A Novel Approach to Query Expansion based on Semantic Similarity Measures

Flora Amato, Aniello De Santo, Francesco Gargiulo, Vincenzo Moscato, Fabio Persia, Antonio Picariello and Giancarlo Sperli
DIETI, University of Naples "Federico II", Naples, Italy

Keywords: Semantic Search, Query Expansion, Information Retrieval.

Abstract: In this paper, we present a framework supporting information retrieval over corpora of documents using an automatic sematic query expansion approach. The main idea is to expand the set of words used as query terms exploiting the notion of *semantic similarity* between the concepts related to the search terms. We leverage existing lexical resources and *similarity metrics* computed among terms to generate - by a proper mapping into a vectorial space - an index for the fast retrieval of a set of terms "semantically correlated" to a given query term. The vector of expanded terms is then exploited in the query stage to retrieve documents that are significantly related to specific combinations of the query terms. Preliminary experimental results concerning efficiency and effectiveness of the proposed approach are reported and discussed.

1 INTRODUCTION

As is known, *semantic search* aims at improving retrieval accuracy by understanding and exploiting the "meaning" of the search terms (*keywords*) in order to generate more relevant results with respect to users' intent and needs.

The real limitation to adopting a semantic engine for searches through the Web lies in the fact that:

- the majority of content available on the Web which could be useful for indexing goals is still in an unstructured shape and spread within a large amount of web pages;
- each search term can assume several meanings depending on the considered semantic domain, thus arising additional *disambiguation* problems.

However, semantic search techniques turn out to be very suitable for retrieval tasks over corpora of semi-structured documents (e.g. clinical records, job offers, etc.) related to specific domains whose lexicon is rather restricted and where the semantic "meaning" to the different terms is usually well-defined.

For this kind of documents, some of the existing *Information Retrieval (IR)* techniques (Rinaldi, 2008; Albanese et al., 2005; Amato et al., 2008) are quite promising candidates to solve the semantic search problem: they can effectively retrieve a set of ranked objects based on their relevance respect to the query in order to satisfy the user needs related to the topic of interest. Indeed, common IR techniques used for such purposes make use of two main approaches: *query expansion* and *ontology-based*.

Systems leveraging query expansion techniques semantically "enrich" the user query adding words that have *semantic relationships* (e.g. synonyms, hypernyms, hyponyms, etc.) with the search terms using as support domain vocabularies or taxonomies or ad-hoc lexical resources, for improving the effectiveness of the information retrieval process (Carpineto and Romano, 2012). Query expansion techniques are quite simple to realize and their performances depend on the "goodness" of expansion phase (Rivas et al., 2014).

On the other hand, ontology-based techniques (Amato et al., 2009; Fernández et al., 2011; Jain and Singh, 2013) exploit ad-hoc ontologies (manually or automatically built) representing the main concepts and relationships for documents' corpora in combination with standard domain ontologies according to the *linked data* paradigm. Then, queries are properly computed mapping the user's request into the related ontology concepts and leveraging the defined data and object properties (Vallet et al., 2005; Castells et al., 2007).

The main challenge when applying ontology-based techniques is surely represented by the initial phase of the ontology building: in fact, the ontology

Amato F., De Santo A., Gargiulo F., Moscato V., Persia F., Picariello A. and Sperlì G..
A Novel Approach to Query Expansion based on Semantic Similarity Measures.
DOI: 10.5220/0005579703440353
In *Proceedings of 4th International Conference on Data Management Technologies and Applications* (KomIS-2015), pages 344-353
ISBN: 978-989-758-103-8

has to be able to effectively capture and describe the entire knowledge related to the corpus of documents. Manual approaches can generate very effective ontologies but are usually time consuming and involve complex activities. In turn, automatic approaches are more efficient but can generate more approximate results that can influence the search effectiveness. Thus, it is our belief that hybrid strategies should be preferred.

In this paper, we present a novel semantic approach to query expansion,with the aim of supporting information retrieval over corpora of documents. In particular, the main idea behind our work is to expand the set of words used as query terms exploiting the notion of *semantic similarity* between the concepts related to search terms and those associated to the available domain vocabularies or taxonomies.

More in details, in the indexing stage we leverage one or more vocabularies or domain taxonomies and a *similarity metric* among terms (e.g. *Wu & Palmer* if we have a taxonomy) to generate by a proper mapping into a *vectorial space* an index for the fast retrieval of the set of terms "semantically correlated" to a given query term. In our implementation, we use the *k-d Tree* and its *k-nearest neighbor* or a *range* search capabilities. The vector of expanded terms is then exploited in the query stage to retrieve documents that contain several combinations of possible query terms. Finally, documents are opportunely ranked with respect to initial user query.

