International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:2, 2007

A New Approach for Flexible Document Categorization

Jebari Chaker, and Ounelli Habib

***Abstract*—**In this paper we propose a new approach for flexible document categorization according to the document type or genre instead of topic. Our approach implements two homogenous classifiers: contextual classifier and logical classifier. The contextual classifier is based on the document URL, whereas, the logical classifier use the logical structure of the document to perform the categorization. The final categorization is obtained by combining contextual and logical categorizations. In our approach, each document is assigned to all predefined categories with different membership degrees. Our experiments demonstrate that our approach is best than other genre categorization approaches.

***Keywords*—**Categorization, combination, flexible, logical structure, genre, category, URL.

I. Introduction

WITH the increase of web documents, it is very difficult to retrieve desired information quickly out of the documents retrieved by a search engine. To improve the search quality, many works propose to classify documents according to their topics [20]. Even if the documents are classified successfully by their subjects, they stayed heterogeneous. For example, the documents grouped by the topic "cinema" can be an actor homepage, a newspaper about a film or an actor, a collection of films posters and so on. So, a user looking for newspapers about cinema should consult all other document types or genres. Therefore, the document genre or type is considered as another categorization criterion. Many definitions of document genre have been proposed in the past [1, 9, 15, 12, 6]. Common to all is that document genre is another document view orthogonal to document topic, say, documents having the same subject can be of different genres. Studies on document genre categorization have been started with Biber works [1, 2, 3]. Since, many studies have been published [4, 6, 8, 9, 10, 12, 14, 17, 19] which most of them concerns textual documents and not suitable for web documents.

In automatic genre categorization a document is represented by a set of features. The values of these features, which extracted from a training corpus, can be used to classify a new document. Selecting suitable features is the core of each genre categorization approach.

Web documents are different from textual document because they have additional proprieties such as, URL and HTML tags.

In our approach we suggest to use the URL to perform contextual categorization and HTML tags to extract the logical structure of the document and then perform logical categorization.

A web document has two kinds of structures: internal structure, which is represented by title and hn tags and external structure, which is represented by hyperlinks. The logical structure is the combination of internal and external structures.

All works on genre categorization are still rely on single-label categorization, witch means that a document is assigned to only one category [17]. In this context, we propose a flexible approach that implements a centroid-based categorization algorithm. Our algorithm assigns an individual document to all predefined categories with different membership degrees.

This paper is organized as follows. In the second section, we explain the proposed approach. In section 3, we discuss experimentations results. Finally, we present our contributions and the future research works.

II. Proposed Approach

Our approach is based on category centroid, where documents are represented using the vector-space model [16]. In this model, each document d is represented by a tfidf vector d_{tfidf} = (tf_1 log(N/df_1), …, tf_i log(N/df_i), …, tf_n log(N/df_n)), where tf_i is the frequency of the ith term in the document, df_i is the number of documents that contain the ith term and N is the number of training documents. For a category c, the centroid is represented by the average for all vectors for the positive examples for this category:

$$C = \frac{1}{|c|} \cdot \sum_{\boldsymbol{d} \in c} \boldsymbol{d}$$

The idea of our approach is to compute the centroid vectors of all categories. So, if you have k categories, this leads to k centroid vectors $\{C_1, \ldots, C_i, \ldots, C_k\}$, where C_i is the centroid for the ith category. For a new document x, our approach compute the similarities between x and all k centroids using the cosine measure. The document x will be assigned to the category corresponding to the most similar. The category of the document x is given by:

$$\arg\max_{i=1,\ldots,k}\left(\cos(x, C_i)\right)$$

Jebari Chaker is a lecturer in King Saud University, college of computer and information sciences, computer science department, and also a PhD candidate in ERPAH research group in Tunis El'Manar University.

Ounelli Habib is a professor in Tunis El'Manar University, Faculty of sciences, computer science department.

Where, the cosine between x and C_i is calculated as follow:

$$\cos(x, C_i) = \frac{x \cdot C_i}{\|x\|_2 \times \|C_i\|_2}$$

Our approach implements contextual and logical classifiers. Both classifiers are based on category centroid model explained above.

