Text Categorization prior to Indexing for the CISMEF Health Catalogue

Alexandrina Rogozan¹, Aurélie Néveol¹², and Stefan J. Darmoni¹²

¹ PSI Laboratory - FRE 2645 CNRS - INSA de Rouen, BP8 avenue de l'Université 76801 Saint-Etienne-du-Rouvray Cedex, France  
{alexandrina.rogozan,aurelie.neveol}@insa-rouen.fr  
² CISMeF et L@stics - Rouen University Hospital and Rouen Medical School 1 rue de Germont, 76031 Rouen, France  
{stefan.darmoni}@chu-rouen.fr  
http://www.chu-rouen.fr/cismef

Abstract. This paper is positioned within the development of an automated indexing system for the CISMeF quality controlled health gateway. For disambiguation purposes, we wish to perform text categorization prior to indexing. Hence, a global approach contrasting with the classical analytical methods based on the analysis of keyword counts extracted from the text is necessary. The use of statistical compression models enables us to proceed avoiding keyword extraction at this stage. Preliminary results show that although this method is not as precise as others in terms of resource categorization, it can significantly benefit indexing.

1 Introduction

Internet has become a very prosperous source of information in numerous fields, including health. The CISMeF project (French acronym of Catalogue and Index of Medical On-Line Resources) was initiated in 1995 in order to meet the users’ need to find precisely what they are looking for among the numerous health documents available online. As a Quality Controlled Health Gateway [1], CISMeF describes and indexes the most important resources of institutional health information in French. It currently contains more than 12,000 resources, and it is updated manually with 50 new resources each week. Indexing is a decisive step for the efficiency of information retrieval within the CISMeF catalogue, and it is also one of the most time consuming tasks for the librarians, demanding high-level documentary skills.

Our research work aims to develop an automatic indexing system that would broaden the CISMeF catalogue coverage while ensuring good indexing quality and achieving high precision and recall rates for information retrieval within CISMeF. For a better approach of automatic indexing, we wish to perform text categorization as a preliminary task.

In fact, the knowledge of the resource medical specialty, we also called context or category, will have a doubly important role in the indexing phase: 1. it will help lexical disambiguation (Pouliquen [2] explains how a lack of such disambiguation leads to systematic indexing errors. For example, several occurrences of the term lutte in a
resource could be related to either MeSH terms *Wrestling* or *Prevention & Control*. Now, if the context is *Sports Medicine* it is highly likely that the appropriate MeSH term is *Wrestling*). It will give more weight to the context related keywords, therefore bringing out the gist of the resource content.

After reviewing the existing methods of text categorization in section 2, a set of medical contexts based on the CISMeF terminology is defined in section 3. Then, a text categorization methodology based on compression models is presented, ongoing experiments are detailed, and their contribution to text categorization is discussed in section 4.

## 2 Global vs Analytical Methods for Text Categorization

Early work of Wiener et al. [3] shows that neural networks and logistic regression are appropriate approaches for topic spotting in documents. Among recent statistical approaches for text categorization, the Support Vector Machines (SVM) are emerging as they provide higher precision than four other learning algorithms, including naïve bayes, bayes nets and decision trees in an experiment conducted by Dumais et al. [4]. However, SVM performances in multi-class problems are limited in terms of speed and algorithm complexity.

Other strategies consist in combining both statistical and linguistic approaches. For instance, Wilcox et al. [5] use data mining and natural language processing tools to extract a pertinent representation of documents, and statistical methods, viz. rule generation, bayesian classifiers, and information retrieval for their categorization. Wilcox’s results confirm that using explicit domain knowledge when available, is the best methodology, because succeed at the best results. Indeed, in recent work we implemented a ruled based algorithm using the semantic properties of the CISMeF terminology for categorization purposes, and obtained 80% precision and 93% recall [6].

However, these categorization techniques, as well as other analytical techniques reviewed by Kosala [7] involve a preliminary representation of documents (i.e. bag of words). The significant words extraction is clearly redundant with the indexing process, and our goal is to identify the context prior to keyword extraction and indexing. This constraint leads us to choose a global approach. Teehan and Harper [8] show that statistical compression models, and in particular PPM (Prediction by Partial Match) models, have performances comparable to those of SVM for text categorization, while using a global approach. Therefore we have decided to adapt them to health resource categorization within CISMeF, after defining a set of medical contexts based on the CISMeF terminology.

## 3 Medical Context Set Based on CISMEF Terminology

In order to identify to which context(s) a given resource belongs, i.e. which medical specialty(ies) it deals with, we need to define a set of medical specialties that would be both complete and relevant for indexing purposes.