# Genre and Domain Processing in an Information Retrieval Perspective

Céline Poudat[1] and Guillaume Cleuziou[2]

[1] CORAL, Centre Orléanais de Recherche en Anthropologie et Linguistique
Orléans - FRANCE
celine.poudat@univ-orleans.fr

[2] LIFO, Laboratoire d'Informatique Fondamentale d'Orléans
Orléans - FRANCE
guillaume.cleuziou@lifo.univ-orleans.fr

**Abstract.** The massive amount of textual data on the Web raises numerous classification problems. Although the notion of domain is widely acknowledged in the IR field, the applicative concept of genre could solve its weaknesses by taking into account the linguistic properties and the document structures of the texts. Two clustering methods are proposed here to illustrate the complementarity of the notions to characterize a closed scientific article corpus. The results are planned to be used in a Web-based application.

## 1 Introduction

The exponential development of the Web has brought about the massive production of extremely diverse textual data which are very difficult to handle. Among the numerous developed for Machine Learning heuristics, the joint use of two corpora is often tried. The first corpus is closed and serves as a training corpus whereas the second is open.

In the prospect of the construction of a Web Information Retrieval (IR) application, the working out of a training corpus has to be seriously thought out. In that respect, we assume that the notion of domain, widely used in IR, might be associated to that of genre. A comparison of two different methods of document clustering, one based on morphosyntactic variables and the other on words has been led on a French corpus of linguistic scientific research articles.

In this paper, we present the results obtained by the coupling of these two approaches on a corpus of restricted size. The conclusive first results we obtained brought us to consider extending the method on more voluminous data within the framework of a domain and genre-based Web application.

## 2 Relevance of a Domain and Genre Coupling in IR

Although it is accepted that IR with textual data has to work with texts rather than with sentences or words, the various variables of discourse analysis (domain,

J.M. Cueva Lovelle et al. (Eds.): ICWE 2003, LNCS 2722, pp. 399–402, 2003.

genre, register, document typology, document structure, etc.) are regrettably little defined and overlap largely. The common assumption that a set of domains (or discourse fields) describing a particular field of knowledge would exist is often worked out in terms of sub-languages and explains the growing success of ontologies. However, the hypothesis is not really founded because it does not allow to solve polysemy nor does it take into account social conditions surrounding texts nor their structure, the latter being essential for IR.

Even if we do not question the importance of domains as regards the processing of the lexicon, and their importance in IR contexts, the notion has to be refined. The notion of genre, which is more and more common in corpus linguistics, turns out to be quite useful. Among the numerous works trying to connect specific features of language with certain types of writing or styles, few studies aim at distinguishing the level of variation of a genre. Yet genres implement recognizable formal processes and they are by definition part of social practices. Consequently, texts of common genre can be indexed in a very relevant way with a set of linguistic and structural variables, obtained thanks to a socially relevant genre corpus [4].

As genres attest specific structures and regulated linguistic properties, they may be implemented within generic profiles, obtained from a characterization of their specificities. Thanks to the coupling of the applicative concepts of genres and domains, i.e. two possible information classifications, we may certainly obtain more conclusive results than the existing systems.

## 3 Materials and Method

### 3.1 The Corpus and Its Constitution

The corpus has been built up according to the domain and genre variables. The choice of the scientific article, considered as a genre, has been determined because of its bureaucratic and structural characteristics, and of its strategic interest in IR and scientific watch applications. The corpus has been reduced to articles belonging to the linguistic field in order to increase domain homogeneity.

247 scientific articles (34 volumes taken from six different French linguistics journals) belonging to the linguistic field have been chosen. As the impact of languages on genres and domains is still little known, the corpus has been restricted to the French language. Texts were all issued between 1994 and 2001 to limit the possibilities of diachronic variations. The notion of sub-domain was not taken into account. Even if we suppose that they will emerge from the soft clustering presented farther, they might emerge from the morphosyntactic-based clustering.

### 3.2 Methodology

*Clustering based on a morphosyntactic labeling of the texts.*
The inductive typological methods [3] used by Corpus Linguistics have been used as they demonstrate that local and global variables can be connected [4]

[5]. Thus it becomes possible to validate or invalidate the intuitions we tend to have on the presence or absence of linguistic markers in certain genres of texts.

Following the example of Biber [1], the use of a quantitative approach seems to be heuristically relevant. The application of multidimensional statistics on genres leads to sound results that can constitute the foundations of a further analysis to check predefined typologies[4].

