A Categorization Method for French Legal Documents on the Web

Guiraude Lame
Centre de Recherche en Infomatique
Ecole nationale des mines de Paris
35, rue Saint-Honoré
F-77305 Fontainebleau
Tel +33164694850 Fax +33164694847
lame@cri.ensmp.fr

1. INTRODUCTION

This paper briefly describes an on-going work in categorizing French legal documents. We used documents from the French official publication *Journal Officiel de la République française, édition Lois et Décrets* (J.O.) which gathers laws, decrees, decisions from various administrations. These documents are published on an internet site http://droit.org which intends to be a web portal for French law. The principle aim of this text categorization system is to determine which subfields of law a given legal document is dealing with. As a result, a thematic access to subfields of law will be provided and text retrieval effectiveness of legal documents improved, as reported by [3].

As our documents deal with the same subject, namely the field of law, categories fitting these documents are very closely related. Vocabularies of the different subfields of law overlap. Consequently, we cannot use unsupervised clustering methods to classify our documents, since these methods are unable to create an adequate fine-grained subject isolation [1]. Our method is therefore a supervised categorization method in which we predefine categories of the documents of the J.O. These categories are designed on the basis of a corpus gathering compendiums (called Codes). These Codes have been officialy created in France to gather all the rules related to a specific subfield of law. Thus each Code describes a particular subfield of law : penal law for Penal Code, trade law for Trade Code…

The first step of our text categorization method is to elaborate an indexing language (Section 2) on the basis of which we obtain representations for both documents and categories. The second step (Section 3) uses similarity measures computed between these documents' and categories' representations. We compared the results given by three different coefficients to select the best one. We present in Section 4 our first experimental results.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Conference '00, Month 1-2, 2000, City, State.
Copyright 2000 ACM 1-58113-000-0/00/0000…$5.00.

2. AN INDEXING LANGUAGE

Our purpose here is to perform a feature selection to obtain appropriate descriptions for our documents and categories. Relevant index terms are selected as content descriptors for documents. We decided to use both multiword and single word indexing since the first approach provides precision and the second effectiveness.

To extract words and multiwords phrases, we are using a tool called Lexter [2]. This tool performs a natural language analysis (namely syntactic analysis) to identify nouns, verbs, adjectives, adverbs, noun and verbal phrases in documents. On these outputs of Lexter, we perform a filtering process that rejects different types of words or phrases. Adverbs, adjectives and prepositions identified and labelled by Lexter are ignored as are phrases and words containing a number. A specific stopword list of 81 elements has also been manually elaborated to filter specific terms such as months, "decree", "minister"… After this filtering process, each term is weighted with its frequency in the considered document. We have chosen in a first experiment to use absolute frequencies of terms.

We apply this process to both documents from the J.O. and Codes. As a result, we obtain for each of these a vector associating index terms and their frequencies. In this way, each document can be represented by an *abstract representation* denoted by a set of pairs called features.

$$D_i = \{(d_{i1}, w_{i1}), (d_{i2}, w_{i2}), \ldots, (d_{in}, w_{in})\}$$

D_i : abstract representation for document i

d_{ij} : index term j in document i

w_{ij} : weight attributed to index term j in document i

Similarly, a category is represented as follows :

$$K_l = \{(k_{l1}, w_{l1}), (k_{l2}, w_{l2}), \ldots, (k_{lm}, w_{lm})\}$$

3. CATEGORIZING ALGORITHM

As seen in Section 1, we predefine categories on the basis of an existing taxonomy of the field of law embodied in the Codes (60 of which are currently avalaible on line). Considering each Code as a category is appropriate but insufficient since this would distinguish too coarsely our documents among different laws : Trade Law, Penal Law… Since each Code is divided in subsections such as Titles, Sections, Chapters, we decided to consider each Title as a category ; these parts of Codes more precisely treat subfields of law such as divorces in Civil Law, traders liability in Trade Law. Each Title of a Code then defines a category for our documents. Each abstract representation of these categories defines a particular weighted vocabulary of subfields of law. The goal is to determine which categories best fit a given document. The similarity score of each category to a given document is used to rank the categories ; the 10 hightest weighted categories are attributed to the document.

