A Simple Solution for Improving the Effectiveness of Traditional Information Retrieval Systems

GIOVANNI PILATO
ICAR - Istituto di CAlcolo e Reti ad alte prestazioni
Italian National Research Council
Viale delle Scienze - 90128 Palermo - Italy

ITALY

GIORGIO VASSALLO, MARIA VASILE, AGNESE AUGELLO, SALVATORE GAGLIO
DINFO - Dipartimento di ingegneria INFOrmatica
University of Palermo
Viale delle Scienze - 90128 Palermo - Italy

ITALY

Abstract: - In this paper we present a system based on the LSA paradigm to improve the performance of a traditional information retrieval system. The proposed system aims to improve both the recall and the precision capabilities of traditional search engines thanks to a semantic query expansion and a subsequent semantic results filtering. A collection of 650 documents has been used to compare the performances of the proposed system with a traditional search engine. Experimental trials show the effectiveness of the proposed solution.

Key-Words: - LSA, Query Expansion, Information Retrieval

1 Introduction

The information retrieval deals with automatically retrieving only those documents that satisfy the need of information of the user minimizing the quantity of irrelevant information[10][13]. Search engines are the main example of IR system. The research is based on a lexical matching: a traditional search engine returns to a user the documents containing the query terms or a their logical combination. Obviously in this context the way in which the need of information is expressed is very important: queries badly expressed can determinate a wrong information retrieval or restrict the amount of relevant information obtained.

Effectiveness is a measure for the quality of information retrieval tools. The lexical approach causes a decrease of the system effectiveness, measured with the Precision, that is the ratio of the number of relevant documents retrieved by the system to the total number of documents retrieved and the Recall, that is the ratio of the number of relevant documents retrieved for a query to the number of documents relevant to that query in the entire document collection[1][13].

A traditional search engine does not take care of words properties: the same word form can have many meanings (polysemy) and many different terms can have the same meaning (synonymy).

In recent years it has been developed a paradigm, called Latent Semantic Analysis (LSA) which is a theory and method for extracting and representing the contextual usage meaning of words by statistical computations applied to a large corpus of text[7]. This methodology allows to deal with the information retrieval problems, using a vector model where documents and queries are coded with vectors that constitute their sub-symbolic representation[12]. Hence is possible to consider these entities independently from their morphological and lexical representation[2].

In this paper we present a system based on the LSA paradigm to improve the performance of a traditional information retrieval system.

A semantic space is automatically generated using a given corpus of texts, and then a sub-symbolic dictionary of words is created. The building of this sub-symbolic and semantic dictionary allows the realization of two modules that interact with a common search engine.

The system aims to improve both the recall and the precision capabilities of classic search engines thanks to a semantic query expansion and a subsequent semantic results filtering.

The first sub-system processes the user query, adding to the query terms other semantically related words. A semantic query expansion is therefore

implemented and, as a consequence, the Recall of the whole system is increased.
The other module processes the documents returned by the search engine in order to semantically filter them, thus selecting only those documents semantically relevant to the query. This strategy increases the Precision of the whole system.
To test the effectiveness of the proposed solution experimental trials have been conducted using the well-known traditional search engine jakarta lucene[14]. A collection of 650 documents given by question-answer pairs extracted by the Internet newsgroups FAQs[15] have been indexed using this search engine, and then the system modules have been used to test the performance improvement of the whole system.
The remainder of the paper is organized as follows: in the next section related works are shown; in section 3 the proposed solution is presented; in section 4 experimental results are illustrated and in section 5 conclusion are given.

