularly serves as a member of international programme committees and the reviewer for prestige international journals and scientific conferences. He is a member of Editorial Board of International Journal of Knowledge Society Research, International Journal of Human Capital and Information Technology Professionals, as well as Social Informatics. He has authored over 130 scholarly and research articles on a variety of areas related to multiple model classification, collective classification, social network analysis and application, knowledge management, collaborative systems, data mining, recommender systems, Information Retrieval, data security, and XML. He also initialized and led over 25 projects chiefly in cooperation with commercial companies, including large international corporations.