International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 8 (2016) pp. 135–144
© MIR Labs, www.mirlabs.net/ijcisim/index.html

The Impact of Term Statistical Relationships on Rocchio's Model Parameters For Pseudo Relevance Feedback

Nesrine Ksentini, Mohamed Tmar, Faiez Gargouri

MIRACL Laboratory
University of Sfax
City ons, B.P. 3023, Sfax, Tunisa
ksentini.nesrine@ieee.org, mohamedtmar@yahoo.fr, faiez.gargouri@gmail.com

Abstract: **Query Expansion using the Pseudo Relevance Feedback based on rocchio's model is a popular technique for reformulating the original user's query. This latter, assumes that most frequent terms in the returned documents are useful to ameliorate the original query and therefore to improve search results which is a challenge today with the proliferation of textual data on the web.**

In this study, we re-examine this assumption and show that it does not hold in reality. Indeed, many expansion terms identified are unrelated to the query and reduce the performance of the retrieval system.

In this paper, we present our method to revisit the rocchio's model parameters in order to take into account the relationships between terms which are defined by our proposed statistical method based on least square optimization.

The evaluation process was performed on CLEF-eHealth-2014 database which is composed of about one million medical and english documents and 50 professional and medical queries. Experimental results show that our proposed method is effective.

Keywords: information retrieval, automatic query expansion, pseudo relevance feedback, rocchio's model, semantic relationships, least square method

I. Introduction

Huge amounts of information are today widely accessible on the Web. At the same time, users are finding it difficult to express and to satisfy their information needs. In this regard, search engine systems are heavily invoked to access this information in an effective way.

Most search engine systems are based only on the users query to select information assumed as relevant in order to meet their needs. This is challenging as users usually use short queries (about three terms) and essentially contain little contextual information.

Query expansion via pseudo relevance feedback (PRF) method is an effective technique for improving user's queries and for boosting the global performance of information retrieval (IR) systems [1, 2, 3, 4].

This method assumes that top returned results in the first-pass retrieval are relevant. Afterwards it uses these feedback results (documents) in order to refine the representation of the original queries through adding the most related terms.

Even if PRF has been shown to be effective in a number of IR tasks in order to improve IR performance, traditional PRF can also fail in some cases [5, 6].

For instance, when some of the feedback documents have many incoherent subjects, terms in these irrelevant results may mislead the feedback model by adding noisy terms to the queries which influence the retrieval performance in a negative way. This negative effect is warranted because the original query was ignored in the process of query expansion. Indeed, the relationships between selected terms and the query terms have been ignored in PRF models [5].

The most commonly used PRF model is rocchio's model which is a classic method for query expansion [6, 7]. It is an effective relevance feedback method based on the relevant judged documents, by selecting the most representative terms and adding them to the user's query.

However, rocchio's model which was stemmed from the SMART Information Retrieval System around the year 1970 [8] and was developed using the Vector Space Model (VSM) do not take into account the relationship between the original query terms and the selected terms from the top ranked documents in the first pass retrieval.

Therefore, we notice the great motivation to make an extension of the rocchio's model in order to take into account the relationships between query terms and studied terms in the top returned documents.

In this paper, we propose how to revisit rocchio's model in order to take into account the semantic relationships between the added terms and the terms of the original query.

The main contributions of this paper are as follows. First, we propose how to determine relationships between terms using the vector space model. Second, we study how to include these defined relations in rocchio's model. Finally, experiments on Clef-Eheath database have been carried out to evaluate our proposed model.

The remainder of this paper is organized as follows.

In section 2, we present a literature review of the different methods to determine semantic relationships between terms and of the existing query expansion techniques. In section 3, we focus on describing the details of the proposed method. In section 4, the evaluation process is presented and discussed. We draw at the end the conclusion and outline future works in section 4.

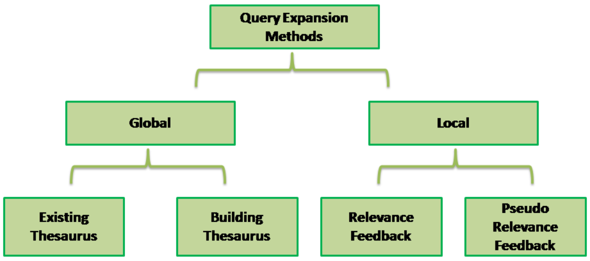


Figure. 1: Classification of query expansion methods

II. Related Works

Query Expansion methods via PRF are very useful and popular techniques which reformulate the original query by adding new terms in order to obtain better results that meet user's needs.

There are a large number of studies on the topics of query expansion and semantic relationships between terms, but here we mainly review works which are most related to our research.

A. *Query Expansion*

Several studies have proposed query expansion techniques, the aim of these techniques is to extend the user's queries with new terms to solve the problem of lexical gap due to what the user wants exactly and how to express this need. Thereafter, the aim is to increase recall of the search system by retrieving more relevant documents [1] [2] [3].

The proposed techniques for query expansion [1] to avoid this problem could be categorized into two main categories illustrated by figure 1:

- Global methods.