The 247 texts were labeled by the Cordial© tagger with a set of 198 morphosyntactic variables. The labeled results were then processed by the SAS© system. A factorial analysis (Principal Component Analysis) coupled with a hierarchical ascending classification (Ward method) was used to classify the individuals. The number of clusters was incremented by five, from five to fifty.

*Concept-based document clustering.*
Recent works have demonstrated that using semantic classes of words (concepts) in the conceptual (or thematic) organization process of a document-based corpus leads to a relevant improvement of the group's final quality and homogeneity [6]. Moreover, the method enables to characterize documents in a limited space, as features focus on word groups rather than on words. The process is divided into two major steps: the construction of semantic word classes and the use of a document clustering algorithm.

The semantic word class construction method differs in two main points from the previous studies: (1) the Web as a contextual resource [2] is more efficient than any training corpus or ontology, (2) the definition of a robust clustering algorithm generating non disjoint clusters (a same object can belong to several classes). Let us underline that the use of soft clustering techniques allows to take polysemy and thematic plurality into account.

The frequent terms of the corpus are extracted and the semantic relations between the terms are quantified from their common use in Web documents via the Mutual Information measure. This measure provides a similarity matrix over the words, which is changed into a correlation matrix which constitutes the entry of the clustering algorithm (TAB.1).

| **Table 1.** | Soft-clustering Algorithm |
|---|---|
| **Input:** | The matrix of similarities between objects |
| **Initialisation:** | Each object is allocated to one cluster |
| **Loop:** | Object extraction from a group |
| or | Object adjunction into a group |
| **Until:** | No adjunction nor extraction is possible |

This ongoing clustering algorithm offers three main advantages: (1) as mentioned before, it is a soft clustering algorithm, enabling to obtain non disjoint groups; (2) contrary to other classical hierarchical algorithms, the groups can be corrected in order to improve their homogeneity whereas a hierarchical representation of clusters is made possible and (3), contrary to most of the existing algorithms, the algorithm is not defined by any factor and there is no a priori decision on the final number of clusters.

Word or concept classes are finally obtained. The documents are represented under a vectorial form with a component for each extracted concept. The occurrences of the concept words are added to each component. A conceptual similarity matrixof the documents, which will be given to the same clustering algorithm, is then obtained from the correlations between the documents vectors. The conceptual characterization step enables to reduce the documents description space. As with the clustering method a same document can belong to different classes, the hypothesis according to which a document can handle different themes is taken into account.

## 4 Results and Perspectives

The very first results obtained from the two different classification methods are on the whole different. This could validate our first hypothesis according to which genres and domains are additional notions which could be jointly used in the framework of an IR application.

In fact, thanks to the concept-based clustering document, it is possible to extract cohesive conceptual groups[1] so as to form thematically relevant document clusters[2]. Thus, the singletons deal with very specific research objects. For instance, we found a text containing a high number of theater play extracts, which was pushed aside by both methods. The genre and domain notions are additional and their coupling allows to obtain a fine characterization of the corpus. This may certainly contribute to increase the quality of the results of a scientific text query on the Web.

We presented here the first incomplete results of an ongoing research, which will be extended to other scientific domains. The structural properties of genres will be taken into account. A similar work on English scientific articles is in progress and the results will be assessed on the Web.

## References

1. BIBER, D. *Variation across Speech and Writing.* Cambridge University Press, Cambridge, 1988.
2. CLEUZIOU, G., CLAVIER, V. and MARTIN, L., *Organisation conceptuelle de mots pour la Recherche d'Information sur le Web.* Conférence francophone d'Apprentissage CAp'02, 2002.
3. KARLGREN, J. *Text genre recognition using discriminant analysis.* International Conference on Computational Linguistics, 1994.
4. MALRIEU, D. et RASTIER, F. *Genres et Variations morphosyntaxiques* in TAL, Vol. 42, num. 2/2001.
5. POUDAT, C. *Characterization of French linguistic research articles with morphosyntactic variables* in Academic discourse - multidisciplinary approaches, 2003.
6. SLONIM N. and TISHBY N. *The power of word clusters for text classification.* In 23rd European Colloquium on Information Retrieval Research. 2001.

[1] Here are some examples of the concepts we obtained: Amphitryon, Molière, Plaute, Sosie/Charcot, schizophrène, etc.

[2] Documents were besides very correlated (0.95 on average).