Multi-catégorisation de textes juridiques et retour de pertinence

Vincent Pisetta, Hakim Hacid et Djamel A. Zighed

article paru dans
First IEEE International Workshop on Mining Complex Data (IEEE MCD'05), Houston, Texas, USA, 2005. pp. 81-84.

Automatic juridical texts classification and relevance feedback

Vincent Pisetta, Hakim Hacid and Djamel.A Zighed
University of Lyon 2
ERIC Laboratory – 5, av. Pierre Mendès-France- 69767 Bron- France
vpisetta@etu.univ-lyon2.fr ; hhacid@eric-univ.lyon2.fr ; zighed@univ-lyon2.fr

***Abstract*— Text retrieval is an important field in automatic information and natural language processing. A large variety of conferences, like TREC, are dedicated to it. The goal of text retrieval systems is to process textual data in order to make easy their retrieval. In this paper, we are interested in text retrieval and especially the automatic juridical texts classification as a way to facilitate their retrieval. We use linguistic tools (terminology extraction) in order to determine concepts presents in the corpus. Those concepts aim to create a low dimensionality space which enabling us to use automatic learning algorithms based on similarity measures as proximity graphs. We compare results extracted from the graph with a more classical method: C4.5. The originality of this work, in addition to the use of real data, is the use of two methods for classifying our data. Promising results are obtained using our methodology and a comparison between proximity graphs and decision trees is shown.**

***Index Terms*— text retrieval, automatic text classification, decision trees, proximity graph.**

I. INTRODUCTION

The general framework of machine learning starts from a training file containing n rows and p columns. The rows represent the items and the columns the features, quantitative or qualitative, observed for each item. In this context, one also supposes that the training sample is relatively high compared to the number of features. Generally speaking, the sample size is about 10 times the number of features to hope to obtain a stability, i.e. a generalization error not much more higher than training error. Moreover, the feature to be predicted is generally supposed with a unique value. It is a feature with actual values in the case of the regression and with discrete methods, called classes, in the case of the classification. In this paper we describe a situation of training which deviates significantly from the traditional framework described above. Indeed, the experimental context does not enable us to immediately have a consequent training sample, each item can belong to several classes simultaneously, and each item, instead of being described by an attribute-values unit, is described by a text in natural English language.

Before describing the approach that we recommend to learn in this context, we, first of all, will point out the problems of the application concerned (section 2). In section 3, we describe the adopted methodological approach. In section four, we describe the stages implemented to format the data and, in particular, the linguistic analysis strategy implemented to extract the main concepts which will have the role of features. We describe then, in section 5, the topological models containing proximity graphs which enable us to manage the multiclass problem. In a comparative aim, we also implement a method based on decision trees which is useful to better identify the discriminant concepts. In section 6, we introduce the results obtained from the linguistic analysis and the machine learning. We detail the obtained performances. In section 7, we conclude and we detail the future issues of this work, in particular the use of association rules.

II. EXPERIMENTS FRAMEWORK

A. *Problematics and corpus description*

This work falls under a project in collaboration with the International Labour Organization (ILO). Several countries signed conventions with the ILO which binds them to the international labour law. More concretely, the agreement relates to two conventions drawn up by the ILO, Convention n°87 and Convention n°98[1]. The conventions contain several laws articles which the signatories commit themself respecting. The countries are subjected, once per year, to an inspection having for goal to check the respect of the conventions' rules. At the end of each inspection, the experts of the ILO deliver a report to the concerned country. The report shows the observed violations and underlines the efforts to be set up in each country, in order to be in adequacy with the conventions.

Datasets (conventions) are expressed in free text without a specific coding of the violations and written in English language by lawyers elected by the ILO.

The goal of our work is to define and to set up data mining methods and tools making it possible to analyse more effectively and more quickly these corpora, which manually, as said before become not suitables. For these reasons, the experts of the ILO wish to have tools allowing the automatic location of the texts containing the violation of one or more

[1] The contents of these conventions and the list of the signatories are accessibles on (http://www.ilo.org/ilolex/cgi-lex/convde.pl?C087 et http://www.ilo.org/ilolex/cgi-lex/convde.pl?C098)

rules (articles) by the concerned countries. After the classification, the experts will be able to synthesize more quickly the difficulties that meet the various countries in the application of these conventions.