A. Contextual Classifier

In this kind of classifier, each document is represented by his URL. Each URL will be processed to remove special characters, which usually used, like /, ., &, ?, :, -, _, etc. In the second step, our classifier removes the more used words, such www, html, htm, http, etc. The obtained URL will be stemmed using the porter stemmer [13]. Finally, we apply the centroid-based model as explained above.

For a new document, contextual categorization is represented by a vector CC = $\{(c_1, \alpha_1), \ldots, (c_i, \alpha_i), \ldots, (c_k, \alpha_k)\}$ where, c_i is the ith category and α_i is similarity between the new document and the centroid of the category c_i. This similarity is calculated using cosine measure as presented above.

B. Logical Classifier

Logical structure is useful for genre identification because is reflect the order of author ideas in a given document [7].

The first step of this classifier is to extract the logical structure of the document, which can be internal, external or both. Internal structure is represented by title and section headers. This kind of structure can be extracted using the text contained in <title> and <Hn> tags. However, external structure is represented by title and hyperlinks and can be extracted using <title> and <A> tags. We notice that logical structure can be the combination of both internal and external structures.

In our approach we have used both internal and external structures. Much experimentation has been proposed to show the importance of each kind of structure.

Like contextual classifier, the text contained in <title>, <hn> and <a> is stemmed using porter stemmer [13]. Finally, we apply the centroid-based model as explained above.

For a new document, logical categorization is represented by a vector LC = $\{(c_1, \beta_1), \ldots, (c_i, \beta_i), \ldots, (c_k, \beta_k)\}$ where, c_i is the ith category and β_i is similarity between the new document and the centroid of the category c_i. This similarity is calculated using cosine measure as presented above.

C. Combination

In our approach we have two homogenous classifiers. To obtain an optimal categorization we should combine the results of contextual and logical classifiers. Many combination methods have been proposed in the literature [21]. They can be sequential if the classifiers outputs are obtained in sequential way, parallel, if the classifiers outputs are present in the same time or mixed. We can classify combination methods according to the type of classifier output, which can be a class, a rank or a measurement. In our case, the outputs of the two classifiers (contextual and logical classifiers) are a set of measurements. In this context many combination rules have been used, such as: maximum, minimum, product, median, linear rule and Bayesian rule. Using these rules, the final categorization FC is defined as follow:

FC = f(CC, LC) =$\{(c_1, f(\alpha_1, \beta_1)), \ldots, (c_i, f(\alpha_i, \beta_i)), \ldots, (c_k, f(\alpha_k, \beta_k))\}$, where k is the number of predefined categories and f is a combination rule such as maximum, minimum, product, etc.

III. Experimentation Setup

Both contextual and logical classifiers are experimented separately. The combination of these classifiers is compared against some classification techniques implemented in the rainbow software package [11].

In this section we first describe the datasets used to perform experimentations. In the second paragraph, we present the experimentations of contextual, logical and combined classifiers. Finally, we compare our approach against some classification techniques implemented in the rainbow package.

A. Datasets

To perform experimentations, we should use a datasets of HTML documents grouped by genres. To evaluate contextual classifier, each document should contain his URL address. According to these conditions, we can use only two datasets, which are KI-04 [12] and WebKB [5] collections. KI-04 corpus[1] is developed by Meyer zu Eissen and Stein and is composed of 1209 web pages distributed over 8 genres. WebKB corpus[2] contains 4518 web pages classified to one of six genres commonly found on computer science department websites (course, department, faculty, project, staff and student) and the category other that not considered in this research. For each dataset, we have removed all empty web pages. After this we have obtained the following datasets presented in Table I and Table II.

TABLE I
Composition of KI-04 Dataset

Category	# of samples
Article	127
Download	151
Link collection	205
Private portrayal	126
Non private portrayal	163
Discussion	127
Help	139
Shop	167
Total	**1205**

[1] http://www.itri.brighton.ac.uk/~Marina.Santini/

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

TABLE II
COMPOSITION OF WEBKB DATASET

Category	# of samples
Student	1541
Faculty	1063
Staff	126
Department	170
Project	474
Course	875
Total	**4249**

To experiment our approach we have proposed to use the accuracy measure for each category or genre. For all categories, we have chosen to use the micro average, because web pages in experimentation datasets are not equally distributed over categories.