2 Related works

Manual query expansion has been revised by a lot of researchers, however the choice of the new terms to be added to the initial query depends on the particular technique of expansion implemented: there are techniques of expansion based on the local relevance feedback and techniques based on the expansion with thesauri[1]. In the first case the idea is that the user's query in a given instant of time, is in some way influenced by the queries formulated in the preceding instants of time. The method therefore expands the query with information deduced by the preceding query and searches[11]. For the use of thesauri there exist two principal typologies[1]: Query Expansion based on a Similarity Thesaurus and Query Expansion based on a Statistical Thesaurus. In both cases the terms that expand the user query are drawn out from a database that memorizes, according to opportune criteria, the relationships of similarity between words[4][6][9].
For information filtering it is intended the procedure of selection to which an informative flow is submitted to prefer only those information that correspond to certain profiles that express the interest of the user[3]. There are many applications of LSA to information filtering problems. Foltz[7] has compared the method of information filtering based on the latent semantic analysis with a method based on the search for word-key, using as whole of texts the netnews. LSA has determined an improvement of the performances of information filtering system of 23% in comparison to the traditional technique.

3 The Proposed solution

In this work a system, based on the automatic creation of a semantic space, according to the LSA paradigm, is presented. The main feature of the proposed system is to improve the performance of a traditional information retrieval tools, like search engines[7].
An off-line procedure is required to create a semantic vector space using the LSA technique[2], where semantically similar words are represented by near vectors according to a properly defined metric. The LSA technique is therefore applied to a collection of documents, a semantic space is created and a vector is associated to each word of the document collection. This set of vectors associated to words will constitute a sub-symbolic dictionary. A database is then built in which both the sub-symbolic coding of all the words contained in the document collection and the semantic relationships between all the possible couples of words is stored. This database can be used for supporting the operations of a user query expansion and subsequent semantic information filtering of the results given by a traditional search engine.
The whole system and its interaction with the user and with a traditional search engine are illustrated in Fig.1.

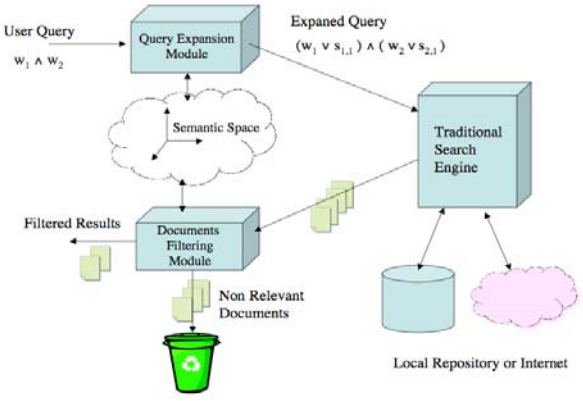

Fig. 1: The proposed solution schema

The system is composed of two modules: the first module, called "Query Expansion Module" realizes a query expansion to increase the retrieval of the "candidate" relevant documents from a repository (which could be also Internet). The expansion module processes the user query, adding to the query terms, given by the user, other semantically related words present in the semantic space created

by the LSA. The expanded query is therefore given to a traditional search engine that, as a consequence, will retrieve more documents related to the semantics of the original query. This approach aims to improve the Recall measure.

The second module, called "*Documents Filtering Module",* processes the documents returned by a traditional search engine. The semantic expansion of the query, increasing the recall measure, lowers the precision of the results given by the traditional search engine. The goal is to semantically filter the retrieved information, selecting only the semantically relevant documents for the query and rejecting the documents considered not important for the user. This approach aims to increase the Precision of the whole system.

3.1 Semantic Space and Sub-symbolic Dictionary Creation

3.1.1 The LSA-Based Semantic Space Creation

The LSA has been used to obtain a sub-symbolic coding [7][12] of the terms from which to deduce the semantic relationships between couples of words.

Starting from a collection of documents it has been defined a term-document matrix $\mathbf{A}_{(mxn)}$, where the generic $[a_{ij}]$ element is the number of occurrences of the i-th word in the j-th document.

