SQX-Lib: Developing a Semantic Query Expansion System in a Media Group

María G. Buey, Angel L. Garrido, Sandra Escudero, Raquel Trillo, Sergio Ilarri, Eduardo Mena

IIS Department, University of Zaragoza, Zaragoza, Spain.
{mgbuey, garrido, sandra.escudero, raqueltl, silarri, emena}@unizar.es

Abstract. Recently, there has been an exponential growth in the amount of digital data stored in repositories. Therefore, the efficient and effective retrieval of information from them has become a key issue. Organizations use traditional architectures and methodologies based on classical relational databases, but these approaches do not consider the semantics of the data or they perform complex ETL processes from relational repositories to triple repositories. Most companies do not carry out this type of migration due to lack of time, money or knowledge.
In this paper we present a system that performs a semantic query expansion to improve information retrieval from traditional relational databases repositories. We have also linked it to an actual system and we have carried out a set of tests in a real Media Group organization. Results are very promising and show the interest of the proposal.

Keywords: information retrieval; query expansion; semantic search

1 Introduction

Search systems in different contexts are adopting keyword-based interfaces due to their success in traditional web search engines, such as Google or Bing. These systems are usually based on the use of inverted indexes and different ranking policies [1]. In the context of keyword-based search on structured sources [2], most approaches retrieve data that exactly match the user keywords and indexed terms. However, users often do not use the same words. So, there exist a term mismatch problem, that reduce the retrieval effectiveness. The main problems are the information lost (low recall) and the retrieval of non-relevant data (low precision).

To deal with this problem, several approaches have been proposed that are applicable in *Automatic Query Expansion (AQE)* solutions: interactive query refinement, relevance feedback, word sense disambiguation, and search results clustering. One of the most natural and successful techniques is to expand the original query with other words that best capture the actual user intent, or that simply produce a more useful query that is more likely to retrieve relevant documents. The survey presented in [3] provides a classification of existing techniques that leverage several data sources and employ sophisticated methods for finding

new features correlated with the query terms, and firmer theoretical foundations and a better understanding of the utility and limitations of AQE. In the last years, basic techniques are being used in conjunction with other mechanisms to increase their effectiveness, including the combination of methods, more active selection of information sources, and discriminative policies. AQE is currently considered a promising technique to improve the retrieval effectiveness of document ranking and it is being adopted in commercial applications, especially for desktop and intranet searches. Although AQE has received a great deal of attention and several approaches have been proposed, very little work has been done to review such studies.

Under these circumstances, we have developed a library called SQX-Lib (Semantic Query eXpansion Library) that exploits linguistic and semantic techniques to improve the quality of the keyword-based search process on the relational repositories of the Heraldo Media Group[1]. A demonstration of its operation (including a video, snapshots, etc.) can be seen at `http://sid.cps.unizar.es/SEMANTICWEB/GENIE/Genie_Downloads.html`. SQX-Lib is focused on automatically and semantically expanding the scope of the searches, and fine-tuning them by taking advantage of Named Entities (NE) present in the query string. For this purpose, it uses lexical resources and dictionaries, and an unsupervised disambiguation method that looks for the accurate meaning of a word.

2 System overview

SQX-Lib can be included in a search engine or browser. The library expands and enriches semantically a given set of user keywords to improve the search process. This library can be used with any data source, as the expanded set of keywords obtained can be translated into a query in the appropriate query language of the data source, such as SQL in our case study, where relational databases are used. The semantic expansion process (Fig. 1) performs three main tasks:

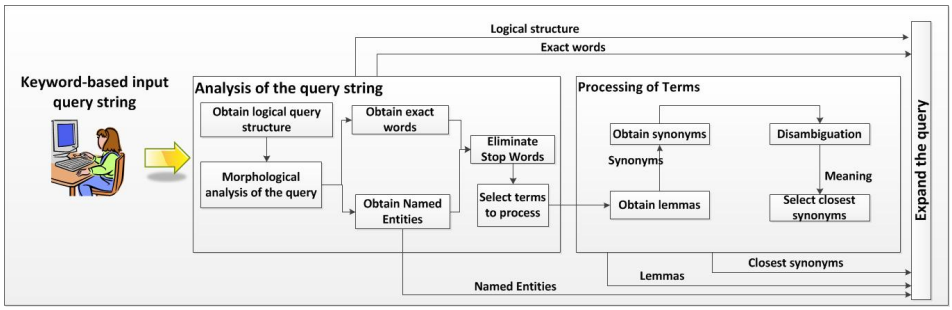

Fig. 1. Processing steps of SQX-Lib

[1] http://www.grupoheraldo.com/

1. *Analysis of the keywords of the query*: SQX-Lib analyzes the keywords introduced as a query. First, it obtains the logical query structure to perform an appropriate construction of the query. It processes the logical operators, parentheses and quotes that could appear in the search string. Then, SQX-Lib performs a morphological analysis of the query string using an NLP tool. Afterwards, the process carries out a second analysis of the query string in order to obtain all the NE. Moreover, it also performs the analysis of the query to obtain exact words, that is the words that have been introduced quoted. These words are searched as they have been written in the query string and they are not be processed to find their semantics. After this, SQX-Lib removes stop words. Finally, it selects the terms (common names or verbs in our case) that are going to be processed to find their semantics to enrich the input query.
2. *Processing of terms*: SQX-Lib obtains the lemmas of the terms chosen in the previous step using an NLP tool. Then, it searches the sets of synonyms for each term selected, by using the lemmas and semantic resources (dictionaries and lexical thesaurus). In our case, the lexical thesaurus used has been EuroWordNet [4] which is a multilingual resource for several European languages. If the terms have more than one set of synonyms, SQX-Lib performs a disambiguation process to select the most appropriate according to the context, and taking into account an adaptation of the Normalized Google Distance (NGD) [5] for a document repository.
3. *Expand the query*: SQX-Lib reconstructs and expands the query from the relevant data extracted in each of the previous tasks while keeping the initial logical structure of the query.

3 Experimental Results

The aim of this work is to improve the user experience on a specific search system. The present case is a search engine that is working over the news repository used everyday by all the departments of the Heraldo Group, a leading Spanish media. The data repository used by the documentation department is a relational database with about 10 million records. The contents and its metadata are in plain text format, and in Spanish. Text fields are fully indexed so that the search engine can quickly locate records using a query based on keywords. Every day the system receives an average of almost one thousand queries.

The standard search mode consists of the introduction of a list of keywords separated by spaces and boolean operators. For experimental evaluation, we have analyzed 29353 real queries, with the following results: 64% of the queries include NE and can be optimized using SQX-Lib, 61% of the queries could take advantage of the lemmatization step, 58% of the queries include one or more common names from which the system is able to obtain synonyms and related words, and finally 13% of the queries could lead to build an incorrect query. So, we have found that about 95% of the queries made to this repository are expandable/optimized by our system.

If we analyze the results of these queries, we obtain the following conclusions: The lemmatization expansion over the query leads to an average improvement of 50% in the recall, the synonym expansion step translates into obtaining about double results on average, and the use of an automatic filter when we found a NE implies a reduction of approximately 20% of the noise. If we use all the options together, we obtain almost four times more results than using the original query, so SQX-Lib successfully minimizes the documentary silence problem. Besides, this is not done at the expense of introducing more noise.

We have carried out an opinion survey among the workers of the Heraldo Group about their use of the system improved with SQX-Lib:

- 70% of the users would include the lemmatization expansion. The rest of users would prefer that this feature be optional.
- 58% like the synonym expansion and they would include it by default.
- 88% agree about the need of a default mechanism to filter the results using the name entities embedded in the query.

In general, all the users are satisfied with the new functions although not all of them agree about whether they must be optional or applied by default. This is due to the different searching style of each department, which in turn is caused by the different types of information that they want to find. In some cases there are few records and it is preferable to use mechanisms to extend the results from the beginning, and in others the opposite occurs because the noise is initially excessive even with a simple query.

4 Demonstration

Our demo will allow to perform queries against an extract of the documentary database, it will show the expanded query, and it will compare the results of the query without expanding with the answers of the query once expanded.

Acknowledgment

This research work has been supported by the CICYT project TIN2010-21387-C02-02 and DGA-FSE. Thank you to Heraldo Group and Jorge Gracia.

References

1. R. Baeza-Yates, B. Ribeiro-Neto, *et al.*, *Modern information retrieval*, vol. 463. ACM press New York, 1999.
2. J. X. Yu, L. Qin, and L. Chang, "Keyword search in databases," *Synthesis Lectures on Data Management*, vol. 1, no. 1, 2009.
3. C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *ACM Computing Surveys (CSUR)*, vol. 44, no. 1, 2012.
4. P. Vossen, *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic Boston, 1998.
5. R. L. Cilibrasi and P. M. Vitanyi, "The Google similarity distance," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 370–383, 2007.