Method of Lexical Enrichment in Information Retrieval System in Arabic

Souheyl Mallat, Department of Computer Sciences, University of Monastir, Monastir, Tunisia

Anis Zouaghi, Department of Computer Sciences, Higher Institute of Applied Science and Technologies Sousse, Sousse University, Sousse, Tunisia

Emna Hkiri, Department of Computer Sciences, University of Monastir, Monastir, Tunisia

Mounir Zrigui, Department of Computer Sciences, University of Monastir, Monastir, Tunisia

ABSTRACT

In this paper, the authors propose a method for lexical enrichment of Arabic queries in order to improve the performance of the information retrieval systems SRI. This method has two types of enrichment: linguistic and contextual. The first one is based on the linguistic analysis (lemmatization, morphological, syntactic and semantic analysis), whose goal is to generate a descriptive list (list-desc). This list contains a set of linguistic lexicon assigned to each significant term in the query. The second enrichment consists in integrating contextual information derived from the corpus documents. It is based on statistical analysis using Salton weighting functions: TF-IDF and TF-IEF. The TF-IDF function is applied on the list-desc and documents in the corpus in order to identify relevant documents. TF-IEF function is made between the list-desc and sentences belonging to the relevant documents to identify relevant sentences. Then, terms in these sentences are weighted, and those with highest weights are considered rich in terms of informative and contextual importance are added to the original query. The authors' lexical enrichment method was evaluated on a corpus of documents belonging to a specialized domain and results show its interest in terms of precision and recall.

Keywords: Arabic NL, Information Retrieval, Lexical Enrichment, Query Enrichment, Weighting

1. INTRODUCTION

The objective of information retrieval system (IRS) is to retrieve relevant documents that meet the needs of users. The user of such system seeks the precision in the answers, and prefers a small number of documents that meet his needs rather than many which contain the answer but drowned in a set of irrelevant documents (Mitra, 1997). In addition, all informations in documents are not always correlated with the user query. This is why the information retrieval

DOI: 10.4018/ijirr.2013100103

systems (IRS) is an important search area today. It should be noted, that the quality of responses obtained by SRI depends not only on the degree of similarity between query / document but also on the query made by the user. Our work is in the optic of improving the performance of SRI, by using lexical enrichment of queries in the domain of environmental pollution. The enrichment consists in first part, in the addition of morphologically, semantically related terms (synonymy, hypernomy, etc.), and lemmas to the key query terms. On the other part, the addition of contextual information by adding new terms related to the context of the initial query, this contextual addition is based on relevant sentences, which are selected from the enrichment corpus by a method of statistical analysis.

This method is based mainly on composed terms which identify the query. These units are more precise and less ambiguous than simple isolated terms (Boulaknadel, 2008). They facilitate the linguistic and statistical treatments on which is based our enrichment method.

The linguistic treatment of text (user query or text corpus) consists of syntactic, morphologic and semantics analysis:

- The morphological module covers the inflectional and derivational variation of significant terms of the query, in order to increase the number of occurrence associated to these terms for the search;
- The syntactical module is based on grammatical labeling, which associates to each word its grammatical category (noun, verb, adjective, particles ...). The purpose of labeling is tracking the terms simple and composed, in order to operate a first treatment of terms disambiguation;
- The semantic module associates synonymy and hyperonymy relations to significant terms of the query. Extraction of these relations is done by using two dictionaries (simple and composed), which express the corresponding domain relations to the simple and composed terms of the initial query.

The result of the linguistic processing is a description of significant terms (list-desc) in the form of a list containing the significant terms with their semantic and morphological variations. This list provides an improvement to the similarity measures between the query and the documents of the corpus in the statistical processing.

The statistical treatment, consists on the similarity determination between the couples (desc-list of the initial query, the documents), and (desc-list of the initial query, the phrases belonging to relevant documents). This measure is based on the weighting functions of Salton TF-IDF and TF-IEF (Salton, 2008). It is defined as criteria of decreased classification of documents, as well as phrases in terms of relevance. This criteria also assigns a weight for each term of relevant sentences, in order to integrate the terms of highest weights (which express the contextual information) to the initial query.

2. PROBLEMATIC

In this work, we are interested in ambiguities that have a direct impact on information retrieval (IR).

Methods which are based on the keywords as a mean for IR, are considered insufficient: for example if the query and a document share a key term, this document can be seen more or less corresponding to the query subject.

This methods insufficiency is due to the fact that the terms used in the query vary morphologically and semantically, compared to documents in the knowledge base. This variation degrades the effectiveness and the precision of IR systems (Yannich, 2000). These changes affect several levels, for example:

- The query does not cover the morphological variations that generate keywords in different numbers, for example "مدرسة" (school) and "مدرستان" (two schools), "خيل" (horse) and "خيول" (horses);

Metadata for Search Engines: What can be learned from e-Sciences?

Magali Roux (2012). *Next Generation Search Engines: Advanced Models for Information Retrieval (pp. 47-77).*

www.igi-global.com/chapter/metadata-search-engines/64420?camid=4v1a

Fuzzy XQuery: A Real Implementation

José Ángel Labbad, Ricardo R. Monascal and Leonid Tineo (2016). *Handbook of Research on Innovative Database Query Processing Techniques (pp. 158-198).*

www.igi-global.com/chapter/fuzzy-xquery/138696?camid=4v1a