The matrix **A** is factorized according to the technique of the singular values decomposition (SVD)[2] in the product of three unique matrices **U**, **Σ**, **V**:

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T \tag{1}$$

The **U** matrix sub-symbolically describes the entities corresponding to the rows of the matrix **A; V** sub-symbolically describes the entities corresponding to the columns of **A**[7]. The truncated SVD is establishes that :

$$\mathbf{A}_k = \sum_{=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \tag{2}$$

$\mathbf{A_k}$ is obtained choosing a number reduced of factors *k*, corresponding to the more significant singular values. $\mathbf{A_k}$ it is the best approximation of **A** according to the criterion of the least squares, and it contains the most meaningful information of **A**.

This entails the selection of the subspace spanned by the most important semantic dimensions. In fact, only a portion of the rows vectors rows of matrix **U** and of the column vectors of **V** contains the more meaningful informative content in order to reconstruct the initial matrix. As a consequence, the effect of the truncated SVD and the resulting dimensionality reduction is the elimination of the noise, due to the local variations in the use of the words.

3.1.2 Sub-symbolic encoding of words

We have defined a generic k-dimensional vector $\mathbf{w_i}$ as a vector coding of the generic i-th word w_i of the dictionary as constituted by the first *k* rows components of the **U** matrix multiplied by the square roots of the corresponding singular values.

$$u_i = \left\{ \underbrace{u_{i,1}, u_{i,2}, \ldots, u_{i,k}}_{\text{most significant components}}, u_{i,k+1} \ldots, u_{i,r} \right\} \tag{3}$$

$$\text{generic word} \Rightarrow w_i = \left\{ u_{i,1}\sqrt{\sigma_1}, u_{i,2}\sqrt{\sigma_2}, \ldots, u_{i,k}\sqrt{\sigma_k} \right\} \tag{4}$$

We have considered the square root of the corresponding singular values, to equally distribute the informative contribution of such factors among the row vectors of **U** and the column vectors of **V**.

The choice of the number *k* of singular values is experimentally determined.

3.1.3 Sub-symbolic coding of documents

A document is constituted by a set of words; therefore it is possible to get the corresponding vector to a specific document as the sum of the vectors that codify the single words. Therefore, if a document *d* is constituted by *n* words, $d=\{w_1, w_2, \ldots, w_n\}$, the vector **d** corresponding to the document will be obtained considering the vector sum brought following:

$$\mathbf{d} = \mathbf{w}_1 + \mathbf{w}_2 + \ldots + \mathbf{w_n} \tag{5}$$

with $\mathbf{w_1}, \mathbf{w_2}, \cdots, \mathbf{w_n}$, vectors corresponding to the respective terms $w_1, w_2, \ldots, w_n$ of the dictionary obtained in the first phase of the proposed solution.

3.1.4 Sub-symbolic Dictionary Creation

The sub-symbolic coding of words obtained with the LSA modifies the semantic comparison between two terms to a simple comparison of vectors. The main problem in this phase is the choice of an index to measure the vector distance and therefore the semantic distance between two words. the vectors. If $\mathbf{w_i}$ and $\mathbf{w_j}$ are the vectors corresponding to two generic terms of the documents collection, we have used the cosine of the angle θ between them:

$$\cos\theta_{ij} = \frac{\mathbf{w_i} \cdot \mathbf{w_j}}{\left\|\mathbf{w_i}\right\| \cdot \left\|\mathbf{w_j}\right\|} \tag{6}$$

Values of cosine near to 1 implicate stronger semantic relationships among terms more meaningful than those corresponding to values of cosine next to 0.

After the definition of the semantic space, the vector coding of the words and the measure of the semantic distance between words mapped in the semantic space, both the sub-symbolic coding of the words and the semantic relationships between couples of words has been memorized in a relational database that will constitute a sub-symbolic semantic dictionary.

3.2 Query Expansion Module

The query expansion sub-system processes the user query to find the meaningful terms to the search eliminating the stop-words. It adds to the query's terms a number *g* (degree of expansion, experimentally determined) of similar words extracted by the sub-symbolic database that memorizes the semantic relationships among couples of words.

The initial terms of the query and the semantically correlated terms according to the implemented metrics are therefore properly combined through logical operators in order to build an expanded query to be given as input to a traditional search engine.

