An approach for automatic query expansion based on NLP and semantics

María G. Buey, Ángel Luis Garrido, and Sergio Ilarri

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

Abstract. Nowadays, there is a huge amount of digital data stored in repositories that are queried by search systems that rely on keyword-based interfaces. Therefore, the retrieval of information from repositories has become an important issue. Organizations usually implement architectures based on relational databases that do not consider the syntax and semantics of the data. To solve this problem, they perform complex Extract, Transform and Load (ETL) processes from relational repositories to triple stores. However, most organizations do not carry out this migration due to lack of time, money and knowledge.
In this paper we present a methodology that performs an automatic query expansion based on natural language processing and semantics to improve information retrieval from relational databases repositories. We have integrated it into an existing system in a real Media Group organization and we have tested it to analyze its effectiveness. Results obtained are 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 classical relational databases, most approaches retrieve data that exactly match the user keywords [2]. However, they do not consider the semantic contents of the keywords and their relationships. This can lead to losing information (low recall). Moreover, search engines return needless data, that are not interesting (low precision). This problem has received a great deal of attention and several approaches have been proposed to solve it, that are applicable in *Automatic Query Expansion (AQE)* solutions. 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. 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. However, very little work has been done to review such studies.

Under these circumstances, we have developed a methodology called SQX-Lib (Semantic Query eXpansion Library). This paper is an extended version of a demo presented in [3]; the extensions include a detailed description of the methodology and a study of an example of its application in a real environment. SQX-Lib has been implemented as a library to encapsulate its functionality so that it can be fit in any environment. It is focused on automatically and semantically expanding the scope of the searches, and fine-tunes them by taking advantage of the Named Entities (NEs) present in the query string. In more detail, SQX-Lib performs three main tasks: *1) Obtain words with common lemmas*, to extract those words that belong to the same family as the keywords entered as a query. *2) Obtain words with similar meanings*, to return records which contain words that are synonymous to the keywords introduced. *3) Refine the queries with named entities.* NEs are considered as a whole. For example, if we search "Real Madrid" the search engine does not return data that contain the words "Madrid" and "real", but those related to that Spanish football team. To carry out these tasks, we have used a parser and a POS (Part of Speech) tagger as natural language processing (NLP) tools, a NE repository, a lexical database and a set of dictionaries for obtaining the meanings and the synonyms of the keywords, and a disambiguation engine designed to eliminate ambiguities.

Therefore, this paper provides two main contributions: Firstly, we present a new multilingual semantic query expansion methodology for information systems based on relational databases; and secondly, we have experimentally tested the use and the impact of our system in a real company environment and we have surveyed the opinion of the users.

This paper is structured as follow. Section 2 describes the state of the art. Section 3 explains the general architecture of our solution and presents the proposed algorithm. Section 4 discusses the results of our study with real data. Finally, Section 5 provides our conclusions and some lines of future work.

2 Related Work

The relative ineffectiveness of information retrieval (IR) systems is largely caused by the inaccuracy of a query formed by a few keywords, and AQE is a well-known method to overcome this limitation by expanding the original query of the user by adding new elements with a similar meaning. An IR system is likely to return good matches when an user query contains multiple specific keywords that accurately describe the information needed. However, the user queries are usually short and the natural language is inherently ambiguous. This is known as the vocabulary problem [4], compounded by synonymy (different words with similar meaning) and polysemy (a word with different meanings). To deal with this problem, several approaches have been proposed, including: interactive query refinement, relevance feedback, word sense disambiguation, and search results clustering. The survey of Carpineto and Romano [5] presents a wide study of AQE and a large number of approaches that handle various data resources and employ very different principles and techniques.

