A Pure URL-based Genre Classification of Web Pages

Chaker Jebari
Information Technology Department
Ibri College of Applied Sciences
Sultanate of Oman
jebarichaker@yahoo.fr

***Abstract*— In this paper, we propose a new approach for multi-label genre classification of web pages that exploits character n-grams extracted from the URL of the web page rather than its content. Using only the URL reduces the time needed for feature extraction since it does not need to download the content of the web page. Our approach deals with the complexity of web pages because it uses a multi-label classification where each web page can be assigned to more than one genre. Moreover, our approach implements a new weighting technique that exploits the structure of the URL. Experiments conducted on a known multi-label dataset show that our approach achieves encouraging results.**

Keywords—web page; multi-label classification; genre; URL structure

I. INTRODUCTION

As the World Wide Web continues to grow exponentially, the classification of web pages becomes increasingly important in web searching. Web page classification, assigns a web page to one or more predefined classes. According to the type of the class, the classification can be divided into sub-problems: topic classification, sentiment classification, genre classification, and so on. Currently, search engines use keywords to classify web pages. Returned web pages are ranked and displayed to the user, who is often not satisfied with the result. For example, searching for the keyword "Java" will provide a list of web pages containing the word "Java" and belonging to different genres such as "tutorial", "exam", "Call for papers", etc. Therefore, web page genre classification could be used to improve the retrieval quality of search engines [15]. For instance, a classifier could be trained on existing web directories and be applied to new pages. At query time the user could be asked to specify one or more desired genres so that the search engine would returns a list of genres under which the web pages would fall.

However, although potentially useful, the concept of "genre" is difficult to define and genre definitions abound. According to [19], the genres found in web pages (also called cyber-genres) are characterized by the triple <content, form, functionality>. The content and form attributes are common to non-digital genres and refers to the text and the layout of the web page respectively. The functionality attribute concerns exclusively digital genres and describes the interaction between the user and the web page.

A web page is a complex object that is composed of different sections belonging to different genres. For example, a conference web page contain information on the conference, topics covered, important dates, contact information and a list of hypertext links to related information. This complex structure need to be captured by a multi-label classification scheme in which a web page can be assigned to multiple genres [17, 24, 6].

This paper has made the following contributions:

- We propose a multi-label classification scheme that fit the real environment in which each web page may fall in more than one genre.
- We used character n-grams extracted only from the URL of the web page. Using only the URL we eliminate the necessity of downloading the web page. It is very useful when the web page content is not available or need more time/space to display.
- In contrast to other genre classification studies, we used a new weighting technique that exploits the URL structure.

The remainder of the paper is organized as follows. Section 2 reviews previous works on genre classification of web pages. Section 3 describes the multi-label classification. Section 4 describes how a web page is represented using only character n-grams extracted from the URL. A new segment-oriented weighting technique is also presented at the end of Section 4. We evaluated the performance of our approach in Section 5. Finally, Section 6 concludes our paper with future research directions.

II. RELATED WORKS ON WEB PAGE GENRE CLASSIFICATION

The previous works on genre classification of web pages differ according to two independent criterion: i) features used to represent the web pages and ii) the classification method.

Many types of features have been proposed for automatic genre classification of web pages. These features can be grouped into four groups. The first group refers to surface features, such as function words, genre specific words, punctuation marks, document length, etc. The second group concerns structural features, such as part-of-speech (POS), tense of verbs, etc. The third group is presentation features, which mainly describe the layout of document. Most of these features concerns HTML documents and cannot be extracted from plain text documents. Among these features we quote the number of specific HTML tags and links. The last group of features is often extracted from metadata elements (URL, description, keywords, etc.) and concerns only structured documents. Once a set of features has been obtained it is necessary to choose a classification method, which are often based on machine learning techniques. Broadly speaking, classification methods can be divided into two main categories: single-label and multi-label methods. In single label methods, a document is associated to only one label, whereas, in multi-label methods, a document is assigned to a set of labels. As stated in the introduction, the need for attributing more than one genre label to a web page is noticed by few authors such as Santini [17], Jebari [5] and Vidulan et al. [24]. In her study, Santini [17] has implemented a zero-to multi-genre classification scheme, where a web page can be associated to zero or multiple genres. Using SVM classifier and a single-label corpus of 1400 web pages equally distributed across 7 genres, Santini reported an accuracy of 0.91. While, using the corpus KI-04[1], she reported an accuracy of about 0.7.

