World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:12, 2007

The Usefulness of Logical Structure in Flexible Document Categorization

Jebari Chaker and Ounalli Habib

Abstract**—** This paper presents a new approach for automatic document categorization. Exploiting the logical structure of the document, our approach assigns a HTML document to one or more categories (thesis, paper, call for papers, email, …). Using a set of training documents, our approach generates a set of rules used to categorize new documents. The approach flexibility is carried out with rule weight association representing your importance in the discrimination between possible categories. This weight is dynamically modified at each new document categorization. The experimentation of the proposed approach provides satisfactory results.

Keywords**—** categorization rule, document categorization, lexible categorization, logical structure.

I. INTRODUCTION

In front of the incredible growth of the Internet, we notice that document categorization is very important in many applications, in particular the information retrieval. Indeed, document categorization can be used in two information retrieval steps: 1) The organization of document collection by category with the intention of improving the efficiency and the effectiveness of the information retrieval process or 2) The organization of the provided documents by category with the intention of accelerating the selection of relevant documents and improving the visualization quality.

In this paper, we propose a new flexible approach for document categorization based on document logical structure. This approach assigns a HTML document to one or more predefined categories (thesis, paper, call for papers, email, …) using the document logical structure.

Our proposed approach can be useful for many other applications:

- Exploiting only terms contained in thematic units[1], extracted using document category, can improve thematic classification accuracy [1].
- Assigning document to one or more categories can facilitate the assimilation and dissemination of great information loads by guiding user search in function of their needs and profiles [2].
- Since, the different automatic document summarization methods depend on document category (thesis, paper, call for papers, email, …). Our approach allows the application of suitable summarization method.

[1] A thematic unit is a logical unit that touchy to announce the theme of the entire document.

This paper is organized as follows. The next section presents some related works in document categorization. The principle of the proposed will be presented in the third section. In the fourth, fifth and sixth sections of this paper we explained the fundamental steps of our approach: generation of categorization rules, categorization of new documents and modification of categorization rules. The experimentation of our approach is also presented in the seventh section. In the conclusion we propose some possible future works.

II. DOCUMENT CATEGORIZATION: RELATED WORKS

The automated document categorization dating back to 60 years, with Maron works [3]. Since then, several authors have proposed different categorization concept definitions. According to Sebastiani [4], the categorization of documents set D consists in assigning each document d belonging to D a category c belonging to a set of predefined categories C.

Automatic document categorization has been used in a number of different applications: automatic indexing for Boolean information retrieval systems, document organization, word sense disambiguation, yahoo-style search categorization [4].

We can distinguish between two kinds of document categorization: thematic and contextual. The thematic categorization aims to identify the document theme using the document content. On the other hand, contextual categorization aims to identify document theme using contextual information like metadata (type, authors, …) [2].

Automatic document categorization can also be used to identify the document type (web page, email, paper, call for papers, …). But in the literature we have a few works that have been devoted to this kind of categorization [5][6][7][8][9][10]. These methods differ in number and kinds of predefined categories that make difficult the comparison between these methods. For example Kevin propose 7 categories for web documents (reportage, editorial, research articles, Reviews, home page, Q&A, spec) [9], Marzin propose 4 categories for web pages (links pages, home pages, web navigators and sales pages) [10].

Several automatic document categorization methods have been proposed in the literature and have been devoted to thematic categorization. These methods can be divided in two approaches: knowledge engineering and machine learning methods. Maron has proposed knowledge engineering approach in 1961 [3]. It based on categorization rules of type IF Condition THEN Category [11][12].

This approach has been abandoned because it needs a manual effort to build and manages the set of categorization rules. To solve this problem the categorization community have been propose in 1980 to use some machine learning techniques [13]. The principle of this last approach consists in automatically generating a categorization function using a set of training documents. This function is used to categorize new documents. Among machine learning algorithms we mention: Rocchio's algorithm [14], K-Nearest Neighbor [15], Decision trees [16], Support Vector Machines [17], Voted classification [18] [19].

III. PRINCIPLE OF THE PROPOSED APPROACH

Our proposed approach assigns a French HTML document to one or more predefined categories (dictionary, patent, book, thesis, memory, report, paper, FAQ, call for papers, web pages, news, email) using the document logical structure.

Our approach is situated in junction of the knowledge engineering and machine learning approaches. Using a set of training documents, our approach allows to automatically generating a categorization function. This function is represented in the form of a set of categorization rules. Contrary to other methods such as decision trees [20][16][21], galois lattice [22] or induction graphs [23][24], where graph transformation in rules is necessary.