If *q* is a given query, constituted by *n* terms:

$$q = \{w_1, w_2, \ldots w_n\} \tag{7}$$

we suppose to consider for every term of the query the *g* (degree of expansion) correlated terms extracted from the data base:

$$w_i = \{s_{i,1}, \ldots w_{i,g}\} \tag{8}$$

with $0 \leq i \leq n$. The terms extracted by the data base are those which correspond the values of cosine with the query's terms as elevated as possible.

The query expanded will have the following expression:

$$(w_1 \vee s_{1,1} \vee \ldots \vee s_{1,g}) \wedge \ldots \wedge (w_n \vee s_{n,1} \vee \ldots \vee s_{n,g}) \tag{9}$$

3.3 Semantic Documents Filtering

The aim of the second module is to filter the documents returned by the search engine so to discard those that are not semantically related to the initial query. This implies that it is necessary to sub-symbolically codify the documents returned by the search engine in a semantic space. In this manner, it is possible to realize a semantic comparison between the vector representing the user query and the vectors representing the documents.

3.3.1 Sub-symbolic encoding of the query

The query is vectorially codified, using the same criterion used for the vectorial coding of the documents.

Is *q* the query constituted by *k* terms, $\mathbf{q}=\{w_1,w_2,\ldots,w_k\}$, for query vector will intend, the vector, **q** obtained in the following way:

$$\mathbf{q} = \mathbf{w}_1 + \mathbf{w}_2 + \ldots + \mathbf{w}_k \tag{10}$$

The evaluation of the semantic relevance has been made using the cosine of the angle of vectors:

$$\cos\theta_{(qd_j)} = \frac{\mathbf{q} \cdot \mathbf{d}_j}{\|\mathbf{q}\| \cdot \|\mathbf{d}_j\|} \tag{11}$$

with $0 \leq j \leq N$, and *N* is the number of the collection documents.

If *C* is the set of documents $\{d_j\}$ returned by the traditional search engine, then *C'*, with $C' \subset C$, is the set of documents returned by the module that implements the semantic filtering. The elements of *C'* are obtained using the following rule:

$$C' = \{\mathbf{d}_i : \cos\theta_{(qd_j)} \geq T\} \tag{12}$$

where T is a threshold.

The filtering module therefore gives to the user a subset of the documents returned by the traditional search engine whose semantic similarity with the user's query overcomes the threshold T.

The choice of the value of T is essential in order to improve the performances of the system: a high value of T can determine the exclusion of relevant documents from the set of returned documents. A low value of T can involve the introduction of noise (non relevant documents) with consequent decreasing of the precision of the proposed system.

The system offers to the user a twofold choice for the value of the threshold T:

- **Static threshold**: the threshold T has a constant value between -1 and 1, and it doesn't depend on the query;
- **Dynamic threshold:** the threshold T has a value that depends on the query results and is evaluated with statistic methods.

  In particular, the value of T is the following:

  $$T = \mu + \sigma \tag{13}$$

  where μ is the average of the values obtained calculating the cosine of the angle $\theta_{(qdj)}$ between the vector representing the query and the vectors corresponding to all the documents returned by the search engine, while σ is their associated standard deviation.

4 Experimental results

The proposed technique has been implemented using a collection of documents constituted by *650* documents given by couples question-answer extracted by the Internet newsgroups FAQ (Frequently Asked Questions)[15].

A single document of the collection corresponds to a specific FAQ item (question-answer pair). The application of the LSA technique to this collection and the following dimensional reduction on which it is based on, allows us to get the semantic space where words can be mapped[2]. The semantic relationships between couples of words are then stored in the semantic dictionary database.
Experimental trials have been conducted using the well-known traditional search engine jakarta lucene[14]. Documents have been indexed using this search engine, and then the system modules have been used to test the performance improvement of the whole system.