In his PhD thesis, Jebari [5] proposed a flexible, incremental, refined and combined approach for genre classification of web pages. The proposed approach exploits the features extracted from three different sources which are: the URL addresses, the title tag, the heading tags and the hypertext links. The experiments conducted on the two known corpora KI-04 and WebKB[2] provides a micro-averaged BEP more than 80%. Moreover, the experiments show that combining all features gave better results than using each feature separately. In comparison with other single label classifiers, Jebari shows that his approach is very fast.

Vidulan et al. [24], presents a multi-class transformation, where each combination of genres is labeled with a single distinct label. Using the AdaBoost classifier and a multi-labeled genre corpus, they achieved a very low F-measure of 0.35. Hence they concluded that their approach failed to properly address multi-genre web pages.

In this section we present some recent works on genre classification of web pages, rather than text genre classification. For each study, we describe the features and the corpus used and the achieved results.

Meyer and Stein [18] used different kinds of features including presentation features (i.e. HTML tag frequencies), classes of words (names, dates, etc.), frequencies of punctuation marks and POS tags. To evaluate their work, Meyer and Stein compiled the KI-04 corpus composed of 800 web pages distributed equitably over 8 genres. Using the corpus KI-04 and a discriminate analysis classifier, they achieved an accuracy of 0.7.

Lim et al. [12] investigated the usefulness of information found in different parts of the webpages (URL, title, body, anchor text etc.). Based on a corpus of 15 genres and using a K-Nearest-Neighbor classifier, they reported that the main body and anchor text information is the most effective.

Kennedy and Shepherd [9] used three sets features to firstly discriminate between home pages from non-home pages. Secondly, they classify home pages into three categories (personal, corporate, and organization). Their feature set comprises features about the content (e.g., common words, Meta tags), form (e.g., number of images), and functionality (e.g., number of links, use of JavaScript). The best reported results were for personal home pages.

Boese and Howe [3] examined the effects of Web page evolution on the task of classifying Web pages by genre. In their study, they exploited the URL and other HTML tags. Using the WebKB dataset and a logistic regression classifier they reported an accuracy of 0.8.

Vidulin et al. [25] used all features used in previous studies. They used 2491 features divided into four groups: surface, structural, presentation and context features. Surface features include function words, genre-specific words, sentence length and so on. Structural features include part-of-speech tags, sentence types and so on. Presentation features describe the formatting of a document through the HTML tags. While context features describe the context in which a web page was found (e.g. URL, hyperlinks, etc.). Vidulin et al. compiled the multi-label corpus MGC[3] containing 1539 web pages belonging to 20 genres. Based on this corpus, they reported a precision of about 0.65.

Kanaris and Stamatatos [8] used character n-grams to identify the genre of web pages. Using the corpus MGC and the SVM method, they reported micro-averaged recall of 0.55, a micro-averaged precision of 0.74 and a micro-averaged F1 of 0.6.

During her thesis study, Mason [13] proposed a centroid-based classifier to classify web pages using n-grams extracted from the textual content. The classifier proposed by Mason builds a genre centroid profile by combining the profiles of training web pages belonging to that genre. Using the MGC corpus, Mason reported an averaged F-measure of 0.84.

More recently, Myriam and David [16] proposed a new genre classification of web pages which is purely based on URL. Their approach is based on the combination of different character n-grams of different lengths. Using an SVM learning technique, they achieved a micro-averaged F1 of 0.46 for Santini corpus.

III. MULTI-LABEL CLASSIFICATION

In traditional single-label classification, a classifier is built and trained using a set of examples associated with just one single label l of a set of disjoint labels L, where |L|>1. Moreover, in multi-label classification, the examples can be associated with a set of labels $Y \subseteq L$. In the literature, different methods have been proposed to be applied to multi-label classification problems. These methods are grouped into two main categories: problem transformation and algorithm transformation [20].