Most IR systems are based on the computation of the importance of terms that occur in the query and in the documents, and on several widely-used ranking models. Linguistic analysis techniques handle language properties such as morphological, lexical, syntactic, and semantic word relationships, to expand or reformulate query terms. Methodologies such as *corpus-specific global* techniques analyze the contents of a database to identify features used in similar ways [6], and *Query-specific local* techniques take advantage of the local context provided by the query using a subset of documents that is returned with the given query [7]. Other current proposal is [8], that uses relational databases combined with machine learning techniques and semantics. Also, substantial effort has been applied to the task of named entity recognition. A named entity is a word or a set of words that correspond with names of people, places, organizations, etc. Named Entity Recognition (NER) is an important task of information extraction systems and it consists of locating proper nouns in a text [9]. Novel approaches, such as [10], try to find more complex names (e.g., film or book titles) in web text. Knowledge-based proposals rely on the combination of a range of knowledge sources and higher-level techniques (e.g., co-reference resolution). Dictionaries and extensive gazetteer lists of first names, company names, and corporate suffixes are often claimed to be a useful resource. Moreover, one of the problems when working with the semantics of words is that there may be ambiguities when there is more than one possible meaning for each word. Word Sense Disambiguation (WSD) is the ability to identify automatically the meaning of words in a context and it is a natural and well-known approach to the vocabulary problem in IR [11]. Depending on how the *similarity relation* between the words and their meanings [12] is computed, we can distinguish different techniques: *measures based on glosses*, where a gloss is a definition or explanation of a word in a dictionary [13]); *measures based on conceptual trees*, which are based on a tree or is-a hierarchy, such as WordNet [14]; and *measures based on the content of the information.* Several experiments suggested that using WordNet may not be effective for IR [15], at least as long as the selection of the correct sense definition (or synset) is imperfect. However, research based on the work using WordNet for AQE has continued using more sophisticated methods [16].

To sum up, AQE has a long history in IR. However, it still suffers from drawbacks that have limited its deployment as a standard component in search systems. But the limitations have started to be addressed in recent research with the exploration of new directions.

3 Architecture

This section describes the process that performs the semantic expansion of the input keyword-based query and its associated named entity recognition method and disambiguation process. The semantic expansion process carries out three main tasks that we summarize below:

1. *Analysis of the keywords of the query*: first, SQX-Lib performs an analysis of the keywords introduced as a query. This task consists of several steps:

- *Obtain the logical query structure*: it performs an appropriate construction of the query with the introduced terms without losing its structure. This task takes into account and processes the logical operators, parentheses and quotes that could appear in the search string.
- *Morphological analysis of the query*: SQX-Lib analyzes syntactically the query string to obtain a preprocessing of the terms that constitute the introduced query, as well as a possible first recognition of the NEs present in it (in case they have been introduced capitalized). To carry out this task it uses the linguistic functionalities that an NLP tool provides. Specifically, for this purpose we have used the Freeling [17] NLP tool.
- *Obtain named entities*: next, it analyzes again the query string in order to get all the NE that are present. The method is described in Section 3.1.
- *Obtain exact words*: it performs another analysis of the query to search words that have been introduced quoted. These words will be searched as they have been written in the query string and they are not processed to find their semantics.
- *Remove stop words*: then, SQX-Lib eliminates those words from the query string that are considered as noise (articles, prepositions, etc.), as they do not provide any relevant information. The information about the type of each word is provided by the NLP tool.
- *Select terms to process*: once it has obtained the structure of the query, the NE and the exact words present in the query, and eliminated the stop words, SQX-Lib selects the words that are processed in order to find their semantics to enrich the input query. In this case, we consider only those words that are common names or verbs.

2. *Processing of terms*: At this stage, SQX-Lib analyzes the selected terms on the previous step. This task consists of four steps:
   - *Obtain lemmas*: it extracts the lemmas of the terms using the Freeling NLP tool described before.
   - *Obtain synonyms*: it searches and obtains the sets of synonyms for each term, by using the extracted lemmas and the semantic resources that a set of dictionaries and a thesaurus offer.
   - *Disambiguation*: if the terms have more than one set of synonyms, SQX-Lib performs a disambiguation process to select the most appropriate ones. This process is described in Section 3.1.
   - *Select the closest synonyms*: then, from the set selected in the previous step, it selects the synonyms that are the closest to each term.
3. *Expanding the query*: finally, SQX-Lib reconstructs and expands the query through the relevant data extracted in the previous tasks while keeping the initial logical structure of the query. For example, if we have the set of keywords "accident highway new york motorbikes", after SQX-Lib has processed the keywords and selected the best terms to expand the query, it collects this information and merges it giving as result the next expanded query "((accident OR mishap) AND (highway OR motorway OR road) AND New York AND (motorbike OR motorcycle OR scooter))"

3.1 Named entity recognition method and Disambiguation process

We have performed a new knowledge-based method for NER that relies on a lexical thesaurus and on a set of resources that contain information extracted before from the text repository. Our method involves three main tasks:

