

A Segment-based Weighting Technique for URL-based Genre Classification of Web Pages

Chaker Jebari

Abstract—We propose a segment-based weighting technique for genre classification of web pages. This technique exploits character n-grams extracted from the URL of the web page rather than its textual content. The main idea of our technique is to segment the URL and assigns a weight for each segment. Experiments conducted on three known genre datasets show that our method achieves encouraging results.

Index Terms—URL, genre classification, web page, segment weight.

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 [18]. 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. Generally speaking, a genre is a category of artistic, musical, or literary composition characterized by a particular style, form, or content, but more specialized characterizations have been proposed [23]. For instance, [14] defined a genre as a bundle of facets, focusing on different textual properties such as brow, narrative, and genre. According to [25], 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 [23, 28, 9].

A broad number of studies on genre classification of web documents have been proposed in the literature [23]. These studies differ mainly with respect to the feature set they use to represent web documents. These features are divided into four types: surface features (function words, genre specific words, punctuation marks, document length, etc), structural features (Parts Of Speech (POS), Tense of verbs, etc), presentation features (number of particular HTML tags and links) and contextual features (URL, keywords, etc.).

Once a set of features has been obtained, it is necessary to choose a classification algorithm. These are often based on machine learning techniques such as Naïve Bayes, K-Nearest Neighbour, Decision trees, Support Vector Machine, Neural Networks, and Centroid-based techniques [20].

It is worth noting that many researchers study the usefulness of URL for genre classification of web documents [21, 22, 28] without deeply exploiting its content and the structure. With respect to this, we proposed in this paper a new approach that represents a web page by a bag of character n-grams (contiguous n-characters) extracted only from the URL. 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. Moreover, we proposed in this paper a new weighting technique that exploits the URL structure.

The remainder of the paper is organized as follows. Section 2 present previous works on genre classification of web pages. Section 3 describes the extraction of character n-grams from the URL. A new segment-oriented weighting technique is also presented at the end of Section 3. The evaluation of our approach is described in Section 4. Finally, Section 5 concludes our paper with future research directions.

Manuscript received on October 9, 2015, accepted for publication on January 30, 2016, published on June 25, 2016.

Chaker Jebari is with College of Applied Sciences, Ibri, Sultanate of Oman (e-mail: jebarichaker@yahoo.fr).

II. PREVIOUS WORK ON GENRE CLASSIFICATION OF WEB PAGES

The previous works on genre classification of web pages differ with respect to the following three factors:

- (1) The list of genres used in the evaluation, called also genre palette
- (2) The features used to represent the web page
- (3) The classification method used to identify the genre of a given web page

The last factor is often based on machine learning techniques such as Naïve Bayes, *K*-Nearest Neighbour, Decision trees, Support Vector Machine, Neural Networks, and Centroid-based techniques. These machine-learning techniques were deeply studied by Mitchell [20]; therefore, we will focus more on the first two factors.

A genre palette can be general if it covers a large part of web or specific if it aims to study a limit number of genres. To study the usefulness of web genres, Meyer and Stein [19] compiled the KI-04 corpus which is composed of 1205 web pages distributed over 8 genres (article, download, link collection, private portrayal, non-private portrayal, discussion, help and shop) (See Table 4). Based on a corpus of 15 genres, Lim et al. [16] investigated the usefulness of information found in different parts of the web page such (title, body, anchor, etc.). Kennedy and Shepherd [13] proposed a more specific and hierarchal genre palette that contains two super-genres (home pages, non-home pages). The super-genre home page is divided into two three sub-genres (personal, corporate and organization pages). In her master thesis, Boese [3] collected a corpus of 343 web documents distributed across 10 structured genres (abstract, call for papers, FAQ, How-to, Hub/sitemap, Job description, Resume/CV, Statistics, Syllabus, Technical paper). Santini [23] compiled manually a corpus of 1400 web pages equally distributed across seven genres (blogs, eshops, FAQs, front pages, listings, personal home pages, search pages). To examine the effects of web evolution on the task of classifying web pages by genre, Boese and Howe [4] used the dataset WebKB [5], which contains 8282 web pages. In this dataset, 4518 web pages belong to one of six functional genres (course, department, faculty homepage, project, staff homepage, and student homepage). The remaining 3764 web pages are assigned to the genre “*Other*.” To evaluate a multi-label genre classification schema, Vidulin et al. [27] used the web site Google Zeitgeist to build the multi-label corpus 20-genre. This corpus contains 1539 web pages belonging to 20 genres (See Table 5). More recently, Priyatam et al. [22] investigated the usefulness of URL to classify web pages into two categories (health and tourism). For this reason, they collected and tagged manually 3000 web pages.