B. Contextual Classifier

For the contextual classifier, we have obtained 0.73 as micro average accuracy for WebKB dataset and 0.78 for KI-04 dataset. These results are explained by made that web pages of the KI-04 collection have been downloaded form different sources, unlike WebKB collection, which have been developed from only computer science departments of four American universities (Cornell, Texas, Washington and Wisconsin).

C. Logical Classifier

The experimentation of the logical classifier is performed by different combinations of title, hn and anchor tags. Different results have been obtained, which are summarized in the following table. For both KI-04 and WebKB datasets, the best accuracy is obtained for the combination of title, Hn and Anchor tags.

TABLE III
ACCURACY OF LOGICAL CLASSIFIER

Tags	KI-04	WebKB
Title	0.79	0.84
Hn	0.78	0.82
Anchor	0.81	0.83
Tilte+Hn	0.77	0.86
Hn+Anchor	0.80	0.76
Title+Hn+Anchor	**0.86**	**0.88**

D. Combined Classifier

In this experimentation we measure the effect of each combination rule on the final accuracy of categorization. For the logical classifier we have used title, Hn and anchor tags because they provide the best accuracy as shown in the table above. The results are presented in the following table. These results show that the minimum rule provide the best result (0.88 for KI-04 corpus and 0.90 for WebKB corpus). This result is better than those obtained by Boese [4] and Meyer zu Eissen [12][17].

TABLE IV
ACCURACY OF COMBINED CLASSIFIER

Combination rule	KI-04	WebKB
Minimum	**0.88**	**0.90**
Maximum	0.87	0.84
Product	0.76	0.77
Median	0.84	0.83
Linear rule	0.86	0.77
Bayesian rule	0.80	0.78

E. Rainbow Results

To compare our approach to other classification methods, we have used the famous rainbow program[3]. As first step, we have extract for each document url, title, Hn and anchor content in separate files, which represents the input of rainbow program to generate models. We notice that rainbow provides a number of data preparation and classification options. In our experimentations we have used tfidf, naïve bayes (NB), knn, svm and tree nodes (tree) as classification method options. The results are presented in the following table (Table V for KI-04 dataset and Table VI for WebKB dataset). From these tables we notice that tfidf method provide best results for all features. But these results are less than obtained with our approach.

TABLE V
RAINBOW ACCURACY FOR KI-04 DATASET

	Tfidf	NB	Knn	SVM	Tree
URL	70.10	67.68	42.22	52.88	67.47
Title	81.19	75.66	62.78	56.93	74.58
Anchor	71.20	64.71	32.33	64.30	65.78
Hn	70.30	42.22	17.18	25.20	42.42
Title+anchor	75.61	65.65	31.30	54.52	61.74
Title+Hn	80.82	78.16	45.92	49.75	76.33
Hn+anchor	74.70	62.96	27.73	43.61	63.97
Title+hn+anchor	84.40	80.20	50.41	55.51	70.14

TABLE VI
RAINBOW ACCURACY FOR WEBKB DATASET

	Tfidf	NB	Knn	SVM	Tree
URL	73.40	69.63	24.55	39.64	70.15
Title	70.23	61.25	45.25	37.29	68.54
Anchor	73.58	71.12	40.48	40.21	68.47
Hn	76.19	59.86	40.47	40.25	47.58
Title+anchor	72.15	70.78	25.46	40.13	70.17
Title+Hn	77.13	80.13	55.23	42.16	79.28
Hn+anchor	77.12	74.78	19.15	41.89	75.46
Title+hn+anchor	84.29	77.74	37.29	43.17	80.19

IV. CONCLUSION

In this paper we have proposed a new approach for flexible document genre categorization. The originality of our approach is the combination of two different classifiers and the use of logical structure of web page. The proposed approach is flexible because it assigns a web page to all categories. Each category is associated with a weight representing the similarity between the document and the

[3] http://www.cs.cmu.edu/~mccallum/bow/rainbow/

given category. Our approach is based on the category centroid generated from the training set.