In our approach, each rule is in the form IF Condition THEN Conclusion, where Conclusion represents the appartenance degrees to all predefined categories. The categorization flexibility is carried out with rule weight association representing your importance in the discrimination between possible categories. This weight is dynamically modified at each new document categorization.

The principle approach is presented in figure I below.

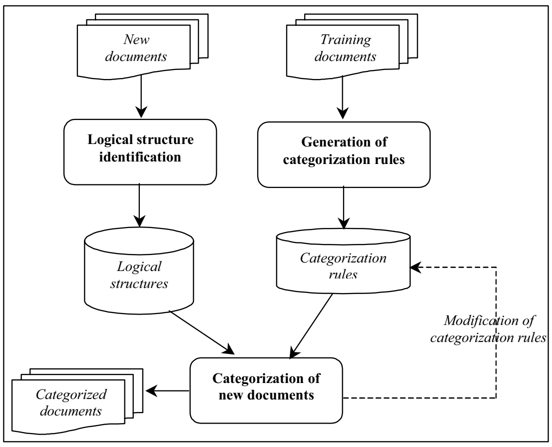

FIGURE I
PRINCIPLE OF PROPOSED APPROACH

A. Training collection

To generate categorization rules, we have collect from web a training set **A** of 1230 HTML documents. Each training document $\mathbf{d_j}$ is represented by: the identification $\mathbf{did_j}$, the category $\mathbf{C_j}$, and the logical structure $\mathbf{sl_j}$. The distribution of the training set **A** on the 12 possible categories is presented in the table I below.

B. Logical structure

A logical structure is represented by a series of logical units ordered one after the other to appear an idea. For each logical unit we have associated a weight between 0 and 1 representing your importance in the logical structure construction. This weight is calculated using the training documents. We have identified 9 possible logical structures (See table II).

TABLE I
NUMBER OF TRAINING DOCUMENTS BY CATEGORY

Notation	Category	# Of training documents by category
C_1	Dictionary	30
C_2	Book	40
C_3	Patent	40
C_4	Thesis	100
C_5	Memory	100
C_6	Report	100
C_7	Paper	120
C_8	FAQ	100
C_9	Call for papers	100
C_{10}	News	160
C_{11}	Web page	180
C_{12}	Email	160

C. Categorization rules

Using the logical structure values, we have identified 9 possible recognition rules. Each rule is of type:

TABLE II
PREDEFINED LOGICAL STRUCTURE

Notation	LOGICAL STRUCTURE	Categorie(s)
SL_1	Titre (1), date et lieu (1), introduction (1), thèmes abordés (1), soumission (1), comité scientifique (1), comité d'organisation (1), dates importantes (1), informations (0.8).	Call for papers
SL_2	Titre (1), auteur(s) (1), affiliation(s) (1), email(s) (1), résumé (1), mots clés (1), introduction (1), texte (1), conclusion (1), remerciements (0.2), références (1).	Paper
SL_3	Titre (1), résumé (1), mots clés (1), abstract (0.5), key words (0.5), dédicaces (0.3), remerciements (0.8), table des matières (1), table des illustrations (0.2), introduction (1), texte (1), conclusion (1), bibliographie (1), annexes (0.4), glossaire (0.2), index (0.2).	Thesis & memory & report

$$\text{IF } SIM(sl_j, sl_i) \geq S_0 \text{ THEN } \{(C_1, \alpha_1), \ldots, (C_{12}, \alpha_{12})\}$$

Where:

- α_i is the appartenance degree to the category C_i. This degree is the proportion of training documents, which belongs to the category C_i.
- S_0 it's the similarity threshold, under this value the categorization the categorization rule cannot be applied. In our case, we have chosen a threshold value as 0.5.
- $SIM(sl_j, sl_i)$ is the similarity between document logical structure sl_j and the predefined logical structure sl_i. This similarity is calculated using this formula:

$$SIM(sl_j, sl_i) = \frac{\sum_{ul_i \in sl_j \cap sl_i} p_i}{\sum_{ul_i \in sl_i} p_i}$$

Where:

- ul_i : A logical unit belonging to the predefined logical structure sl_i.
- p_i : The weight assigned to the logical unit ul_i.