The paper is organized as follows. Section 2 shows a brief motivating example for our work. Section 3 describes the of some recent approaches for automatic query expansion problem. Section 4 illustrates the document semantic retrieval framework with several implementation details. Section 5 reports some preliminary experimental results, while Section 6 finally discusses some conclusions and the future work.

2 MOTIVATING EXAMPLE

The diffusion of Web-centric services for the job market is rapidly expanding all over the world, and enabling a significant part of the job demand to be routed through Web portals, services, and applications. As a consequence, the Web is increasingly used by both employers and job seekers to advertise demand and supply, and it is becoming increasingly important to find new ways of enhancing the recruitment-related activities.

In particular, there is the need for job seekers to retrieve job offers most representative of their curriculum and skills as fast as possible, especially considering the increasing competition due to the ongoing financial crisis. Moreover, to be effective the job research must also be as *complete* as possible, in order to help people make the best choice for them according to their backgrounds, skills and ways of life.

Clearly, simple *syntactic queries* are highly inadequate to this goal: for instance, if a job seeker is a computer science engineer and he should type just '*Computer Science Engineer*' while performing his search, he would definitely miss offers which do not esplicitly contain these words either in their titles or in their description.

Our work is grounded in the belief that this issue could be better addressed by exploiting *semantic search* approaches instead. In fact, a *query expansion* approach based on semantic techniques would allow the seeker to also retrieve the offers related to words like *Java Programmer, Engineer, Linux Systems* rather than just *Computer Science Engineer*.

Thus, the main aim of this paper is to design and develop a semantic framework for performing query expansion tasks on large corpora of documents. Specifically, we will focus the experimental phase on a set of job related announcements in Italy.

3 STATE OF THE ART

Search Engines are essential tools for most computer users in a wide variety of contexts. As a result, information retrieval has become an important field of research over the last 30 years (Amato et al., 2014). IR (Sagayam et al., 2012) is the process related to the organization and retrieval of information from a large number of documental corpus.To this purpose, many document indexing and retrieval systems have been proposed which have been shown to be generally effective. However, a deeper analysis reveals that even though these techniques improve average performance, there is often wide variation in the impact of each chosen technique on the retrieval effectiveness for specific queries. Particularly, the problem known as the *vocabulary problem* - consisting in the indexers and the users not using the same word set (Furnas et al., 1987) - is the most critical issue for retrieval effectiveness. Furthermore, synonymy (different words with the same or similar meanings, such as *tv* and *television*) together with word inflections (such as with plural forms,*television* versus *televisions*), may result in a failure to retrieve relevant documents, with a decrease in recall (the ability of the system to retrieve all relevant documents) while polysemy (same word with different meanings, such as *java*) may cause re-

trieval of erroneous or irrelevant documents, thus implying a decrease in precision (the ability of the system to retrieve only relevant documents) (Moscato et al., 2010b; Moscato et al., 2010a).

To cope with the *vocabulary problem* , different approaches have been proposed: among these we can list interactive query refinement, relevance feedback, word sense disambiguation and search results clustering. One of the most common techniques is to expand the original - typically short - query through the addition of related keywords to it. These additional keywords (or *expansion terms*) generally increase the likelihood of a match between the query and relevant documents during retrieval, thereby improving user satisfaction.

Automatic Query Expansion (AQE) has a long history in IR, as has been suggested since 1960 by Maron and Kuhns (Maron and Kuhns, 1960).

Following (Carpineto and Romano, 2012) AQE techniques can be classified into five main groups according to the conceptual paradigm used for finding the expansion features: *linguistic methods*, *corpus-specific statistical approaches*, *query-specific statistical approaches*, *search log analysis* and *Web data*. Linguistic methods leverage global language properties such as morphological, lexical, syntactic and semantic word relationships to expand or reformulate query terms. They are typically based on dictionaries, thesauri, or other similar knowledge representation sources such as WordNet. These techniques are usually generated independently of the full query and of the content of the database being searched, and they are usually more sensitive to word sense ambiguity.