1. *Collection of the NEs identified by NLP (Natural Language Processing):* First, it collects the named entities identified previously by the NLP tool.
2. *Identification of hidden NEs:* Secondly, it identifies the named entities that are hidden in the text, i.e. those NEs which have not been identified by the NLP tool because they have been written in lowercase. This happen because the NLP tool only identifies NE written in uppercase. So, for this purpose, for each word in the text, the method tests if it is possible the word is an NE by consulting an NE repository constructed before. Then, it also tests if it is a common name in EuroWordNet. In case the word could be both a named entity and a common name, the process has to disambiguate its meaning.
3. *Identification of embedded NEs:* Next, the method analyzes the named entities found to identify if some of them are embedded within another, i.e., whether multiple named entities recognized constitute a greater one. For this purpose, it identifies groups of named entities (two or more named entities that appear together without another word involved). Then, it calculates the joint and separate occurrences of each named entity of a group, and for each set of occurrences it disambiguates between them considering the other query keywords. Finally, it selects the most likely ocurrence of a named entity from each group.

Taking the previous example "accident highway new york motorbikes", "new york" appears in lowercase and Freeling do not identify it as a named entity. Our algorithm first identifies "new" and "york" as possible named entities, and then it finds that they compose a single named entity "New York".

To carry out the disambiguation of keyword meanings, we have implemented a disambiguation engine that includes a set of methods and techniques that compute the semantic relation between words. It uses a set of dictionaries with semantic information of words and a lexical thesaurus to find the possible sets of synonyms for a word. The lexical thesaurus used has been EuroWordNet [18], which is a multilingual database for several European languages, such as Spanish, French, Italian or German. This process combines the techniques of measures based on conceptual trees and measures based on the content of the information. It uses a given word as input and a set of lists of meanings of that word provided from the aforementioned lexical resources. Specifically, it calculates the degree of semantic similarity between a word and each possible list of meanings by adding up the weight obtained for each term that appears in a list of meanings. This weight is calculated by taking into account the `TF-IDF` of the term, the length of the shortest path in the EuroWordNet hierarchy between the term with the word that is being analyzed, the measure of specifity that the term has in the hierarchy, and an adaptation of the Normalized Google Distance (NGD) [19]. That is, it takes into account the relation that it has with the word and with the set of documents. Once the weight is computed for each possible list of meanings, the disambiguation engine returns the set whose weight is the smallest.

4 Case study

The present case is the search engine that is working over the news repository used everyday by all the departments of the Heraldo Group[1], a leading Spanish media. We have carried out this case study due to the difficulty we have found when comparing our proposal with other existing ones, since there are not similar approaches using the Spanish language.

The data repository used by the documentation department of the Heraldo Group is a relational database with about 10 million records. It contains news, interviews, articles, photos, infographics, videos and entire pages corresponding to publications of the group. The contents and its metadata are in Spanish plain text format. Text fields are fully indexed, so 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.

Although the system provides advanced search options, users mainly use the basic search interface, which is very similar to that of many search engines that can be found in the Internet, such as Google. This search interface is indeed very fast and practical, but of course, it also has many shortcomings regarding the quality of the results, as explained in Section 1. We find three major groups of queries using this interface: general text-based queries (61%), documentation department professional queries (8%), and queries based on thesauri and specific database fields (31%). The really expandable queries are the text-based ones. Therefore, we decided to link SQX-Lib to this type of search, which would ensure a greater coverage and quality of results for the input queries.

For experimental evaluation, we have carried out a semi-automatic process considering the set of the last queries performed on the system. First, we have discarded the queries made by the documentation department because this kind of queries are based on thesaurus tags. Likewise, we have also discarded all the queries that are using database fields different from the text content, like the date, the author, etc. So, finally, we have select a sampling of 1,000 real queries made by users to perform our analysis. We have applied our methodology and we have obtained the following results: 64% of these queries include NEs and could be optimized using SQX-Lib; 61% of these queries could take advantage of the lemmatization step; 58% of these queries include one or more common names from which the system is able to obtain synonyms and related words; and 13% of the queries could lead to build an incorrect query. We have reached this conclusion by noting that the results of certain expanded queries were incorrect and did not contain the desired information due to language idioms.