Example:

IF $SIM(sl_j, sl_3) \geq 0.5$ THEN {(dictionary, 0.00), (**book**, 0.70), (**patent**, 0.60), (**thesis**, 1.00), (**memory**, 1.00), (**report**, 1.00), (**paper**, 0.15), (FAQ, 0.00), (call for papers, 0.00), (news, 0.00), (web page, 0.00), (e-mail, 0.00)}

V. CATEGORIZATION OF NEW DOCUMENTS

At each new document d_j, the categorization process identifies the document logical structure sl_j using <Hn>. After this preprocessing, the categorization process allows the selection of the adequate categorization rule by comparing document logical rule with predefined logical structures. The application of the suitable rule provide the following set of possible categories representing the rule conclusion:

$$Categorization=\{(C_1, \alpha_1), \ldots, (C_{12}, \alpha_{12})\}$$

In general, we choose the category having the highest appartenance degree.

Example:

If Categorization={(dictionary, 0.00), (**book**, 1.00), (**patent**, 0.20), (**thesis**, 0.30), (**memory**, 0.10), (**report**, 0.50), (**paper**, 0.15), (FAQ, 0.00), (call for papers, 0.00), (news, 0.00), (web page, 0.00), (e-mail, 0.00)}.

We choose the category "book" because he has the maximum weight.

VI. MODIFICATION OF CATEGORIZATION RULES

After each new categorization, we should update the set of rules. This modification is summarized in two fundamental points, which are:

- Remove rules, which their conclusions are equal to 0. In other words, the rules whose all their appartenance degrees to all possible categories are equal to 0.
- Since, the proportion of training documents verifying Conclusion rules will be modified. We should recalculate the appartenance degrees for all rules.

VII. EXPERIMENTATION

To experiment any categorization method you have two possible techniques: comparing the obtained categorization with another categorizations given by another categorization methods or comparing the obtained categorization with manual or *reference* categorization.

In our case, the comparison with other approaches is impossible because all the proposed approaches don't use the same number and kinds of predefined categories (see section 2). So we have chosen the second technique.

Our proposed approach has been implemented in the CFD system. To experiment this system, we have used a corpus of 615 HTML documents belonging to the possible categories (see table III).

TABLE III
DISTRIBUTION OF TESTING DOCUMENTS BY CATEGORY

Notation	Category	# Of training documents by category
C_1	Dictionary	10
C_2	Book	10
C_3	Patent	10
C_4	Thesis	30
C_5	Memory	35
C_6	Report	50
C_7	Paper	70
C_8	FAQ	70
C_9	Call for papers	60
C_{10}	News	100
C_{11}	Web page	90
C_{12}	Email	80

For each testing document $\mathbf{d_j}$. We have identified the logical structure sl_j. Exploiting this logical structure, we have obtained the following results presented in table IV.

TABLE IV
RECALL, PRECISION, ACCURACY AND ERROR BY CATEGORY

Category	Recall	Precision	Accuracy	Error
Dictionary	0.66	0.67	0.65	0.35
Book	0.79	0.80	0.77	0.23
Patent	0.76	0.78	0.75	0.25
Thesis	0.85	0.88	0.85	0.15
Memory	0.87	0.90	0.88	0.12
Report	0.84	0.85	0.82	0.18
Paper	0.91	0.93	0.90	0.10
FAQ	0.71	0.72	0.70	0.30
Call for papers	0.80	0.82	0.80	0.20
News	0.62	0.65	0.60	0.40
Web page	0.58	0.60	0.55	0.45
Email	0.76	0.77	0.75	0.25

From the table 4 we notice that recall, precision, accuracy and error values are acceptable for all categories. These remarks confirm that logical structure is very important for document categorization. In particular for strongly structured documents (documents who's logical structure is explicit and very easy to extract). For example: academic documents (thesis, memory, report, paper), call for papers, email. We have obtained a recall average value of 0.87, a precision average value of 0.94, an accuracy average value of 0.84 and an error average value of 0.16. These results are satisfactory.

VIII. CONCLUSION AND FUTURE WORKS

In this paper, we have proposed a new approach for document categorization. This approach exploits document logical structure. Using a set of training documents, our approach allows the generation of a set of categorization rules. Each rule is of the type IF Condition THEN Conclusion. The Conclusion of each rule represents the appartenance degrees to possible categories.

The experimentation provides satisfactory results especially for strongly structured documents.

In this research, we have used only HTML documents. In the future works, we propose:

- The integration of new electronic formats (SGML, XML, …) to exploit the meta data provided by the Dublin Core[2] norm.
- The integration of this approach in the process of information retrieval to improve their performance.

ACKNOWLEDGMENT

Thanks to my thesis supervisor Dr. Ounalli Habib and all me colleagues in King Saud University in Saudi Arabia. My thanks also to all my family.