In (Buey et al., 2014), Buey *et al.* present a semantic query expansion methodology called *SQX-Lib*, combining techniques such as lemmatization, NER and semantics for information extraction from a relational repository. The proposal includes a disambiguation engine that calculates the semantic relations between words and - in case any ambiguity is found - selects the best meaning from those available for the specific word. SQX-Lib has been integrated by a major Media Group in Spain.

Dalton et al. (Dalton et al., 2014) propose an AQE method based on annotations of entities from large general purpose knowledge bases, such as *Freebase* and the *Google Knowledge Graph*. They proposed a new technique, called *Entity Query Feature Expansion* (EQFE) which enriches the query with features from entities and their links to knowledge bases, including structured attributes and text. One limitation of this work is that it depends upon the success and accuracy of the entity annotations and linking.

Bouchoucha et al. (Bouchoucha et al., 2014) present a unified framework to integrate multiple resources for a *Diversified Query Expansion*. By implementing two functions, one to generate candidate expansion terms and the other to compute the similarity between two terms, any resource can be plugged into the framework. Experimental results show that combining several complementary resources performs better than using one single resource.

In (Pal et al., 2014), a new way of using WordNet for query expansion is proposed with a combination of three AQE methods that take into account different aspects of a candidate expansion term's usefulness. For each candidate expansion term, this method considers its distribution, its statistical association with query terms, and also its semantic relation with the query.

Corpus-specific statistical approaches analyze the contents of a full database to identify features used in similar ways. Most early statistical approaches to AQE were corpus-specific and generated correlations between pairs of terms by exploiting term co-occurrence, either at the document level, or to better handle topic drift, in more restricted contexts such as paragraphs, sentences, or small neighborhoods.

In (Huang et al., 2013), the authors propose a new method - called *AdapCOT* - applyingadaptive co-training to select feedback documents for boosting effectiveness. Co-training is an effective technique for classification over limited training data, which is particularly suitable for selecting feedback documents. The proposed AdapCOT method makes use of a small set of training documents, and labels the feedback documents according to their quality through an iterative process.

Colace et al. (Colace et al., 2015) use a minimal relevance feedback to expand the initial query with a structured representation composed of weighted pairs of words. Such a structure is obtained from the relevance feedback through a method for pairs of words selection based on the Probabilistic Topic Model. The proposed approach computes the expanded queries considering only endogenous knowledge.

Query-specific techniques take advantage of the local context provided by the query. They can be more effective than corpus-specific techniques because the latter might be based on features that are frequent in the collection but irrelevant for the query at hand.

Ermakova et al. (Ermakova et al., 2014) propose a AQE method that estimates the importance of expansion candidate terms by the strength of their relation to the query terms. The method combines local analysis and global analysis of texts.

Yang *et al.* (Yang et al., 2014) suggest a method that applies a linguistic filter and a C-value method to

Figure 1: Workflow.

extend the query terms, and then uses the Normalized Google Distance-based method to calculate the term weight and choose Top-N terms as extended query. The authors claim that the *Normalized Google Distance* (NGD) with some global factors enhance the relevance between initial query and extended query, and improve the accuracy of the search results of the expert finding system.

Search log analysis paradigm is based on analysis of search logs. The idea is to mine query associations that have been implicitly suggested by Web users.

Finally, Web data techniques rely on the presence of anchor texts. Anchor texts and real user search queries are very similar because most anchor texts are succinct descriptions of the destination page.

4 DOCUMENT RETRIEVAL FRAMEWORK

In this section, we first provide a brief overview of the AQE process workflow which, as shown in Figure 1, can be divided into the following three steps: *Keyword Extraction & Document Tagging*, *Semantic Query Expansion* and *Document Retrieval*.

The basic idea behind our proposal is that the results of a user's query on a document corpus from a specific domain could be significantly improved by expanding the set of words used to search the corpus of documents at hand through a process exploiting semantic proximity between concepts.

Thus, the input of this process is a *Document Corpus* - a set of documents related to a relevant topic- which will be firstly processed through a *Keyword extraction & Document Tagging* step.

In fact, we need not only to identify the concepts (also, keywords) that can be relevant to each document, but we also to limit the set of words that users could use to query the corpus, thus identifying all possible words that need to be semantically linked together.

Therefore, in a first step each document belonging to the corpus is parsed by proper NLP algorithms in order to identify the salient words in the text and link them to the document itself enriching its content through a *tagging* process.