Many types of features have been proposed for automatic genre categorization. These features can be grouped on 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 Parts 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.

With respect to plain text document representation, Kessler et al. [14] have used four types of features to classify the Brown corpus by genre. The first types are structural features, which includes counts of functional words, sentences, etc. The second types are lexical features, which includes the existence of specific words or symbols. The third kinds of features are character level features, such as punctuation marks. The last kind concerns derivative features, which are derived from character level and lexical features. These four features sets can be grouped on two sets, structural features and surface features. Karlgren [12] have used twenty features: count of functional words, POS count, textual count (e.g. the count of characters, the count of words, number of words per sentence, etc.), and count of images and links. Stamatos et al. [24] identified genre based on the most English common words. They have used the fifty most frequent words on the BNC corpus and the eight frequent punctuation marks (period, comma, colon, semicolon, quotes, parenthesis, question mark, and hyphen). Dewdney et al. [6] have adopted two features sets: BOW (Bag of Words) and presentation features. They used a total of 89 features including layout features, linguistic features, verb tenses, etc. Finn and Kushmerick [7] used a total of 152 features to differentiate between subjective vs. objective news articles and positive vs. negative movie reviews. Most of these features were the frequency of genre-specific words.

With respect to web page representation, Meyer and Stein [19] used different kinds of features including presentation features (i.e. HTML tag frequencies), classes of words (names, dates, etc.), and frequencies of punctuation marks and POS tags. Lim et al. [16] introduced new sets of features specific to web documents, which are extracted from URL and HTML tags such as title, anchors, etc. First, Kennedy and Shepherd [13] used three sets features to 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).

Vidulin et al. [27] used 2,491 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.).

Kim and Ross [15] used image, style, and textual features to classify PDF documents by genre. The image features were extracted from the visual layout of the first page of the PDF document. The style features are represented by a set of genre-prolific words, while textual features are represented by a bag of words extracted from the content of the PDF document. Kim and Ross pointed out that some PDF are textually inaccessible due to password protection, and that image features would be especially useful in this case.

In his PhD thesis, Jebari [8] 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 show that combining all features gave better results than using each feature separately. Kanaris and Stamatatos [11] used character n -grams and HTML tags to identify the genre of web pages. They stated that character n -grams are language-independent and easily extracted while they can be adapted to the properties of the still evolving web genres and the noisy environment of the web. In her thesis study, Mason [17] used character n -grams extracted from the textual content to identify the genre of a web page. Recently, Myriam and David [21] proposed a new genre classification of web pages that is purely based on URL. Their approach is based on the combination of different character n -grams of different lengths. More recently, Priyatam et al. [22] used character n -grams extracted from the URL to classify Indian web pages into two genres: sport and health. They aim to improve the Indian search engine Sandhan.

III. 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 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 words with one character.

After that, we extracted from each word all character n -grams. For example, from the word “*JAVA*” we can extract one character 4grams (*JAVA*), two character 3-grams (*JAV*, *AVA*) and 3 character 2-grams (*JA*, *AV*, *VA*).

To reduce the time needed for training and testing, we removed the words and character n -grams that appear in less than 10 web page URLs.

B. Segment-based 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, which is more suitable to represent a web page. To implement this idea, we proposed a new weighting technique termed SWT. In this technique, the weight for a given character n -gram C_i in a URL U_j is defined as follows:

$$SWT(C_i, U_j) = \sum_s W(s) \cdot TF(C_i, s, U_j) \tag{1}$$

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 U_j
- $W(s)$ is the weight assigned to the segment s and is defined as follows:

$$W(s) = \begin{cases} \alpha & \text{if } s = DOMN \\ \beta & \text{if } s = DOCP \\ \lambda & \text{if } s = DOCN \end{cases} \tag{2}$$

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

IV. EVALUATION

A. Datasets

To evaluate our approach, we used three datasets: KI-04, 20-Genre and KRYS-I. These datasets are unbalanced, which means that the web documents are not equally distributed among genres.

1) KI-04

This dataset was built following a palette of eight genres suggested by a user study on genre usefulness. It includes 1295 web pages, but only 800 web pages (100 per genre) were used in the experiment described in Meyer and Stein [19]. In the experiments described in this paper, I have used 1205 web pages because we have excluded empty web pages and error messages; see Table I.