4.1 Evaluation Strategies

The IR system proposed has been evaluated, using the precision (percentage of relevant documents on the documents retrieved by the system), and recall (percentage of relevant documents retrieved on the relevant documents of the whole collection), for a test set of *50* queries.
For every query the values of precision have been calculated in correspondence to standard recall levels [10%, 20%, 30% ,…, 100%] using the following strategy of interpolation: the precision of the system to a given level of recall r_i is equal to the maximum value of the precisions $p(r_i)$ calculated to the superior or equivalent levels of recall of the considered level[1]:

$$p(r_i) = \max\{p(r_j)\} \qquad (14)$$

with $j \geq i$.
The obtained (precision, recall) results have been reported in a Cartesian diagram.

4.2 Performance Comparison

To verify the efficiency of the proposed system, it has been tested the choice of the threshold T, then a lot of search strategies have been implemented and compared.

4.2.1 Analysis of the Influence of the Threshold T

To estimate the influence of T, in figure 2 it is reported the average Precision/Recall diagram obtained using, respectively:

- A standard search engine (labelled as "Lucene"), whithout neither any expansion of the query, nor any filtering of the resuts;
- The proposed system with query expansion with *g=3* and a static filtering threshold fixed to 0.5 (label "T=0.5")
- The proposed system with query expansion with *g=3* and a dynamic filtering threshold according to eq. 13 (label "Dynamic T")

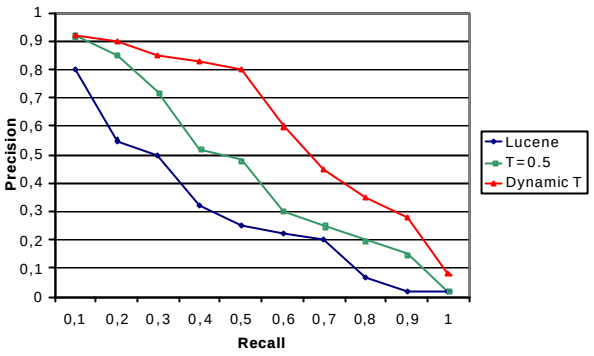

Fig. 2: Precision vs Recall diagram of a traditional search engine ("Lucene"), the proposed system with a fixed treshold ("T=0.5"), and the proposed system with a dynamic treshold ("Dynamic T")

As it can be seen, the choice of a dynamic treshold T allows a clear emprovement of a traditional search engine: from figure 2 it can be seen that for the same value of recall, we obtain a higher value of precision and for the same value of precision we obtain a higher recall value

4.2.2 Analysis of Query Expansion and Results Filtering

To test the semantic expansion of the query many strategies of search have been implemented and the retrieval performances of the system have been evaluated using the same set of queries:

1. search without expansion ("Lucene") (degree of expansion= 0).
2. search with expansion associating to every meaningful term of the query respectively one (QE g=1), three (QE g=3), five (QE g=5) terms extracted by the semantic database (degrees of expansion= 1, 2, 5).

In figure 3 the average values of precision obtained in correspondence to the standards levels of recall are reported.
The shape of the curves expresses the typical link of inverse proportionality that characterizes the values of precision and recall. We observe that to the increase of the expansion degree, for the same value of precision, the corresponding value of recall is higher. Therefore the semantic expansion of the query increases the performances of a traditional IR system: it increases the recall, because it allows recovering also those documents that are semantically but not lexically related to the words used in the query. Besides, the shapes of the curves in figure 3 show a general improvement of the

performances: for the same value of recall we can observe a clear increase of precision values.

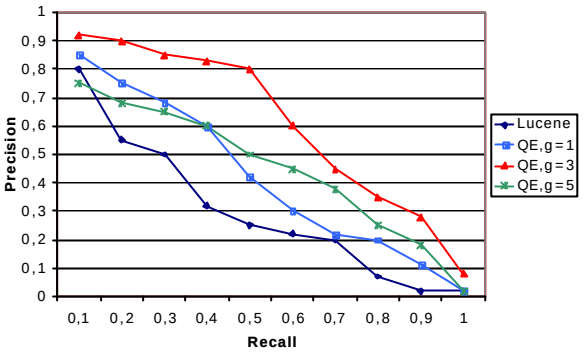

Fig. 3: Precision vs Recall diagram of a traditional search engine ("Lucene"), the proposed system with expansion of the query with 1 ("QE,g=1"), 3 ("QE,g=3"), and 5 ("QE,g=5") semantic similar terms given by the automatically created semantic space.