The Experimentations have demonstrated that our approach provides results as good as those obtained by rainbow classifiers. In the future we hope to integrate our approach in a web search engine.

References

[1] Biber, D. Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, 62(2), 1986, 384-413.

[2] Biber, D. The multidimensional approach to linguistic analyses of genre variation: an overview of methodology and finding. *Computers in humanities*, 26(5-6), 1992, 331-347.

[3] Biber, D. Dimensions of register variation: a cross-linguistic comparison. Cambridge, England: *Cambridge University Press*, 1995.

[4] Boese, E. S and Howe, A. E. Effects of web document evolution on genre classification. *In proceeding of 5th conference information and knowledge management*, Berlin, Germany, 2005.

[5] Craven, M., DiPasque, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K. and Slattery, S. Learning to extract symbolic knowledge from the word wide web. *In proceeding of the 15th national/10th conference on artificial intelligence/innovative applications of artificial intelligence*. Madison, W, 1998.

[6] Dewdney, N., Vaness-Dikema, C. and Macmillan, R. 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] Jebari, C., Ounalli, H. The Usefulness of Logical Structure in Flexible Document Categorization. *In Proceeding of the International Conference on Computational Intelligence*, Istanbul, Turkey. *International Journal of Information Technology*. 2004.

[8] Karlgren, J. and Cutting, D. Recognizing Text Genre with Simple Metrics Using Discriminant Analysis. *In Proceedings of the 15th International Conference on Computational Linguistics* (COLING 1994). Kyoto (Japan), 1994.

[9] Kessler, B., Numberg, G. and Shutze, H. Automatic Detection of Text Genre. *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, Spain, 1997.

[10] Lee, Y. and Myaeng, S. Text Genre Classification with Genre-Revealing and Subject-Revealing Features. *In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (SIGIR 2002). 2002, Tampere, Finland, 2002.

[11] McCallum, A. *Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering*, 1996 (http://www.cs.cmu.edu/~mccallum/bow).

[12] Meyer zu Eissen, S. and Stein, B. Genre Classification of Web Pages: User Study and Feasibility Analysis. In Biundo S., Fruhwirth T. and Palm G. (eds.). KI2004: Advances in Artificial Intelligence. *Springer. Berlin-Heidelberg-New York*, 2004, 256-269.

[13] Porter, M. An algorithm for suffix stripping. *Program*, 14(3), 1980.

[14] Rauber, A. and Muller-Kogler, A. Integrating Automatic Genre Analysis into Digital Libraries. *In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (JCDL 2001), 2001, Roanoke, Virginia (USA).

[15] Roussinov, D., Crowston, K., Nilan, M., Kwasnik, B., Cai, J., Liu, X. Genre based navigation on the web. *In proceedings of the 34th Hawaiin International Conference on System Sciences*, Hawaii, 2001. *IEEE Computer Press*.

[16] Salton, G. Automatic Text Processing: The transformation, analysis and retrieval of information by computer. 1989, *Addison-Wesley*.

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

[18] Sebastiani, F. Machine learning in automated text categorization, *ACM Computing Surveys*, 34(1), 2002, 1-47.

[19] Stamatatos, E., Fokatakis, N. and Kokkinakis, G. Text Genre Detection Using Common Word Frequencies. *In Proceedings of the 18th International Conference on Computational Linguistics* (COLING 2000). 2000. Saarbrücken (Germany).

[20] Wang, Y., and Kitsuregawa, M. Evaluating contents-link coupled web page clustering for web search results. *In proceeding of 11th international conference on information and knowledge management*, 2002, 499-506.

[21] Zouari H., Heutte L., Lecourtier L. and Alimi A. Un panorama des méthodes de combinaison de classifieurs en reconnaissance de formes. *In 13ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et d'Intelligence Artificielle RFIA'02*, Angers, France, vol. 2, 2002, 499-508.