TABLE I
COMPOSITION OF KI-04 DATASET

Genre	# of web pages
Article	127
Download	151
Link collection	205
Private portrayal	126
Non-private portrayal	163
Discussion	127
Help	139
Shop	167

2) 20-Genre

This dataset, gathered from internet, consists of 1539 English web pages classified into 20 genres as shown in the following table. In this dataset, each web page was assigned by labelers to primary, secondary, and final genres. Among 1539 web pages, 1059 are labeled with one genre, 438 with two genres, 39 with three genres and 3 with four genres [28]; see Table II.

TABLE II
COMPOSITION OF 20-GENRE DATASET

Genre	# web pages	Genre	# web pages
Blog	77	Gateway	77
Children	105	Index	227
Commercial	121	Informative	225
Community	82	Journalistic	186
Content delivery	138	Official	55
Entertainment	76	Personal	113
Error message	79	Poetry	72
FAQ	70	Adult	68
Shopping	66	Prose fiction	67
User input	84	Scientific	76

3) KRYS-I

This dataset was built between 2005 and 2008 by Kim and Ross [15]. It consists of 6494 PDF documents labeled independently by two kinds of people (students and secretaries). Each document was assigned to 1, 2 or 3 genres.

A set of 70 genres has been defined, which can be classified into 10 groups (Book, Article, Short Composition, Serial, Correspondence, Treatise, Information Structure, Evidential Document, Visually Dominant Document and Other Functional Document). After removing inaccessible documents, we obtain 5339 documents (See Table III).

TABLE III
COMPOSITION OF KRYS-I DATASET

Genre	# docs	Genre	# docs
Resume/CV	124	Menu	102
Speech transcript	119	Comics	110
Poems	64	Essay	305
Chart	154	Catalog	143
Technical manual	168	Artwork	65
Memo	146	Abstract	155
Dramatic script	57	Financial record	130
Sheet music	39	Miscellaneous report	307
Research article	323	Fact sheet	228
Advertisement	103	Slides	132
Periodicals	128	Interview	107
Project description	130	Manual	237
Magazine article	254	Product description	206
Thesis	200	Forum discussion	124
Bibliographical sketch	113	Receipt	101
Letter	177	Technical report	256
Poetry book	121	Journals	139
Table calendar	108	Handbook	365
Diagram	149	Project proposal	129
Raw data	179	Minutes	116
Regulations	156	Form	271
Telegram	13	Book of fiction	24
Operational report	218	List	100
Legal proceedings	129	Contract	58
Questionnaire	135	Guideline	223
Other research article	475	Other book	62
Email	120	Slips	20
Review	198	Announcement	79
Conference proceedings	148	Appeal propaganda	40
Graph	114	Fictional piece	48
Poster	160	Order	42

B. Experimental setup

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. Moreover, to keep the dimensionality of the problem in a reasonable level, we removed character n -grams that appear in less than 5 web page URLs. Table IV shows the number of words and character n -grams extracted from the datasets KI-04, 20-Genre and KRYS-I.

TABLE IV
NUMBER OF WORDS AND 2-4 GRAMS EXTRACTED
FROM KI-04, 20-GENRE AND KRYS-I DATASETS

	# words	# 2grams	# 3grams	# 4grams
KI-04	120	120	117	229
20-Genre	139	185	210	317
KRYS-I	230	203	236	345

The experimentation of our approach is conducted using Naïve Bayes, IBk, J48, and SMO classifiers implemented in the Weka toolkit. Due to the small number of web pages in each genre, we followed the 3-cross-validation procedure, which consists of randomly splitting each dataset into three equal parts. Then we used two parts for training and the remaining one part for testing. This process is performed three times and the final performance in terms of micro-averaged accuracy is the average of the three individual performance figures.

It is worth noting that in this study we evaluated our method using English web pages. Moreover, since character n-grams are language independent, our method can be used to classify non-english web pages.

C. Results and discussion

To evaluate our approach, we conducted two experiments. The aim of the first experiment is compare the classification accuracy obtained using words and character n-grams. While, the second experiments aims to identify the more suitable values of weighting parameters used to achieve the best accuracy.

1) Experiment 1

In this experiment we evaluated our approach using two kinds of features: words and 2-4 grams. Note that in this experiment the weighting parameters are equal to 1. Table V illustrated the achieved results for different datasets and machine learning techniques.