In general, we have found that 95% of these queries are expandable, that is, they can take advantage of SQX-Lib and be optimized by our system. So, we obtained the next general conclusions by observing the results obtained:

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

- The lemmatization expansion over the query leads to an average improvement of 50% in the recall.
- The semantic expansion step translates into obtaining about double results on average.
- The use of an automatic filter when we found a NE implies a reduction of an average of 20% of documentary noise.
- If we use all the options together, on average 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.

Moreover, we have carried out an opinion survey among the workers of the Heraldo Group about their use of the system improved with SQX-Lib in their daily work. They were asked about what features they would like to have in the search system by default. We have seen that 70% of the users would include the lemmatization expansion despite the rest of users prefer this feature to be optional, 58% of the users like the semantic expansion and they would also include it by default, and 88% of the users 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 it must be optional or applied by default. This is due to the different search style of each department, which is consequence of the different types of information that they want to find.

5 Conclusions and Future Work

In this paper, we have presented a semantic query expansion methodology called SQX-Lib, that combines different techniques, such as lemmatization, NER and semantics, for information extraction from a relational repository. The process includes a disambiguation engine that calculates the semantic relation between words in case it finds ambiguities and selects the best meaning for those words. We have integrated SQX-Lib in a real major Media Group in Spain and we have carried out a case study in this real environment. In our evaluation we have found that about 95% of the user keyword-based queries are optimized by our system, and obtaining up to four times more results than in a normal search. Moreover, we also have carried out an opinion survey in the company about the user experience using SQX-Lib in their daily work.

This is a prototype designed to give a solution to the particular environment of Heraldo Group, as a future work remains generalizing the solution in other languages, such as English, to enable its evaluation with standard test packages like TREC[2]. This should also facilitate the comparison of our solution with others approximations.

[2] http://trec.nist.gov/

Acknowledgment

This research work has been supported by the CICYT project TIN2010-21387-C02-02 and DGA-FSE. We thank Raquel Trillo and Jorge Gracia for their help.

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. M. G. Buey, A. L. Garrido, S. Escudero, R. Trillo, S. Ilarri, and E. Mena, "SQX-Lib: Developing a semantic query expansion system in a media group," in *European Conference on Information Retrieval*, pp. 780–784, 2014.
4. G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary problem in human-system communication," *Communications of the ACM*, vol. 30, no. 11, pp. 964–971, 1987.
5. C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *ACM Computing Surveys*, vol. 44, no. 1, 2012.
6. Y. Qiu and H.-P. Frei, "Concept based query expansion," in *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 160–169, 1993.
7. C. Buckley, G. Salton, and J. Allan, "Automatic retrieval with locality information using SMART," pp. 59–72, 1993.
8. S. Bergamaschi, F. Guerra, M. Interlandi, R. Trillo-Lado, and Y. Velegrakis, "QUEST: A keyword search system for relational data based on semantic and machine learning techniques," *Proceedings of the VLDB Endowment*, vol. 6, no. 12, pp. 1222–1225, 2013.
9. S. Sekine and E. Ranchhod, *Named Entities: Recognition, Classification and Use.* John Benjamins, 2009.
10. D. Downey, M. Broadhead, and O. Etzioni, "Locating complex named entities in web text.," in *IJCAI*, vol. 7, pp. 2733–2739, 2007.
11. M. Sanderson, "Retrieving with good sense," *Information retrieval*, vol. 2, no. 1, pp. 49–69, 2000.
12. R. Navigli, "Word sense disambiguation: A survey," *ACM Computing Surveys*, vol. 41, no. 2, p. 10, 2009.
13. F. Vasilescu, P. Langlais, and G. Lapalme, "Evaluating variants of the lesk approach for disambiguating words.," in *LREC*, 2004.
14. G. A. Miller, "WordNet: a lexical database for English," *Communications of ACM*, vol. 38, no. 11, pp. 39–41, 1995.
15. E. M. Voorhees, "Using WordNet to disambiguate word senses for text retrieval," pp. 171–180, 1993.
16. H. Schütze and J. O. Pedersen, "Information retrieval based on word senses," 1995.
17. X. Carreras, I. Chao, L. Padró, and M. Padró, "FreeLing: An open-source suite of language analyzers," in *Fourth International Conference on Language Resources and Evaluation*, pp. 239–242, European Language Resources Association, 2004.
18. P. Vossen, *EuroWordNet: A multilingual database with lexical semantic networks.* Kluwer Academic Boston, 1998.
19. 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.