In order to achieve this, the ILO provides a database containing 1315 texts, 834 texts relative to the Convention N°87 containing 17 violations (rules) and 481 texts relative to the Convention N°98 describing 10 violations. Each text, corresponds to an annual report addressed by the experts of the ILO about specific country. Each text describes the violataed rules related to each convention and the methods of this violation. The texts are sorted by convention (n°87 and n°98), date and country. Let us note, that a text can report the violation of several rules.

III. Methodology

A. General principle

In order to set up methods able to identify the violations in a corpus, we use machine learning techniques. Our choice is motivated by the state of the data. Indeed, we are in a supervised learning case in which the predictive features represent the coordinates of each text (extracted by text mining techniques) and the classes to be predicted are represented by the violations assigned to each text. Thus, we can build a precition model, which, once evaluated and considered to be acceptable by the user, can be used as an automatic categorization tool for the new texts.

In order to build a prediction model in a supervised learning context, the principle consists in providing to the training algorithm a dataset containing pre-classified texts called training set. In our case, because each text can violate one or more rules, we are in a multiclass learning problem. More formally, let ω be a text of the total corpus Ω and $c_i \; i = 1, \ldots, k$ the conventions rule sensitive to be violated. Then, $C(\omega) = \{c_1, c_2, ..., c_k\}$ expresses that the text contains violations relating to the rules $\{c_1, c_2, ..., c_k\}$ of convention v_a. Let us note in this context that it is difficult and extremely expensive to manually carry out this categorization in only one draft. We suggested to the experts to annotate about seventy texts by convention (exactly 71 for convention n°87 and 65 for convention n°98) which will be used as an initial learning datasets. Let us call this first corpus Ω_1.

Let Ψ be a training algorithm. The result of the training is a model, noted M, and a generalization error ε estimated on a sample test or by cross-validation.

$$\Psi(\Omega, C) = (M, \varepsilon)$$

The application of the model M on a new unknown dataset Ω' with relatively modest size allows to predict for each unknown text the rules which would be violated, for example $\mathrm{M}(\omega') = \{c_1, ..., c_k\}$. The relevance feedback allows the user to evaluate each labelled text belonging to the unknown dataset. The user U can then validate completely or partially the labelling suggested by the model. If, for a text ω', the prediction is considered to be erroneous by the user U then, the concerned text is inserted in the training sample $\Omega_{i+1} = \Omega_i \cup \omega'$ for another iteration. This operation is renewed each time that an item is considered badly labelled. We reiterate the training process with the new sample thus created. We obtain a new model M' which one hopes for a lower error rate $(\varepsilon' < \varepsilon)$. A new unknown sample of modest size to facilitate a manual checking is made up. The reiteration of this process should lead to an iterative improvement of the model. The question which arises consequently is the choice of an algorithm able to manage the multi-labelling. The proximity graphs[17], which belong to the instance based learning methods, allow that. In order to use these methods, the texts have to be transformed into a vector representation. Then, each text could then be considered as a point in a multidimensional space $\mathbb{R}^p$.

The coordinates of a text in this space will be $\mathrm{X}(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_p(\omega))$. What do these variables represent, how are they extracted ?

IV. REPRESENTATION SPACE

A. Terminology extraction

We choose to create our representation space by concepts extraction. One of the advantages of this technique compared to other methods as the N-grams, or the co-occurences matrices of words, is the significant reduction of the dimensionnality, allowing the use of classifiers based on similarity measures. Various applications based on this principle gave interesting results[9]. Two different methods exist for the concepts construction: by training or by extraction.

The first one is a statistical method and aims to search the most discriminant words according to a class to be predicted. The words are then gathered into concepts on the basis of their co-occurences or thanks to association rules[9]. The second method is a linguistic method and consists in extracting the terminology from the corpus and gathering the extracted terms according to their semantic proximity. Our preference goes towards the linguistic techniques.

The choice of linguistic analysis is justified by the fact that this method makes it possible to prevent the polysemia and raises ambiguities related to the context [6]. It also deals with small textual units [11]. Moreover, our learning database contains few examples, it then seems difficult to use the training techniques described above. Our work is carried out in collaboration with experts of the legal field, which is an additional reason to use the linguistic techniques. We use a data processing sequence inspired by [1].

After a standardization phase (conservation or not of the capital letters, proper names, etc.), we carry out a grammatical tagging of the corpus using BRILL [3]. The goal of this operation is to associate to each word its grammatical tag. The next operation is the terminology extraction. We use for this a tool called EXIT [14]. Terminology extraction passes by the search for candidate-terms. Candidate-terms are sets of two or more adjacent units (lexical, syntatcic or semantic). We group then the candidate-terms, extracted according to their semantic proximity, in order to locate the concepts present in the texts.