At this point, the *Semantic Query Expansion* step will allow us to semantically enrich the user query in order to improve the performance of the document retrieval process by enlarging its semantic scope. In particular, there are three main phases of the semantic expansion process:

1. In the first phase, semantic distances between the keywords recognized in the first step are computed, based on different semantic similarity measures (Resnik, 1999) - such as *Wu and Palmer*, *Resnik* or *Leacock and Chodorow*. A formalized description of the knowledge of the domain - such as ad hoc ontologies, taxonomies or vocabularies - is needed for these measures to work effectively;
2. then, the keywords and their related distances - computed as output of previous bloc - are embedded in a data structure accurately chosen to efficiently support information retrieval tasks on large sets of data: for instance, m-tree, kd-tree, or different kinds of NO-SQL databases, such graph-databases, are particularly appropriate for these kind of applications;
3. The last phase is the actual query expansion step, which allows to enrich the scope of the user query through appropriate operations performed on the data structure.

Finally, the *Documental Retrieval* phase returns a set of documents (*Query Results*) significantly related to the user query by exploiting the semantic expansion operated through the data structure and the tagging of the documents resulting from the NLP stage.

4.1 System Architecture

The proposed system has been developed according to the architecture shown in Figure 2. In particular we can distinguish three main sections in the framework structure.

The first section is represented by an *NLP Module* devoted to extracting significant keywords from the set of documents we are interested in querying and, subsequently, to tagging each document with the related relevant keywords.

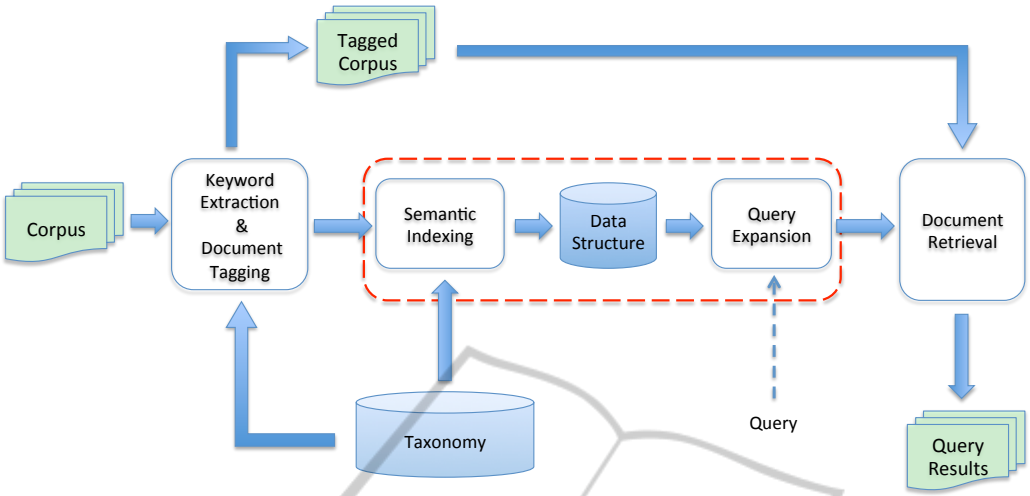

Figure 2: System Architecture.

A second section is represented by the *Semantic Expansion Module*, which is undoubtedly the central feature of our proposal. It is composed by a subcomponent devoted to building (in the indexing stage) a semantic indexing of the extracted keywords exploiting a chosen taxonomy or vocabulary, and of a subcomponent devoted to the actual query expansion task (in the querying stage).

Synthetically, this macro-module takes as its inputs the set of keywords extracted from the corpus of documents and the query the user wants to perform on documents' set, while its output is a set of keywords semantically linked to the original query.

Finally, a *Document Retrieval Module* takes as input the expanded query and the tagged corpus and returns the documents matching the set of expanded keywords.

In what follows, we will describe each of these module in more detail.

4.1.1 NLP Module

The first thing we need for our framework to work efficiently is to extract a set of significant keywords from the documents we are interested in retrieving. These keywords are those which can be chosen by users to express their queries on the document set.

Therefore, we want the *NLP module* to implement an extraction technique able to retrieve significant keywords from the documents, while keeping the set of possible keyword restricted to those terms users will adopt while querying the corpus.

In this work, we decided to implement a methodology previously proposed by some of the authors in the context of ontology population for e-health applications (Amato et al., 2015b). In particular, the work was aimed at using techniques from natural language processing for populating and enriching an ontology descriptive of the domain of Medical records in Campania region starting from a set of Medical records and from a simple Ontology for the medical record domains built with the help of a domain experts.