REFERENCES

[1] C. Jebari & al., Catégorisation d'un document électronique en vue d'une meilleure classification thématique, *GEI'2002*, Hammamet, Tunisie, 2002.
[2] V. Chanana & al., A new context-based information retrieval system, *Accepted in 3rd WSEAS Int. Conf. On Artificial Intelligence, Knowledge Engineering, Data Bases (AIKED 2004)*, Salzburg, Austria, February 13-15, 2004.
[3] M. Maron, Automatic Indexing: An Experimental Inquiry, *Journal of the Association for Computing Machinery*, 1961, 8(3): pp. 404 – 417.
[4] F. Sebastiani, Machine Learning in Automated Text Categorization, *ACM Computing Surveys*, Pisa, Italy, 2002.
[5] J. Karlgren and D. Cutting, Recognizing Text Genres with Simple Metrics Using Discriminant Analysis, *Proc. Of COLING1994*, Kyoto, 1994.
[6] L. Yong-Bae and Sung Hyon, Automatic Identification of Text Genres and Their Roles in Subject-Based Categorization, *In Proceedings of the 37th Hawaii International Conference on System Sciences*, 2004.
[7] B. Kessler & al., Automatic Detection of Text Genre, *ACL'97, pages 32 – 38*, July 1997.
[8] E. Stamatatos, Text Genre Detection Using Common Word Frequencies, *Proc. Of the 18th International Conference on COLING2000*, 2000.
[9] C. Kevin and W. Marie, Reproduced and emergent genres of communication on the world-wide web, In *Proceedings of the 30th Hawaii International Conference on System Sciences (HICSS-30)*, Institute of Electrical and Electronics Engineers, 1997.
[10] A. Marzin & al., Classification de pages web en genre, *Journée d'études ATALA'2004*, Grenoble, France, janvier 2004.
[11] C. Apte & al., Automated learning of decision rules for text categorization, *ACM Transactions on Information Systems*, 1994, 12(3): pp. 233 – 251.
[12] P.J. Hayes, CONSTRUE/TIS: a system for content-based indexing of a database of news stories, *In Proceedings of IAAI-90, 2nd Conference on Innovative Applications of Artificial Intelligence*, 1990, pp. 1 – 5.
[13] T. Mitchell, *Machine Learning*, McGraw Hill International editions, Computer Science series, ISBN 0-07-042807-7, 1997.
[14] J. J. Rocchio, Relevance Feedback in Information Retrieval, *In the SMART retrieval system*, G. Salton, pp. 313 – 323, Prentice Hall, Inc., 1971.
[15] R.O. Duda & al., Pattern Classification and Scene Analysis, *John Wiley & Sons*, 1973.
[16] L. Breiman and al., *Classification and Regression Trees*, Belmont, CA: Wadsworth, 1984.
[17] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer – Verlag, 1995.
[18] L. Breiman, Bagging predictors, *Machine Learning*. Vol. 24, 1996, pp. 123 – 140.
[19] Y. Freund and Shapire, Experiments with a new boosting algorithm, *In Proceeding of 13th international conference on Machine Learning*, 1996, pp. 148 – 156.
[20] J.R. Quinlan, C4.5: Programming for machine Learning, *Morgan Kaufman*, 1993.
[21] J.R. Quinlan, *Learning efficient classification procedures and their application to chess and games,* In R. S. Michalski, J. G. Carbonell and T. M. Mitchell editors, Machine Learning: An Artificial Intelligence Approach. Vol. 1, pp. 463 – 482, 1983.
[22] E. Mephu Nguifo, Treillis de Galois et Classification Supervisée, Séminaire LIMOS, Clermont – Ferrand, 7 mars 2002.
[23] R. Rakotomalala, Graphes d'Induction, Thèse de doctorat de l'université Claude Bernard – Lyon I, décembre 1997.
[24] D.A. Zighed et al., *SIPINA : Méthode et logiciel*, Editions Alexandre Lacassagne, Mathématiques appliquées n°2, 1992.

Jebari Chaker, King Saud University, College of Computer and Information Sciences, Computer Sciences Department, PO. BOX 51178 Riyadh 11543, Kingdom of Saudi Arabia (email: jebarichaker@yahoo.fr). Mr. J. Chaker is a computer sciences lecturer in King Saud University, KSA. His research is about document categorization (document type recognition) and information retrieval.

Ounalli Habib, Université de Tunis El'Manar, Faculté des Sciences de Tunis, Département d'Informatique, Campus Universitaire 2092 Tunis, Tunisie (email: habib.ounelli@fst.rnu.tn). Dr. O. Habib is a permanent member of ERPAH (Equipe de Recherche en Programmation Algorithmique et Heuristique). He is a doctor in Faculty of Sciences, Tunisia.

[2] See http://dublincore.org.