In order to choose the candidate-terms and to perform concepts creation, we use the same methodology as [2]. We distinguish two stages :

1) First, we study the terms whom appearance frequency is higher than a threshold *l*. Initially, we fix a high threshold. This preliminary stage makes it possible to locate the main conceptual axis.

2) The semantically close candidate-terms are gathered using tools such as WordNet [10]. The recourse to the expert is of primary importance here.

The phases 1 and 2 are iterative, we increase very quickly the representation by examining the candidate-terms whom appearance frequency in the corpus is lower than the threshold *l*.

B. Vector Encoding

This level raises the problem of the choice of the representation model. We choose the vectoriel model [15] which appears to us more adapted than the Boolean model. The reason is that it seems simplistic to apply a binary logic to an information search. The vectoriel model proposes to represent a document on the dimensions represented by the words. We adapted it to represent a document by a vector of concepts. And, instead of represent it according to the frequency of the concept in the document, we use weighting TF x IDF[16].

$$TF \times IDF_{X_i(\omega)} = TF_{X_i(\omega)} \log\left(\frac{n}{DF_{X_i}}\right)$$

With:

- X_i a concept;
- ω a document;
- $TF_{X_i(\omega)}$ the absolute appearance frequency of the concept in a document ω ;
- DF_{X_i} the number of documents in the learning dataset containing the concept X_i;
- n the number of documents in the learning dataset.

At the end of this operation, we obtain a learning dataset which we represent in a tabular form.

We will now use the TF x IDF scores (which can be calculated on all the texts) in order to predict violations in unlabelled texts.

V. Modelisation and generalization tools

We have defined the representation space of our documents. The objective is now to use training algorithm with an aim of automatically classifying the documents. In our study, we are brought to classify multi-labelled texts. Two possibilities are then offered: a global approach and a binary approach consisting in dividing the global problem into *m* subproblems.

We performed the two approaches. We use relative neighborhood graph in the global approach and decision trees in the binary approach.

A. Global approach

The representation space created is a low dimensionnality space. Then, it seems reasonnable to use classifiers based on similarity measures. We choosed proximity graphs which comes from computationnal geometry [12]. Proximity graphs have some advantages on k-NN. They define better the proximity between two items [5]. There are several graphs models. We choosed relative neighborhood graph which is a compromise between number of neighbors and algorithmic complexity.

We apply a simple decision function. The main idea is that an unlabelled text inherits from the properties of its neighbors contained in the training database. Let K be the number of neighbors of the text to be labelled ω' and c_i a rule. The probability that the new text to be labelled contains a violation of c_i is:

$$P(c_i \mid \omega') = \frac{\textit{Number of neighbors of } \omega' \textit{ which contains a violation of } c_i}{K}$$

B. Binary approach

The objective here is to predict the presence or the absence of each rule. We create consequently as many trees as there are rules. More formally, we consider each rule as a boolean attribute $c_i = \{0,1\}$. If the convention v_a contains k rules, we create k trees. Each tree is then a model M_i envisaging the presence or the absence of a rule c_i. We obtain thus k models which return c_i if rule *i* is considered violated, $\varnothing$ elsewhere. Discrimination is carried out by the C4.5 algorithm[13].

VI. Results and methods

We will illsutrate the obtained results on the Convention n°98. Our labelled dataset is composed by only 65 texts. We consequently decide to entirely preserve it during the execution of the training algorithms. In order to have a more precise idea on the real error rate, we carry out 200 bootstrap replications of the training sample. Results are illustrated hereafter.

A. C4.5 results

TABLE 1 C4.5 RESULTS

	Learning			200 bootstrap		
	% well classified	recall	precision	% well classified	recall	precision
Violation 1	95%	100%	91%	94%	93%	94%
Violation 2	98%	96%	100%	98%	96%	99%
Violation 3	100%	100%	100%	95%	42%	100%
Violation 4	98%	86%	100%	89%	38%	96%
Violation 5	91%	91%	90%	87%	86%	87%
Violation 6	94%	93%	95%	80%	78%	81%
Violation 7	98%	80%	100%	91%	36%	95%
Violation 8	89%	94%	83%	86%	90%	82%
Violation 9	94%	79%	100%	89%	82%	92%
Violation 10	100%	100%	100%	96%	30%	99%

We observe excellent classification rates in training (between 89% and 100%). Sometimes, one observes a collapse of the recall at the time of the bootstrap. The reason of this collapse is simple. Indeed, the training dataset used here comprises few examples. It thus happens that some violations are few represented in the dataset. It is consequently difficult for C4.5 to determine the "profiles" of texts containing these violations. On the other hand, the results are very encouraging in the recognition violations where manpower presence-absence are balanced enough. The good classification rates remain stable on the bootstraped samples.