This result finds a justification in the fact that the documents retrieved by the search engine are semantically filtered to select only those documents that are semantically associated to the user query.
The degree of expansion cannot be however excessively increased, we can observe as, for instance, the introduction of a number of correlated terms equal to 5 deteriorates the positive effects introduced by the expansion. This behaviour is due to the fact that increasing too many extra terms increases also the possibility of introducing noise-words that lead to a higher number of non-relevant documents returned by the system.

5 Conclusions

A system based on the LSA paradigm to improve the performance of a traditional information retrieval system has been presented.
The system has therefore the advantage to be used in conjunction with other, pre-existing, information retrieval systems and is capable to improve their effectiveness in terms of both precision and recall. Experimental trials show the effectiveness of the proposed solution.
Future work will regard the improvement of the proposed technique, a more accurate study of the expansion and filtering techniques and their implementation in other information retrieval related tasks.

References:

[1] R.Baeza-Yates Berthier Ribeiro-Neto, *Modern Information Retrieval.* Addison Wesley, 1999.

[2] M.W.Berry, S.T. Dumais and G.W. O'Brien, Using Linear Algebra for Intelligent Information Retrieval, 1994.

[3] Jos'e De P'erez, Maritza L. Calder'on, Cristina N. Gonz'alez, Towards an Information filtering system in the web integrating collaborative and content based Techniques, *Proceedings of the first Latin American Web Congress*, 2003.

[4] W. Hersh, S. Price, L. Donohoe, Assessing Thesaurus-Based Query Expansion using the UMLS metathesaurus, *Proceedings of the AMIA 2000 Annual Symposium, Los Angeles*, 2000.

[5] Hust, S. Klink, M. Junker and A. Dengel, Query expansion for web information retrieval, *in: S. Schubert, B. Reusch, N. Jesse (edts) Proceedings of Web Information Retrieval Workshop, 32nd Annual Conf. Of the German Informatics Society, Dortmund,* Germany (2002), pp. 176-180.

[6] H. Imai, N. Collier, Jun'ichi Tsujii, A Combined Query Expansion Approach for Information Retrieval, *in Proc. of Genome Informatics*, Tokyo, Japan, 1999, pp292-293.

[7] T.K.Landauer, P.W. Foltz and D. Laham, An Introduction to Latent Semantic Analysis, *Discourse Processes*, 1998, pp. 259-284.

[8] Paul Ogilvie, Jamie Callan.: The Effectiveness of Query Expansion for Distributed Information Retrieval, *In Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management, Atlanta, Georgia, USA*, November 5-10, 2001.

[9] M. Sahlgren, Jussi Karlgren, R Coster and T. Jarvinen, Automatic Query Expansion Using Random Indexing. *Swedish Institute of Computer Scinece* , 2002.

[10] G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[11] Xuehua Shen, ChengXiang Zhai, Exploiting Query History for Document Ranking in Interactive Information Retrieval, (poster). *In Proceedings of 26th Annual International ACM SIGIR Conference*, 2003.

[12] G. Vassallo, G. Pilato, A. Maggio, A. Puglisi, S. Gaglio, Sub-Symbolic Encoding of Words, *Lecture Notes in Artificial Intelligence*, No. 2829, 2003, pp. 449-461.

[13] C.J.Van Rijsbergen B.Sc.,Ph.D.,M.B.C.S., *Information Retrieval*, Butterworths, 1975.

[14] http://jakarta.apache.org/lucene/

[15] Usenet FAQ archives, www.faqs.org