[1] http://www.webis.de/research/corpora

[2] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo- 20/www/data

[3] http://dis.ijs.si/mitjal/genre/

Problem transformation methods are algorithm independent and transforms a multi-label learning problem into one or more single-label learning problems. The most popular transformation method is called Binary Relevance (BR) that builds L binary classifiers, one for each different label l. For the classification of a new instance, BR outputs the union of the labels l_i that are positively predicted by the L classifiers [22]. A simpler and less common problem transformation method is called Label Power Set (LP) [22]. LP method considers each unique set of labels that exists in a multi-label training set as one of the labels of a new single-label classification task. Given a new instance, the single-label classifier of LP outputs the most likely label, which is actually a set of labels [14]. Exploiting the LP method, Random k-labelsets method (RA*k*EL) constructs an ensemble of LP classifiers [21]. Each LP classifiers is trained using a different small random subset of the set of labels. An average decision is calculated for each label, and the final decision is positive for a given label if the average decision is larger than a given threshold.

The algorithm transformation methods extend existing learning algorithms to deal with multi-label data directly. Several transformation methods have been proposed in the literature such as BR-SVM, MLKNN and BPMLL.

BR-SVM is an improvement of the SVM classifier [7]. This improvement concerns the margin of SVMs in multi-label classification problems. BR-SVM improves the margin by i) removing very similar negative training instances which are within a threshold distance from the learnt hyper-plane, and ii) removing negative training instances of a complete class if it is very similar to the positive class, based on a confusion matrix that is estimated using any fast and moderately accurate classifier on a held out validation set.

The most recent adapted algorithm called MLKNN has been proposed by Zhang and Zhou [26]. This algorithm transforms the original data set into |L| data sets D_l that contain all examples of the original data set, labeled with the label l if the example belongs to l and $\neg l$ otherwise. After that, MLKNN applies the KNN algorithm for each label λ_j and considers those that are labeled at least with the label λ_j as positive and the rest as negative. MLKNN can be extended to produce a ranking of the labels as an output.

BPMLL extends basic back-propagation algorithm by introducing a new global error function that captures the characteristics of multi label learning [27].

IV. Web page representation

The representation of a web page is the main step in automatic genre classification. The first paragraph of this section describes the extraction of features from the URL and the second paragraph presents a new Weighting technique that exploits the URL segments.

A. Feature extraction

Often, features for classifying web pages are extracted from its content, which needs more time since it requires downloading it previously [1]. To deal with this issue, we decided in this paper to represent a web page by its URL, since every web page possesses a URL, which is a relatively small string (therefore easy to handle). A URL can be divided into the following segments: Domain Name, Document Path, Document Name and Query string [2].

For example for the URL: *http://www.math.rwth-aachen.de/~Greg.Gamble/cv.pdf*, we can extract the following segments:

- Domain name (DOMN): www.math.rwth-aachen.de
- Document path (DOCP) : ~Greg.Gamble
- Document name and query string (DOCN): cv.pdf

For each URL segment we performed some pre-processing, which is consist into:

- Removing special characters (_,.,:,?,$,%) and digits.
- Removing common words (for example the word “www” from the domain name and the words “pdf”, “html”, etc. from the document name)
- Removing generic top-level domains (.edu, .uk, .org, .com, etc.) from the domain name
- Removing one-character words.

After that we extracted from each URL string all character n-grams. We only consider 2-grams, 3-grams and 4-grams as candidate *n*-grams since they can capture both sub-word and inter-word information and keep the dimensionality of the problem in a reasonable level.

B. Structure-Oriented Weighting Technique

Term Frequency does not exploit the structural information present in the URL. For exploiting URL structure we must consider not only the number of occurrences of character n-gram in the URL but also the URL segment the character n-grams are present in. The idea of the proposed weighting technique is to assign greater importance to character n-grams that belong to the URL segment that are more suitable for representing a web page. To implement this idea, we proposed a new weighting technique, named “SWT”. In this technique, the weight for a given character n-gram C_i in a URL U_j is defined as follows:

$$SWT(C_i, U_j) = \sum_s W(s) \cdot TF(C_i, s, U_j)$$

Where

- $TF(C_i, s, U_j)$ denotes the number of times the character n-gram C_i occurs in the segment *s* of the URL *Uj*.
- *W(s)* is the weight assigned to the segment s and is defined as follows:

$$W(s) = \begin{cases} \alpha & if \quad s = DOMN \\ \beta & if \quad s = DOCP \\ \lambda & if \quad s = DOCN \end{cases}$$

Where the values of the weighting parameters α, β and λ are determined using an experimental study.

V. Evaluation

A. Corpus

In our approach we used the corpus MGC [25] (See Table 1). For the best of my knowledge, MGC is the only multi-label genre corpus available at the moment.