Essentially, the authors start from the set of medical records available to them in free text, and obtain structured text in the form of RDF triples through lexicon-grammar based NLP techniques exploiting the NooJ environment [1] for domain specific information retrieval. The local grammars needed for the information retrieval task can be built with the help of both linguists and domain experts.

This approach is particularly interesting to us, since it allows the retrieval of keywords through the definition of lexicon-grammars built upon a taxonomy specifically chosen for the domain of interest. For each document, this methodology will produce a set of relevant keywords descriptive of the document and limited to the vocabulary of the taxonomy. Once a list of keywords is extracted from the document, it is also added to the document description through a tagging procedure. The tagging of the documents will later ease the querying of the corpus and the retrieval task.

4.1.2 Semantic Expansion Module

Let us recall that the purpose of our framework is to implement an effective way for semantically expanding a user query on a set of documents, in order to improve the retrieval of significant elements in large, domain specific corpora.

Thus, the set of keywords produced by the *NLP*

[1] http://www.nooj4nlp.net/

module and the taxonomy/vocabulary used to guide the NLP task are used as input to a *Semantic Expansion Module*. This module implements the central features of our framework: namely, the building of a data structure indexing in a vectorial space the extracted keywords by means of semantic distances and the expansion of a user generated query through the retrieval of a set of semantically related keywords.

First of all, we are interested in recognizing the semantic relationships between each every word in the keyword list and in using them to compute *distances* between those words.

Since by construction each term in the keyword set is matched by a concept in the taxonomy/vocabulary used in the NLP stage, we can use the available taxonomy to compute semantic distances between the keywords.

There are several metrics that can be exploited to compute a semantic distance between two concepts in a taxonomy, every one of them having its advantages and disadvantages (Resnik, 2011). Although the main features of the proposed framework are not dependent on the metric chosen to compute the distances, the results of the document retrieval phase could indeed be affected by the goodness of the metrics, often dependent on the available related taxonomy/vocabulary. While the choice of the metric to adopt may vary from domain to domain, in this first implementation we decided to perform our experiments exploiting a distance based on the *Wu-Palmer* similarity measure .

At this point, having applied a semantic distance to the word list, we have put the whole set of words into a metric space. There are several possible data structures which will allow us to exploit this space for targeted information retrieval. However, to fully exploit the potential of querying a list of words semantically linked to each other, we chose to implement a semantic index.

In particular, in this paper we propose to use *SemTree*, a semantic index for supporting retrieval of information from huge document collections, proposed by some of the authors in (Amato et al., 2015a). Essentially, SemTree is a distributed version of Kd-Trees for information retrieval, leveraging the mapping of words in a vector space by means of a proper semantic distance between concepts in a taxonomy.

In a first stage, our experiments will be run using a sequential implementation of a standard kd-tree. However, using the distributed implementation of SemTree, we will also present several experiments on a distributed cluster showing the efficiency of the distributed approach in managing big amounts of data.

Note that, to effectively implement this index, we first need to map the set of words from a metric space to a vector space. This can be done through the use of several existing algorithms, such as *FastMap* or *MDS* (Faloutsos and Lin, 1995).

In fact, both these methods are able to map objects into points in some n-dimensional space (where n is user-defined, and in this application will be set to 3), such that the dis-similarities between objects are preserved.

Experiments on real and synthetic data have shown that FastMap is faster than MDS (being linear, as opposed to quadratic, on the database size N), while the latter is more accurate in preserving distances between objects and the overall structure of the data-set. Thus, feeding the list of word and an accurately chosen semantic metric to one of these algorithm will produce a set of points in an n-dimensional space, each point representing one of the words from the original list.

4.1.3 Document Retrieval Module

Leveraging a chosen semantic measure and the kd-tree based index, we are now able to produce - through a simple search on the tree - a set of words which are semantically similar to the one the user choose to query.

Furthermore, both the kd-tree's k-nearest and *range* query produce results ordered according to their relevance to the queried term. In particular, a k-nearest query on the tree would produce the set of k terms in the index which are closest to the query's original word. The list of words returned by an interrogation of the tree is ranked according to the proximity of each word to the initial term.

This ranking can be exploited in the document retrieval phase, where the rank of each word in the list is used to assign a *significancy score*, representing the related documents' probability of being relevant with respect to the expanded query. The actual query on the corpus of interest is performed at this point.

Although the problem of ranking the results of an expanded query is complex and - as shown in Section 3 -currently widely discussed in the existing literature, in this first evaluation we decided to perform a weighted *AND* combining-operation of the results, exploiting each word's *semantic proximity*.

