Chapter 56
Towards a Statistical Approach to the Analysis, the Indexing, and the Semantic Search of Medical Videoconferences

Ameni Yengui
University of Sfax, Tunisia

Mahmoud Neji
University of Sfax, Tunisia

ABSTRACT

In this article, the authors introduce their OSSVIRI information retrieval system which composed of three modules. In the analysis module, they have proposed a statistical technique exploiting the word frequency in order to extract the simple, compound and specific terms from the documents. In the indexing module, the authors used the ontology to associate the terms with their concepts, retrieve the relations between them and disambiguate the concepts to improve the sematic content of the documents. The concepts and relations are represented as a conceptual graph. In the research module, the authors have proposed a technique of users' query reformulation based on external resources and users' profiles and a technique of pairing based on the combined expansion of the requests and the documents guided by the context of the requirement in information and the documentary contents. This system is validated using the metrics from the research information and comparisons with existing statistical approach. The authors show that their approach achieves good results.

1. INTRODUCTION

Nowadays, there is a continuous development of information technology. These new technologies enhanced the rapid development of material production technology and information management. The progress of production of information tools, such as video conferencing, has enabled the production of

DOI: 10.4018/978-1-5225-5191-1.ch056

a huge amount of information. This rapid increase in the volume of information has created the problem of "how to find information that interests us in this great mass of information"

To treat this problem, the IRS has been developed in order to select from a volume of information, the pertinent information vis-à-vis an information need. These SRI aim to connect two representations; one is of the user's needs, and the other of the document content using a correspondence function.

A semantic information retrieval system is essentially composed of three modules: the analysis module, the indexing module and the search module. We found that indexing is essentially an interpretative task of analysis and reformulation that requires a lot of knowledge and expertise. In addition, the result of a search -in the index base- should consider all the "imprecise" the user's needs. The existing problem is then how to keep the semantic continuity and give exact answers to the users' needs.

Our objective is to represent every document and every request by a set of concepts that make up a language core of the representation of information and which are defined in an external semantic resource (RS). We therefore propose a statistical method for detecting these concepts. The extraction of the terms denoting these concepts is made on the basis of a support corpus to extract the simple terms and on a statistical measure to retrieve the compound terms. Next, these terms are transformed into concepts using the semantic resources. To meet the users' need, the request representation is compared to that of each document with a similarity measure. The relevant documents are classified according to a measure of relevance.

In this paper, first, we present the state of the art on the indexing methods using external semantic resources, methods of reformulations of queries and pairing query / document methods. Second, we explain the different stages of our approach. Third, we present an experimental validation of our approach and, finally, we end up with a conclusion.

2. BACKGROUND

The modules that are mostly studied are the indexing and the information retrieval. In this part, we deal with some previous work.

2.1. Indexing

To index is a primarily interpretative task of analysis and reformulation that requires a lot of knowledge and expertise. In addition, the search results -in index base- should consider all the "imprecise" users' needs. The existing problem is how to keep the semantic continuity and to give exact answers to users' needs.

The indexation of audiovisual documents is a major challenge for many actors and users of audiovisual materials, such as journalists, documentary makers, analysts of movies, film historians, teachers of cinematography, etc. The implementation of this index is often considered unsatisfactory from a qualitative point of view. Different actors in the field of broadcasting employ documentalists for manual and textual indexation in their funds, which is time consuming and too informal to be fully exploited by an information system.

The state of the art recommended to control this semantic continuity is the use of documentary languages. The latter make it possible to control the expressions serving as an index as well as to provide reference frameworks which guide the interpretation and indexing.