TABLE I. COMPOSITION OF MGC CORPUS

Genre	# web pages	Genre	# web pages
Blog	83	Index	308
Adult	79	Informative	318
Children’s	113	Journalistic	206
Commercial/Promotional	193	Official	85
Community	82	Personal	133
Content Delivery	207	Poetry	76
Entertainment	126	Prose Fiction	75
Error Message	90	Scientific	98
FAQ	71	Shopping	81
Gateway	119	User Input	96

B. Experimental setup

The experimentation of our method is conducted using four different multi-label classification methods (RA*k*EL, BR-SVM, MLKNN and BPMLL) presented in section 3. These methods are implemented in the Mulan toolkit[4].

The evaluation of multi-label classifiers requires different evaluation metrics from those used in single-label classifiers [4]. In this study, we used Hamming Loss, One-Error, Ranking Loss, Coverage and Micro-averaged precision metrics. Due to the small number of web pages in each genre, we followed the 3-cross-validation procedure.

VI. Results and discussion

In this section we describe the conducted experiments and show and discuss obtained results.

A. Experiment1

The objective of the first experiment is to measure our approach without exploiting the URL structure. For this purpose we fixed the values of all weighting parameters to 1. Table 2 reports the experimental results achieved using the classifiers RA*k*EL, BR-SVM, MLKNN and BPMLL using character n-grams of length between 2 and 4 extracted from the URL. The best result on each metric is shown in bold face. The value following ± gives the standard deviation.

TABLE II. CLASSIFICATION PERFORMANCE PROVIDED BY DIFFERENT MULTI-LABEL CLASSIFIERS USING CHARACTER N-GRAMS OF LENGTH BETWEEN 2 AND 4.

	Ra*k*EL	**BR-SVM**	**MLKNN**	**BPMLL**
HamLoss	0.083±5.866	0.085±0.003	**0.081**±2.105	0.917±6.674
OneError	0.905±0.029	0.767±0.037	**0.693**±0.011	0.936±0.016
RankLoss	0.550±0.013	0.488±0.024	**0.304**±0.006	0.519±0.033
Coverage	12.346±0.191	11.273±0.441	**7.689**±0.231	11.750±0.588
Micro-Precision	**0.851**±0.120	0.478±0.064	0.704±0.084	0.083±6.674

To make a clear view of the performance between two classifiers C1 and C2, a partial order “>” is defined for each evaluation metric. Based on this order, the notation $C1>C2$ means that the performance of classifier C1 is statistically better than that of classifier C2 on the specific metric (based on two-tailed paired *t*-test at 5% significance level). In order to give an overall performance of a classifier, a score is assigned to it. This score is calculated for each evaluation metric and for each possible pair of classifiers C1 and $C2$. If $C1>C2$, then a positive score +1 and a negative score -1 are assigned to C1 and C2 respectively. For each classifier, the accumulated score on all evaluation metrics, gave a total order ">" for this classifier. The partial and total order in terms of each evaluation metrics using different multi-label classifiers is summarized in Table 3.

TABLE III. PERFORMANCE ORDER BETWEEN EACH MULTI-LABEL CLASSIFIER USING CHARACTER N-GRAMS OF LENGTH BETWEEN 2 AND 4 IN TERMS OF EACH EVALUATION METRIC.

	RA*k*EL:C1, BR-SVM:C2, MLKNN:C3, BPMLL:C4
HamLoss	C1>C2, C1<C3, C1>C4, C2<C3, C2>C4, C3>C4
OneError	C1<C2, C1<C3, C1>C4, C2<C3, C2>C4, C3>C4
RankLoss	C1>C2, C1<C3, C1<C4, C2<C3, C2>C4, C3>C4
Coverage	C1<C2, C1<C3, C1<C4, C2<C3, C2>C4, C3>C4
Micro-Precision	C1>C2, C1<C3, C1>C4, C2<C3, C2>C4, C3>C4
Total Order	**MLKNN(13)>BR-SVM(-1)>RA*k*EL(-3)>BPMLL(-10)**

As shown in the above table, we can say that the classifier MLKNN achieves the best results with respect to all experimentation metrics, followed by BR-SVM, Rakel and BPMLL.

B. Experiment2

The purpose of this experiment is to compare the performance achieved by MLKNN classifier using words and character grams of length between 2 and 4. In this experiment, we fixed the values of all weighting parameters to 1. The obtained results are shown in Table 4.