Let us consider a list of expanded query terms $t_1, \ldots, t_n$, numbered according to their relevance to the original query term. Then, the expanded query is performed as:

$$(t_1 \text{ AND } \ldots \text{ AND } t_m) \text{ AND } (t_{m+1} \text{ OR } \ldots \text{ OR } t_n)$$

where $t_1, \ldots, t_m$ are the terms that have a semantic proximity respect to the query term greater than a

given threshold.

Significant documents are retrieved from the tagged corpus through a matching phase with the keyword list, each documents having been assigned a list of *tags* representing the words more descriptive of the document contents.

Obviously, the output to the user query is a ranked set of documents, presented with significancy scores.

5 PRELIMINARY EXPERIMENTAL RESULTS

In order to evaluate the goodness of our proposal, we hereby present a preliminary experimental evaluation of our protocol's retrieval effectiveness, using standard evaluation metrics, such as precision and recall, and efficiency.

5.1 Experimental Setup for Job Announcements Retrieval

First of all, in order to estimate our framework effectiveness in retrieving relevant documents starting from a single-term queries, we considered a dataset of about 1000 documents concerning job announcements in Italy, crawled from commonly used job search engines.

In order to efficiently conduct the NLP phase, we also exploited the Italian version of the *ESCO taxonomy*[2]. Thus, our first step was to process the corpus of untagged documents as described in the previous sections; then we performed 100 k-nearest queries - each of them starting from the input of a single job related term chosen from those listed in the ESCO taxonomy.

Furthermore, we decided to evaluate the framework efficiency focusing on the index's retrieval times - both for its sequential implementation and for its distributed implementation, as presented in (Amato et al., 2015a).

This choice of focusing on the index searching times is supported by the fact that the NLP processing phase and the index building have to be performed just once for each considered document corpus, thus the delays associated to these phases can be considered non-significant for an initial evaluation of the framework.

5.2 Effectiveness

The effectiveness has been computed using the *Precision/Recall* metrics, by comparing the output of our algorithms against a ground truth provided by human annotators.

Here, we present a preliminary evaluation of our approach effectiveness with respect to the described case study. In particular, we want to show feasibility of the proposed framework in automatically finding job announcements significantly related to the *job term* a user chose to query.

Specifically, we performed 100 different k-nearest queries using as query points a set of concepts extracted from the ESCO taxonomy (i.e. *Ingegnere Elettronico, Farmacista, Softwarista, Giardiniere*).

In order to evaluate the output of our framework we asked a group of human testers to perform a similar retrieval task by hand, namely searching the tagged document corpus for job announcements semantically related to the job they were asked to retrieve, each job corresponding to one of the query we performed automatically through our framework.

Denoting with T the set of documents returned by the k-nearest query related to a given target job (i.e. the user searching for *ingegnere*) and with T^* the set of documents provided by the human ground truth when searching for the same job, Precision (P) and Recall (R) were computed using:

$$P = \frac{|T \cap T^*|}{|T|}$$

$$R = \frac{|T \cap T^*|}{|T^*|}$$

Figure 3 shows the average Precision and Recall values for the 100 k-nearest query cases we performed on the the tagged document corpus of job announcements, when varying k.

As expected, the lower is k, the higher is P and the lower is R; *vice-versa*, when the value of k is high, the value of R grows up while the value of P decreases.

5.3 Efficiency and Considerations on Scalability Issues

In order to get an accurate evaluation on the efficiency of our index for a big data scenario, we used SCOPE [3], a General Purpose Supercomputing Infrastructure.

SCOPE is an infrastructure provided by the Information System Center of University of Naples, composed of about 300 computing nodes (64 bit Intel Xeon with 8 GB RAM).

In particular, we used at most 65 processors for evaluating the performances of our index holding a maximum number of *8,000,000* points.