To perform this discriminant analysis, features of categories and documents abstract representations are compared, using different similarity measures. In our study, three classical similarity measures, namely Jaccard coefficient, cosine coefficient and Dice coefficient (see [4]), have been experimentaly compared to determine which one gives better results with our sets of documents and categories. The Dice coefficient has been selected (see Section 4).

Dice coefficient :

$$S_{D,K} = \frac{2 \sum_{p \in terms(D_i) \cap terms(K_j)} weight_{D_i}(p) \times weight_{K_j}(p)}{\sum_{p \in terms(D_i) \cap terms(K_j)} weight_{D_i}(p)^2 + \sum_{p \in terms(K_j) \cap terms(K_j)} weight_{K_j}(p)^2}$$

$$terms(\{..(d_i, w_i)..\}) = \{..d_i...\}$$
$$weight_{\{...(d_i, w_i)...\}}(d_i) = w_i$$

4. EXPERIMENTAL RESULTS

To assess the accuracy of our technique, we selected 26 Codes corresponding to 1339 subfields of law to categorize 101 documents from the J.O. corresponding to all French laws for 2000 and until March 2001. Early experimental results suggest that Jaccard indices often yield irrelevant results, attributing most of the time the same subset of categories to documents. The cosine and Dice indices give better results and an expert of the legal domain confirmed that Dice performs accurate categorizations of our documents.

Evaluation is, in our work, difficult since we do not have any benchmark to compare our work with others. We thus asked a legal domain expert to validate our results. Since our first experiments rely on 26 Codes (compared with the total of 60), it was hard for our expert to determine which of the 10 categories were relevant in our results for each law. We thus asked her which categories were correctly identified by our system. For the total of 101 laws above described, a cursory analysis of our experiment underlines that our system is unable to correctly determine subfields for laws ratifying international treaties. It is mainly because these laws (50 in our 101 set) are not thematically related with French law. We then decided to skip these laws from our evaluation set. For the 51 remaining laws, these dealing with national law, three subsets can be identified. For the Subset 1 (see Figure 1 below), our system failed since less than five subfields were considered relevant by our expert. She noticed that the laws from Subset 1 either contain few words (one or two sentences) or cannot be categorized in any subfields of French law since they are not related with law. This is the case, for instance, of the 2000-44 law instituting a national day in memory of the French State racial and antisemite crimes victims. Subset 2 mainly gathers laws semantically related with subfields defined in Codes we still have not taken into account. Subset 3 gathers laws (16 for a total of 51) well categorized since more than 70% of the identified categories are relevant. Nevertheless, some problems still remain, related to language ambiguities.

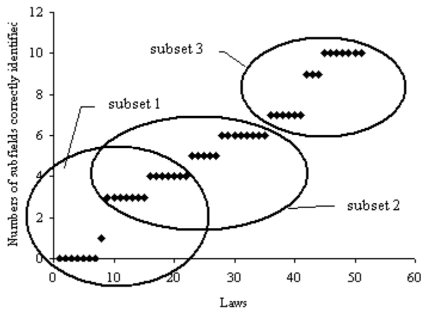

Figure 1 Evaluation for national laws

5. FUTURE WORK

The first limit of our system is that it relies on documents themselves : some cannot be related to subfields of French law and don't have to. The second limit stems from language ambiguities unavoidibly occurring in documents. To solve these problems, a terminological analysis in which each category's abstract representation defines the specific vocabulary for the concerned subfield of law has to be performed. Adjustments could thus be made on the weights attributed to each term of the categories. We intend to deal with these issues next, also as include the Codes that haven't been analyzed yet.

6. REFERENCES

[1] Aggarwal, C., Gates, S., Yu, P. On the merits of building categorization systems by supervised clustering. In Proceedings of SIKDD'99 (San Diego, CA, 1999), ACM Press, p. 352

[2] Bourigault, D. Lexter, un Logiciel d'Extraction de TERminologie. Application à l'extraction des connaissances à partir de textes. PhD Thesis, Mathématiques, informatique appliquées aux sciences de l'homme. EHESS (Ecole des Hautes Etudes en Sciences Sociales) , Paris, 1994

[3] Lam, W., Ruiz, M., Srinivasan, P. Automatic text categorization and its application to text retrieval. In IEEE Transactions on Knowledge and Data engineering, 11(6):865-879

[4] Salton, G. Automatic Text Processing : The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley, 1989