TABLE V
CLASSIFICATION ACCURACY USING DIFFERENT MACHINE LEARNING
TECHNIQUES AND EXPLOITING WORDS AND CHARACTER N-GRAMS

	Words			2-4 grams		
	KI-04	20-Genre	KRYS-I	KI-04	20-Genre	KRYS-I
NB	0.71	0.64	0.63	0.72	0.73	0.68
KNN	0.75	0.55	0.62	0.87	0.88	0.82
SVM	0.77	0.81	0.73	0.87	0.88	0.84
DT	0.64	0.62	0.59	0.66	0.70	0.57

It is clear from Table V that using 2-4 grams we can achieve better results than using words. However, the best result is reported by the SVM classifier followed by KNN, Naïve bayes and then decision trees (DT). Using character 2-4 grams, the SVM achieves the highest accuracy values of 0.87, 0.88 and 0.84 for KI-04, 20-Genre and KRYS-I datasets respectively.

As shown in Table V, the overall accuracy differs from dataset to another. Using the KRYS-I dataset we obtained the lowest accuracy, while the 20-Genre dataset achieves the highest accuracy. This performance variation is due to the huge number of overlapping genres and generic web pages.

As noted by Meyer and Stein [19], multi-genre web pages are web pages where two or more genres overlap without

creating a specific and more standardized genre. For example in KI-04 dataset, the genres shop and portrayal overlap. In KRYS-I dataset, the genres manual, technical manual and guideline overlap. The 20-genre dataset reports the highest accuracy because it contains few number of overlapping genres and generic web pages. Therefore, it is clear that using datasets with a big number of generic or noise web pages and overlapping genres reduces the classification performance.

As illustrated in the Table V, the SVM reported the highest accuracy due to its less over-fitting and its robustness to noise web pages. However, SVM method is more complicated and runs slowly [10, 26].

2) Experiment 2

This experiment aims to identify the appropriate values of the weighting parameters in order to achieve the best performance. In this experiment we used SVM as a machine learning technique and 2-4 grams as a classification features. The reported results are shown in Table VI.

TABLE VI
PERFORMANCE WITH DIFFERENT CONFIGURATIONS

α	β	λ	KI-04	20-Genre	KRYS-I
0	0	1	0.77	0.56	0.66
0	1	0	0.85	0.89	0.82
1	0	0	0.79	0.85	0.80
1	1	1	0.87	0.88	0.84
1	2	3	0.62	0.68	0.84
1	3	2	0.65	0.72	0.86
2	1	3	0.66	0.73	0.70
2	3	1	0.88	0.90	0.86
3	1	2	0.73	0.88	0.84
3	2	1	0.71	0.72	0.78

As shown in Table VI, it is clear that the segment DOCP captures more information about the genre of the web page than the segments DOCN and DOMN. Therefore, assigning the highest weight to the DOCP segment, followed by DOCN segment, then DOMN segment achieves the best result. From our experiment, the best result is reported using the values of 2, 3 and 1 for the weighting parameters α , β and λ , respectively.

V. CONCLUSION AND FUTURE WORK

In this paper, we suggested a method for genre classification of web pages that uses character n-grams extracted only from the URL of the web page. Hence, our method eliminates the necessity of downloading the content of the web page because. Moreover, our method uses a new weighting technique based on URL segmentation. Conducted experiments using three known datasets show that our method provides encouraging results. As future work, we plan to deal with generic and overlapping genres by proposing a multi-label classification where a web page can be assigned to more

than one genre. Moreover, we plan to evaluate our approach using multi-lingual datasets with large number of examples.

ACKNOWLEDGMENT

Author would like to thank the anonymous reviewers for their suggested comments to improve the quality of this manuscript.