TABLE IV. PERFORMANCE OF MLKNN CLASSIFIER USING DIFFERENT FEATURES IN TERMS OF DIFFERENT EVALUATION METRICS

	Words	**2gram**	**3gram**	**4gram**	**2-4gram**
HamLoss	0.082±0.001	0.082±9.494E-4	**0.081**±0.001	**0.081**±9.886E-4	**0.081**±2.105
OneError	0.697±0.002	0.740±0.212	0.704±0.024	0.708±0.009	**0.693**±0.011
RankLoss	**0.296**±0.009	0.328±0.006	0.300±0.010	0.311±0.004	0.304±0.006
Coverage	**7.554**±0.245	8.180±0.057	7.620±0.227	7.825±0.057	7.689±0.231
Micro-Precision	0.625±0.041	**0.732**±0.138	0.690±0.107	0.644±0.057	0.704±0.084

To give an overall performance on all features, we calculated the partial and the total orders as explained in the previous experiment (See Table 5). It is clear from the Table 5 that using character grams of length between 2 and 4, we achieved better results than using words and each character n-grams separately.

[4] http://mulan.sourceforge.net/index.html

TABLE V. PERFORMANCE ORDER USING DIFFERENT FEATURES EXTRACTED FROM THE URL IN TERMS OF EACH EVALUATION METRIC.

	Words:W, 2-gram:G2, 3-gram: G3, 4-gram: G4, 2-4 gram: G24
HamLoss	W<G3, W<G4, W<G24, G2<G3, G2<G4, G2<G24
OneError	W>G2, W>G3, W>G4, W<G24, G2<G3, G2<G4, G2<G24, G3>G4, G3<G24, G4<G24
RankLoss	W>G2, W>G3, W>G4, W>G24, G2<G3, G2<G4, G2<G24, G3>G4, G3>G24, G4<G24
Coverage	W>G2, W>G3, W>G4, W>G24, G2<G3, G2<G4, G2<G24, G3>G4, G3>G24, G4<G24
Micro-Precision	W<G2, W<G3, W<G4, W<G24, G2>G3, G2>G4, G2>G24, G3>G4, G3<G24, G4<G24
Total Order	**G24(8) > G3(6) > W(3) > G4(-6) > G2(-11)**

C. Experiment3

In this experiment we will test our weighting technique SWT with different values of weighting parameters. The results reported using MLKNN and character grams of length between 2 and 4 are illustrated in Table 6.

TABLE VI. CLASSIFICATION PERFORMANCE FOR DIFFERENT VALUES OF THE WEIGHTING PARAMETERS USING MLKNN CLASSIFIER AND ALL CHARACTER N-GRAMS BETWEEN 2 AND 4

α	**β**	**Λ**	**HamLoss**	**OneError**	**RankLoss**	**Coverage**	**Micro-P**
1	0	0	0.084	0.780	0.357	8.737	0.435
0	1	0	0.084	**0.726**	**0.340**	**8.453**	**0.419**
0	0	1	**0.083**	0.783	0.372	9.129	0.513
1	1	1	0.081	0.693	0.304	7.689	0.704
1	2	3	0.082	0.708	0.303	7.689	0.627
2	1	3	0.083	0.731	0.309	7.723	0.536
2	3	1	**0.081**	**0.687**	**0.300**	**7.606**	**0.706**
3	1	2	0.081	0.711	0.307	7.700	0.679
3	2	1	0.081	0.713	0.308	7.726	0.698
1	3	2	0.082	0.700	0.300	7.607	0.647

It is clear from this table that the segment DOCP captures more information about the genre of the web page than the segments DOCN and DOMN. In our experiment, the best result is reported using the values of 2, 3 and 1 for the weighting parameters α, β and λ respectively.

VII. CONCLUSION AND FUTURE WORKS

In this paper, we suggested a multi-label genre classification of web pages, which is more suitable for the complexity of web pages because it can assign a web page more than one genre. With regards to classification features, our method uses character n-grams extracted only from the URL of the web page. Moreover, our method uses a new weighting technique based on URL segmentation. Conducted experiments using a multi-label corpus show that our method provides encouraging results. In the future, we plan to evaluate our approach using big multi-label and multi-lingual corpora.

REFERENCES

[1] E. Baykan, M. Henzinger, L. Marian, and I. Weber, "Purely URL based topic classification," *In the Proceedings of the 18th International Conference on World Wide Web,* Madrid, Spain, 2009.