[2] https://ec.europa.eu/esco/home

[3] www.scope.unina.it

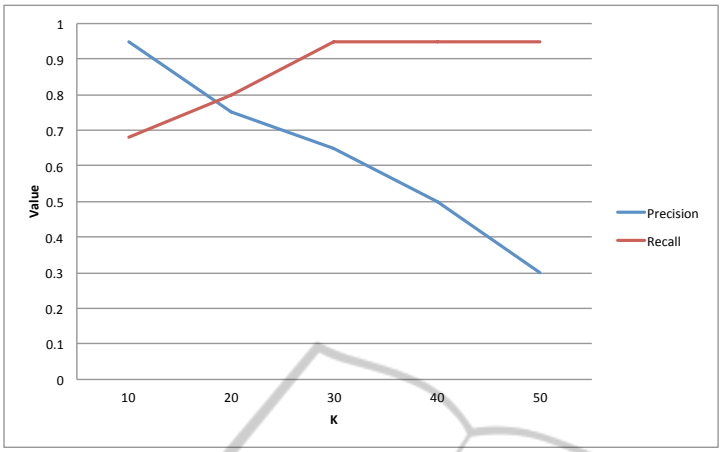

Figure 3: Average Effectiveness Measures on 100 Queries

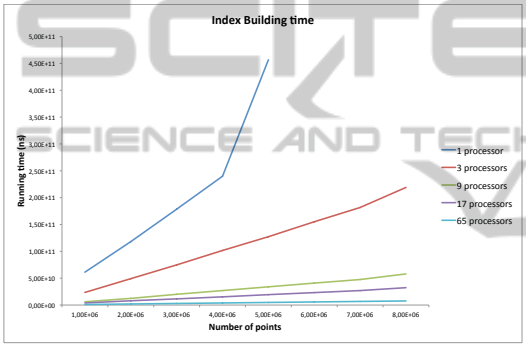

Figure 4: Index Building Time in SCOPE

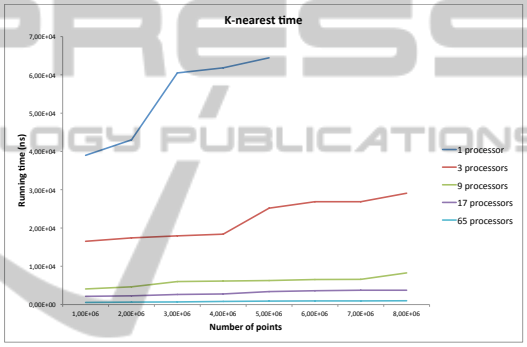

Figure 5: K-Nearest Time in SCOPE

More in details, Figure 4 shows the running time achieved for building the index when varying the size of the input data and the number of used processors of the supercomputing infrastructure.

Figure 5 reports the running time of the *distributed K-nearest algorithm* when varying the number of processors.

Similarly, Figure 6 shows the running time of our algorithm for *distributed range query*. All the listed case studies confirm the positive trend achieved during the preliminary experimentation (Amato et al., 2015a).

As expected, the most costly procedure is the *index building*; however, the achieved running times are very good and become extremely low whether we decide to use 65 processors, even if our index holds *8,000,000* points.

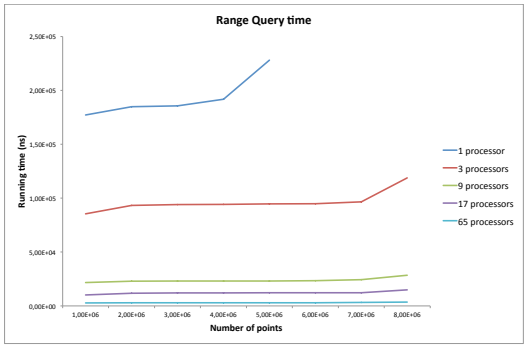

Figure 6: Range Query Time in SCOPE

6 CONCLUSIONS

In this paper, we proposed a document retrieval framework based on semantic query expansion techniques, developed exploiting state-of-the-art NLP techniques and a combination of well known semantic metrics within a kd-tree indexing structure.

The idea behind this work is that expanding a user query on a corpus of documents with a set of semantically related words (within the documents theirselves) would increase both the number of documents returned by the query, and the relevance of documents to the original query.

Thus, we firstly presented a general architecture

for our framework and then proposed a preliminary experimental evaluation both in terms of efficiency and effectiveness, considering as case study a corpus of job announcements in Italy.

The experimental evaluation showed that our approach offers good results in terms of retrieval precision and recall. Furthermore, we evaluated the execution time of the core data structure - the kd-tree- both in its sequential version and in a distributed one, showing that the proposed approach is easily scalable.

REFERENCES

Albanese, M., Capasso, P., Picariello, A., and Rinaldi, A. M. (2005). Information retrieval from the web: an interactive paradigm. *Advances in Multimedia Information Systems*, pages 17–32.

Amato, F., De Santo, A., Gargiulo, F., Moscato, V., Persia, F., Picariello, A., and Poccia, S. (2015a). Semindex: an index for supporting semantic retrieval of documents. In *Proceedings of the IEEE DESWeb ICDE 2015*.