B. Comparison of the two classifiers

We do not have a sample test with labelled texts for the previously quoted reasons, however, we can note a strong similarity between C4.5 and RNG (Table 2). To evaluate the predictive similarity between the two methods, we have extracted 20 unlabelled texts and we have applied the RNG and C4.5 with an aim of predicting the violations contained in these texts.

TABLE 2 COMPARISON BETWEEN RNG AND C4.5

	% agreement					
	100%	80%	75%	66%	50%	33%
Nb of countries	9	1	3	3	2	2

*there is agreement if the two methods identify the same violation in the same text

In spite of the low size of our learning dataset, the two methods show rather similar results (sixty percent have a percentage of agreement equal to or higher than 66%), which are encouraging. We have consequently strong reasons to think that the increase in the number of examples will result in an even more significant agreement.

VII. Conclusion and Future Work

The finality of this work is to propose a predictive model able to determine the violations of several countries concerning two conventions of law the labour. A machine learning approach was adopted. Initially (data preparation), we extracted, thanks to the linguistic techniques, a set of candidates-terms which then allow us to create concepts related to the studied corpus. The purpose of this operation is to reduce the dimensionnality of the representation space of the texts. We were able thus to use the proximity graphs, in addition to C4.5. The results seem to be interesting insofar as the two prediction methods give similar results in spite of a training dataset comprising few examples. We now plan to increase the learning database size with an aim of improving the prediction and of leading to more robust results.

References

[1]Amrani. A, Azé. J, Heitz. T, Kodratoff. Y, Roche. M (2004). From the texts to the concept they contain : a chain of linguistic treatments. TREC 2004, p.1-5.

[2]Baneyx.A and Charlet.J and Jaulent.M.C. (2005). Construction d'ontologies médicales fondée sur l'extraction terminologique à partir de textes : application au domaine de la pneumologie. Journées Francophones d'Informatique Médicale (JFIM) 2005, p.1-6.

[3]Brill. E. (1995). Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. Computational Linguistics, 21(4):543-565.

[4]Bourigault. D, Jacquemin. C. (1999). Term Extraction + Term Clustering: An Integrated Platform for Computer-Aided Terminology . Proceedings of the European Chapter of the Association for Computational Linguistics (EACL'99), Bergen, pages 15-22.

[5]Clech. J. (2004). Contribution Méthodologique à la Fouille de Données Complexes. Thèse, Laboratoire ERIC, Université Lumière Lyon 2.

[6]Fluhr C. (2000). Indexation et recherche d'information textuelle. Ingénierie des langues. Hermes, 2000

Institute,University of Calgary.

[8]Hajek. P, Havranek. T, Chytil. M (1983) GUHA Method. Academia, Prague.

[9]Kumps. N, Francq.P, Delchambre. A. (2004). Création d'un espace conceptuel par analyse de données contextuelles. JADT 2004, (International Conference on Statistical Analysis of Textual Data), p.683-691.

[10] Miller. A. G, Fellbaum.C, Tengi. R, Wolff.S, Wakefield.P, Langone. H, Haskell. B (2005) WordNet 2.1, Cognitive Science Laboratory, Princeton University.

[11]Pouliquen. B, Delamarre. D, Le Beux. P. (2002). Indexation de textes médicaux par extraction de concepts, et ses utilisations. JADT 2002, p.617-627.

[12]Preparata. F, Shamos. M. (1985). Computational Geometry An Introduction. New-York, Springer.

[13]Quinlan. J.R. (1993). C4.5 : Programs for Machine Learning. San Mateo, CA

[14]Roche. M and Heitz.T and Matte-Tailliez.O and Kodratoff.Y. (2004). EXIT: Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés. Proceedings of JADT 2004 (International Conference on Statistical Analysis of Textual Data), volume 2, p. 946-956.

[15]Salton G. (1971) The SMART retrieval system. Experiment in automatic document processing. Prentice Hall. Englewood Cliffs. New Jersey.

[16]Salton G. et Buckley C. (1988). Term weighting approaches in automatic text retrieval. Information Processing and Management, (l) 24, n° 5, p. 513-523.

[17]Toussaint. G.T. (1980). The relative neighborhood graphs in a finite planar set. Pattern recognition, 12:261–268.