Abstract

Today we face a fundamental problem in information retrieval. Multi-lingual information resources are becoming accessible throughout the world at an unprecedented rate. Although research has demonstrated that new algorithms and user interfaces are desirable to deal with the particularities of multilingual corpora, we are barely beginning to see effective results.

Starting from our previous work on InfoRadar, this paper describes the preliminary design of InfoRadar-cl, a prototype search system supporting the anticipatory user interface supported by InfoRadar, but redesigned to support multilingual queries and results. A selected collection of refined queries in different languages is displayed to the user as an interactive query hierarchy and multi-lingual documents are automatically classified using queries as categories.

This paper also briefly reviews the current state of the art in multilingual information retrieval, presents some methods for query translation, and some details about semantic disambiguation that we plan to incorporate into the InfoRadar-cl design.

1. Introduction

Information retrieval is “the science of matching information needs to documents” (Brinton, 1991). Our research addresses an inherent problem in information retrieval. Formulating precise and effective information retrieval queries has always been a difficult task, even for experienced users. Hence, simple queries often return extremely large and imprecise result sets when used with large corpora.

The Cross-Language information retrieval problem consists of providing the best possible set of multi-lingual relevant documents in response to a query expressed in one language.

Initially, we presents InfoRadar, the first anticipatory user interface for search result visualization that exploits Query Lookahead, a technique that eagerly evaluates refined queries automatically generated from an initial user query in a given language [Velez99].

Query Lookahead has the potential of improving search systems in at least two novel ways. First, it enables the deployment of anticipatory user interfaces capable of presenting the result sets of automatically generated refined queries ahead of time. Refined queries serve as categories upon which a large and imprecise result set can be organized. Second, Query Lookahead has the potential of improving the effectiveness of feature (e.g. term) selection algorithms. These algorithms can be improved by exploiting information about the result set induced by each potential term when combined with the user query.

InfoRadar supports a novel user interaction model based on Visual Query Hierarchies. As shown in Figure 1, the central idea is to organize a result set in a hierarchy using queries as basis for forming the various categories upon which the documents are organized. Visual query hierarchies help users quickly focus on their
particular information needs. The task of formulating an effective query requires the user to predict which terms appear in documents relevant to the information needed. This often requires extensive knowledge about the document collection, which it’s difficult to obtain in large document corpora. In addition, users want to avoid retrieving irrelevant documents due to a query that is under-specified or contains ambiguous terms. As a result, users of an information retrieval system with a large corpus are often faced with the task of manually sifting through very large and often inappropriate result sets. For complementary information about InfoRadar the reader is referred to [Velez97], [Velez99] and [Velez01].

Figure 1: Query Hierarchy Panel showing expanded query nodes

2. A New Scenario: CLIR Approach

The first implementation of InfoRadar was only applicable to English language texts. In the actual scenario, the document collection is conforming by web pages indexed from UPRM University of Puerto Rico at Mayagüez sites. These documents could be written either in Spanish or English language. Therefore, the research conducted during the fall of 2001 focus on Cross-Language Information Retrieval (CLIR).

In this section, we present a review of the important changes in the software. These changes will allow the new multilingual support of the InfoRadar system. The review begins with the important concepts of data indexing. Next is a definition of multilingual interactive query hierarchies. Finally, we present the query translation method used in this approach.

2.1. Indexing sub-system

To process queries, the search engine requires a collection of data structures holding information about the indexed documents. The InfoRadar indexing module generates these data structures. The indexing process is an off-line feature and a transactional procedure.

Figure 2. Indexing phases in the InfoRadar-cl indexing module
Initially, there were four phases in the indexing process: scanning, stopping, stemming, and file inversion. In the new approach, we have the same phases but with a few modifications. Figure 2 shows a flow diagram of the indexing process.

The InfoRadar implementation actually integrates scanning and stopping into a single software module. Scanning identifies and separates the terms in each document, and classifies the document in one of the supported languages. Stopping removes terms that are too common to be of any use for retrieval purposes. The indexing module reads a list of stop terms for each language and integrates them into a finite state automaton that recognizes and identifies terms in the input stream. The Stemming phase is language-dependent, we use an implementation of the Porter stemming algorithm for each language. Stemming detects minor variations of the same term and makes sure that only one variation of each term is indexed. Finally, the File inversion phase generates a mapping from each term to the list of documents that contains that term.

2.2. Multilingual query hierarchies

A multilingual visual query hierarchy is a tree of refined multilingual queries together with their individual result sets. A user provides an initial seed query. Refined queries will combine terms appearing in the result set with the terms in the seed query and the translated seed query. The descendants of a query provide a succinct hierarchical categorization of the collection of documents matching both queries.

Hence, the multilingual visual query hierarchy will present a refined set of categories in only one language with related documents in both languages. For instance, if the original query is space, some of the categories in the hierarchy could be booster, disaster and astronaut. Then, in the category astronaut, we find documents that matched the query (space ^ astronaut) but also we find documents that matched the query (espacio ^ astronauta).²

2.3. Query translation method

The main idea is that the user can simply find documents written in any language using queries expressed in one language. Then, we must translate each query term and then perform a retrieval process in the language of the document. We consider three different methods: using a machine translation system, using a bilingual dictionary or terminology base, or using a statistical/probabilistic model based on parallel texts.

Machine translation systems tend to spend a lot of effort trying to produce syntactically correct sentences and also are expected to select one of many possible translations (synonymous or related words). This could be a problem, because is possible often end up selecting the wrong translations. [Winiwarter00] presents an implementation based in the machine translation environment to guarantee high translation performance in cross-lingual information retrieval.

In other hand, we can use a dictionary for query translation also known as a machine-readable dictionary (MRD). Nevertheless, the problem remains, because some words receive many translations that correspond to different meanings. There are many studies about that, but it is still very difficult to determine the correct meaning of the word in a query. The effective use of dictionaries might improve the retrieval performance as demonstrated by [Sadat01] et al.

The third approach is to determine translational equivalence automatically, on the basis of a corpus of parallel texts (a corpus made up of source text and their translations). [Nie99] et al. provides more details about the subject.

In our proposed CLIR system, query translation can be performed with a simple stemming process of query terms. The next step is a term-by-term translation using initially a MRD with the wider coverage possible. Obviously, the quality of the translation depends on the quality of the dictionary. An outstanding dictionary in machine-readable format could be very expensive, so, we opted to use a free dictionary available on the Internet. The interesting thing in this dictionary is that new translations can be added.

² The ^ symbol represents logical conjunction.
But, we also have to deal with the semantic disambiguation. Semantic disambiguation involves identifying the correct sense or meaning of each term in a particular query. Successful semantic disambiguation is still an important open research topic by itself.

3. Preliminary Results and Future Work

After finishing some updates in the InfoRadar software we have started a first case study with documents entirely written in Spanish. But, the most obvious problem with the lexical structure of written Spanish is its accent rules. Modifying the data structures of valid characters, the Spanish semantic can be enabled.

At the current state it is possible for the user to apply queries in only one language at a time, so it had to give a truncated result set of related documents.

Future work will concentrate in the impact of a cross-lingual coverage in the existing algorithms as Query Refinement and Query Hierarchies, evaluate performance and compare the result against the original monolingual performance. Amongst the aims of the research work is that the user can perform multilingual queries and obtain various categories upon which the documents in different languages are organized. Also, we need apply a semantic disambiguation technique to select the best translation terms from the terms found in the machine-readable dictionary entries.

References