Amato, F., De Santo, A., Moscato, V., Picariello, A., Serpico, D., and Sperli, G. (2015b). A lexicon-grammar based methodology for ontology population in e-health applications. In *The 9-th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS-2015). Blumenau, Brazil.*

Amato, F., Mazzeo, A., Moscato, V., and Picariello, A. (2009). A system for semantic retrieval and long-term preservation of multimedia documents in the e-government domain. *International Journal of Web and Grid Services*, 5(4):323–338.

Amato, F., Mazzeo, A., Moscato, V., and Picariello, A. (2014). Exploiting cloud technologies and context information for recommending touristic paths. In *Intelligent Distributed Computing VII*, pages 281–287. Springer.

Amato, F., Mazzeo, A., Penta, A., and Picariello, A. (2008). Knowledge representation and management for e-government documents. *IFIP International Federation for Information Processing*, 280:31–40.

Bouchoucha, A., Liu, X., and Nie, J.-Y. (2014). Integrating multiple resources for diversified query expansion. In *Advances in Information Retrieval*, pages 437–442. Springer.

Buey, M. G., Garrido, Á. L., and Ilarri, S. (2014). An approach for automatic query expansion based on nlp and semantics. In *Database and Expert Systems Applications*, pages 349–356. Springer.

Carpineto, C. and Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, 44(1):1.

Castells, P., Fernandez, M., and Vallet, D. (2007). An adaptation of the vector-space model for ontology-based information retrieval. *Knowledge and Data Engineering, IEEE Transactions on*, 19(2):261–272.

Colace, F., De Santo, M., Greco, L., and Napoletano, P. (2015). Weighted word pairs for query expansion. *Information Processing & Management*, 51(1):179–193.

Dalton, J., Dietz, L., and Allan, J. (2014). Entity query feature expansion using knowledge base links. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 365–374. ACM.

Ermakova, L., Mothe, J., and Ovchinnikova, I. (2014). Query expansion in information retrieval: What can we learn from a deep analysis of queries? In *International Conference on Computational Linguistics-Dialogue 2014*, volume 20, pages pp–162.

Faloutsos, C. and Lin, K.-I. (1995). *FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets*, volume 24. ACM.

Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., and Motta, E. (2011). Semantically enhanced information retrieval: an ontology-based approach. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):434–452.

Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971.

Huang, J. X., Miao, J., and He, B. (2013). High performance query expansion using adaptive co-training. *Information Processing & Management*, 49(2):441–453.

Jain, V. and Singh, M. (2013). Ontology based information retrieval in semantic web: A survey. *International Journal of Information Technology and Computer Science (IJITCS)*, 5(10):62.

Maron, M. E. and Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)*, 7(3):216–244.

Moscato, V., Picariello, A., and Rinaldi, A. M. (2010a). A combined relevance feedback approach for user recommendation in e-commerce applications. In *Advances in Computer-Human Interactions, 2010. ACHI'10. Third International Conference on*, pages 209–214. IEEE.

Moscato, V., Picariello, A., and Rinaldi, A. M. (2010b). A recommendation strategy based on user behavior in digital ecosystems. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, pages 25–32. ACM.

Pal, D., Mitra, M., and Datta, K. (2014). Improving query expansion using wordnet. *Journal of the Association for Information Science and Technology*, 65(12):2469–2478.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language.

Resnik, P. (2011). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *CoRR*, abs/1105.5444.

Rinaldi, A. M. (2008). A content-based approach for document representation and retrieval. In *Proceedings of the eighth ACM symposium on Document engineering*, pages 106–109. ACM.

Rivas, A. R., Iglesias, E. L., and Borrajo, L. (2014). Study of query expansion techniques and their application in the biomedical information retrieval. *The Scientific World Journal*, 2014.

Sagayam, R., Srinivasan, S., and Roshni, S. (2012). A survey of text mining: Retrieval, extraction and indexing techniques. *International Journal Of Computational Engineering Research*, 2(5).

Vallet, D., Fernández, M., and Castells, P. (2005). An ontology-based information retrieval model. In *The Semantic Web: Research and Applications*, pages 455–470. Springer.

Yang, K.-H., Lin, Y.-L., and Chuang, C.-T. (2014). Using google distance for query expansion in expert finding. In *Digital Information Management (ICDIM), 2014 Ninth International Conference on*, pages 104–109. IEEE.