REFERENCES

- [1] E. Baykan, M. Henzinger, L. Marian, I. Weber, "Purely URL based topic classification," in *Proceedings of the 18th International Conference on World Wide Web*. Madrid, Spain, 2009.
- [2] T. Berners-Lee, R. T. Fielding, L. Masinter, "Uniform Resource Identifier (URI): Generic Syntax," *Internet Society*. RFC 3986, STD 66.
- [3] E. S. Boese, (2005). *Stereotyping the web: Genre Classification of Web Documents*. M.Sc. Dissertation. Colorado State University, USA, 1998.
- [4] E. S. Boese, A. E. Howe, "Effects of web document evolution on genre classification," in *Proceedings of the CIKM'05*, 2005.
- [5] M. Craven, D. DiPasque, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery, "Learning to extract symbolic knowledge from the World Wide Web," in *Proceeding of the 15th national / 10th conference on artificial intelligence / innovative applications of artificial intelligence*, Madison, 1998.
- [6] N. Dewdney, C. Vaness-Dikema, and R. Macmillan, "The form is the Substance: Classification of Genres in Text," in *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and 10th Conference of the European Chapter of the Association for Computational Linguistics*, Toulouse, France, 2001.
- [7] A. Finn and N. Kushmerick, "Learning to classify documents according to genre," in *Proceedings of the Workshop Doing it with Style: Computational Approaches to Style Analysis and Synthesis*, held in conjunction with IJCAI 2003, Acapulco, Mexico, 2003.
- [8] C. Jebari, *Une nouvelle approche de catégorisation flexible et incrémentale de pages web par genres*. Ph.D. Dissertation, Tunis El Manar University, Tunisia, 2008.
- [9] C. Jebari, W. Arif, "A Multi-label and Adaptive Genre Classification of Web Pages," in *Proceedings of ICMLA*, 2012.
- [10] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," in *Proceedings of 10th European Conference on Machine Learning*, 1998.
- [11] I. Kanaris and E. Stamatatos, "Learning to recognize webpage genres," *Information Processing and Management Journal*, 45 (5) 499–512.
- [12] J. Karlgren, "Stylistic experiments in information retrieval," in *Natural Language Information Retrieval* (ed. T. Strzalkowski), pp. 147–166, 1999.
- [13] M. Kennedy, Shepherd, "Automatic Identification of Home Pages on the Web," in *Proc. of the 38th Hawaii International Conference on System Sciences*, 2005.
- [14] B. Kessler, G. Nunberg, and H. Schütze, "Automatic detection of text genre," in *Proceedings of the 35th ACL / 8th EACL*, 32–38, 1997.
- [15] Y. Kim and S. Ross, "Examining Variations of Prominent Features in Genre Classification," in *Proceedings of HICSS Conference*, 2008.
- [16] C. S. Lim, K. J. Lee, G. C. Kim, "Multiple Sets of Features for Automatic Genre Classification of Web Documents," *Information Processing and Management*. 41 (5) 1263–1276, 2005.
- [17] J. Mason, *An n-gram-based Approach to the Automatic Classification of Web Pages by Genre*. Ph.D. Dissertation, Dalhousie University, Canada, 2009.
- [18] Z. E. Meyer, *On Information Need and Categorizing Search*. Ph.D. Dissertation. Paderborn University, Germany, 2007.
- [19] Z. E. Meyer, B. Stein, "Genre classification of web pages: User study and feasibility analysis," in *Proceedings KI 2004: Advances in Artificial Intelligence*. pp. 256–269, 2004.
- [20] T. Mitchell, *Machine learning*. McGraw-Hill, 1997.
- [21] M. Abramson, D. W. Aha, "What's in a URL? Genre Classification from URLs. Intelligent Techniques for Web Personalization and Recommender Systems," *AAAI Technical Report*. WS-12-09, 2012.
- [22] P. N. Priyatam, S. Iyengar, K. Perumal, and V. Varma, "Don't Use a Lot When Little Will Do: Genre Identification Using URLs," in *14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013)*, University of the Aegean, Samos, Greece; *Research in Computing Science* 70, pp. 233–243, 2013.
- [23] M. Santini, *Automatic identification of genre in web pages*. Ph.D. Dissertation. Brighton University, UK, 2007.
- [24] E. Stamatatos, N. Fakotakis, G. Kokkinakis, "Text Genre Detection Using Common Word Frequencies," in *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrücken, Germany, 2000.
- [25] M. A. Shepherd, C. Watters, "Evolution of cybergenre," in *Proceedings of the 31st Hawaii International Conference on System Sciences*, 1998.
- [26] V. Vapnik, *The Nature of Statistical Machine Learning Theory*. Springer, 1995.
- [27] V. Vidulin, M. Luštrek, M. Gams, "Using Genres to Improve Search Engines," in *1st International Workshop: Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, Borovets, Bulgaria, pp. 45–51. 2007.
- [28] V. Vidulin, M. Luštrek, M. Gams, "Multi-Label Approaches to Web Genre Identification," *Journal of Language and Computational Linguistics*, 24 (1) 97–114, 2009.
- [29] I. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2005.