[2] T. Berners-Lee, RT. Fielding, and L. Masinter, "*Uniform Resource Identifier (URI): Generic Syntax",* Internet Society. RFC 3986; STD 66; 2005, 1998.

[3] E. S. Boese and A. E. Howe, "Effect of web document evolution on genre classification", *In Proceedings of the 14th ACM International conference on Information and knowledge Management*, 2005.

[4] R. Cerri, R.R. Silva, and A. C. Carvalho, "Comparing Methods for Multilabel Classification of Proteins Using Machine Learning Techniques" . BSB 2009, LNCS 5676. 109-120, 2013.

[5] C. Jebari, "A new Centroid-based Approach for Genre Categorization on Web Pages", *Journal of Language Technology and Computational Linguistics*, Volume 24. Number 1. 73-96, 2009.

[6] C. Jebari and A. Wani, "A Multi-label and Adapative Genre Classification of Web Pages", *In Proceedings of ICMLA 2012, 20102*.

[7] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classication", *In Proceedings of PAKDD '04*. 22-30, 2004.

[8] I. Kanaris and E. Stamatatos, "Learning to recognize webpage genres", *Information Processing and Management Journal*, 45(5): 499-512, 2009.

[9] M. Kennedy and M. Shepherd, "Automatic Identification of Home Pages on the Web", *Proc. of the 38th Hawaii International Conference on System Sciences, 2005*.

[10] B. Kessler, G. Nunberg, and H. Schütze, "Automatic detection of text genre", *Proceedings of the 35th ACL/8th EACL*, 32-38, 1997.

[11] Y. Kim, and S. Ross, "Examining Variations of Prominent Features in Genre Classification", *In Proceedings of HICSS Conference, 2008.*

[12] C.S. Lim, K.J. Lee, and G.C. Kim, "Multiple Sets of Features for Automatic Genre Classification of Web Documents", *Information Processing and Management.* 41(5). 1263-1276, 2005.

[13] J. Mason, "*An n-gram-based Approach to the Automatic Classification of Web Pages by Genre*", Ph.D. Dissertation, Dalhousie University, Canada, 2009.

[14] A. McCallum, "Multi-label text classification with a mixture model trained by em". *Proc of AAAI' 99 Workshop on Text Learning, 1999.*

[15] Z. E. Meyer, "*On Information Need and Categorizing Search",* Ph.D. Dissertation. Paderborn University, Germany, 2007.

[16] A. Myriam and W. A. David, "*What's in a URL? Genre Classification from URLs*", Intelligent Techniques for Web Personalization and Recommender Systems. AAAI Technical Report. WS-12-09, 2012.

[17] M. Santini, "*Automatic identification of genre in web pages",* Ph.D. Dissertation. Brighton University, UK, 2007.

[18] B. Stein and Z. E. Meyer, "Retrieval Models for Genre Classification", Scandivian Journal of Information Systems (SJIS), 20:1, pp. 91-117. 2008.

[19] M.A. Shepherd and C. Watters, "Evolution of cybergenre", In Proceedings of the 31nd Hawaiian International Conference on System Sciences, 1998.

[20] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview", *International Journal of Data Warehousing and Mining*, 3(3):1-13, 2007.

[21] G. Tsoumakas and I.P. Vlahavas, "Random k -Labelsets: An Ensemble Method for Multilabel Classification", *In Proceedings of ECML.* 406-417, 2007.

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

[23] G. Tsoumakas, I. Katakis and I. Vlahavas, "Random k-Labelsets for Multi-Label Classification", *IEEE Transaction on Knowledge and Data Engineering, 2010*.

[24] V. Vedrana, L. Mitja and G. Matjaž, "Multi-Label Approaches to Web Genre Identification", *Journal of Language and Computational Linguistics*, 24(1):97-114, 2009.

[25] V. Vidulin, M. Luštrek and M. Gams, "Using Genres to Improve Search Engines", *1st International Workshop: Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, (pp. 45-51). Borovest, Bulgaria, 2007.

[26] M. L. Zhang and Z.H. Zhou, "A k-nearest neighbor based algorithm for multi-label classification", *In Proceedings of the 1st IEEE International Conference on Granular Computing (GrC'05)*. 718-721, 2005.

[27] M. L. Zhang and Z.H. Zhou, "Multi-label neural networks with applications to functional genomics and text categorization", *IEEE Transactions on Knowledge and Data Engineering 18*. 1338–135, 2006.