- Local methods.

The most used methods of query expansion are global analysis, which does not take into account the results obtained from the original query. This method is based on using some form of thesaurus either existing or manually or automatically generated [9].

One of the most common existing lexical-semantic resources is WordNet [10] which is an external lexical database developed by linguists in the Cognitive Science Laboratory at Princeton University (Hearst, 1998).

Its aim is to identify, classify and relate in different ways the lexical and semantic content of the English language. The information on these semantic resources (nouns, adjectives, verbs) is grouped into synonym sets called synsets where each group presents a distinct concept and is interlinked with conceptual and lexical semantic relations such as meronymy, hypernymy, $\cdots$.

Therefore, the query is expanded by related and synonym words from the thesaurus which appears in the groups of terms that express the initial query.

For the manually-built thesaurus, human editors have built up sets of synonymous names for concepts or sets of related terms by using statistical approaches to calculate and determine the related keywords (will be detailed in the following sub-section).

As a solution to reduce the cost of a manual thesaurus, automatic generation of tresaurus by analyzing the document collection was set up [1] [11]. The idea is to calculate a co-occurrence weight based on the similarity between two terms. Starting from a term-document matrix X, where each cell $X(t,d)$ presents the tf-idf (Term Frequency-Inverse of Document Frequency) weight of term t_i in the document d_j ($w(t,d)$), we calculate the matrix $A = (X \times X^T)$ where $A(i,j)$ is the score of similarity between terms i and j.

Although some terms in the thesaurus are at least suggestive or good, others are bad. The quality of the associations is generally a problem and may cause a query drift. For that, taking into account the results returned for a query in the query expansion process is quite efficient to improve search results.

The latter is studied in the local methods of query expansion (see figure 1).

For relevance feedback, the user is asked to participate in the research process to improve the returned results by marking or evaluating some of the returned documents as either relevant or irrelevant. Then, the system formulates a new query based on the user's feedback [1].

In Pseudo relevance feedback, also called blind relevance feedback, which is an automatic method for local analysis, the user is not asked in the retrieval process to give his relevance judgments. In this method, the top k documents returned by the first search process are assumed as relevant [1, 2, 3, 4].

The purpose of this assumption is to find suitable terms to expand the original user query from these top documents (for example by selecting high weighted terms (tf-idf)) and to return better results to the user without any interaction.

This method is generally effective and it tends to be better than the manual local analysis method and the global analysis method.

In [2], the authors proposed a query expansion method based on PRF and equi-frequency of partition of the documents using the tf-idf scores.

Indeed, the authors assumed that relevant information can be found within a document near the main theme. They divided the document into sections (paragraphs and lines) and they proposed a method which tries to extract the keywords that are closer to the central idea of the document.

In fact, the expansion terms are obtained by equi-frequency in partition of the documents returned from the first-pass retrieval (PRF) and by using tf-idf scores.

In [3], tf-idf is measured to extract keywords that appear frequently in top returned documents and which are ranked in descending order of their weights.

Indeed, the authors applied two query expansion methods in sequence to reformulate the initial query. The first method is to apply the similarity thesaurus based on expansion, and the other method was to apply local feedback method. The similarity thesaurus whose used, based on terms similarity in the top returned documents and queries, is the sum of the weighted relevance of the term to each term in the query.

After that, the queries were expanded by adding top n relevant terms, which are most similar to the query concept, rather than selecting terms that are similar to the query terms.

The rocchio's model is a basic and a classic framework for implementing PRF to improve the query representation [6, 7, 8]. It makes a way how to incorporate relevance feedback information in the first-pass retrieval into the VSM in IR. In PRF, the factor of irrelevant documents in the Rocchio equation is ignored.

Indeed, the new representation of the query is finally refined by taking a linear combination of the initial query vector with the vectors of returned relevant documents.

The formula of this method is as follows:

$$\vec{Q_1} = \alpha * \vec{Q_0} + \beta/|R| * \sum_{\vec{d} \in R} \vec{d} \qquad (1)$$

Where:
$\vec{Q_1}$ represents the new query vector.
$\vec{Q_0}$ represents the original query vector.
$|R|$ represents the set of returned documents assumed as relevants.
$\vec{d}$ represents the document weight vector in the R set (represented as a weighted terms vector).
α represents the original query weight.
β represents the related documents weight.

we notice that α and β are constant values which control how much we rely the original query and the feedback information. In practice, we set α at 1, and β is defined by experiments until we get better performance.

However, this model does not take into account the relationships between the original query terms and the new added terms. Also, weight of added terms is always the same (β is a constant).

Intuitively, the exploitation of semantic relationships between terms and the variation of weight added terms could be included in the rocchio's model [12].

B. Semantic Relationships between Terms

Measuring similarity and relatedness between terms in the corpus becomes decisive in order to ameliorate search results [10]. Earlier approaches that have been investigating the latter idea can be classified into two main categories: those based on pre-available knowledge (ontology such as wordnet, thesauri, etc.) which are detailed in the previous sub-section, and those inducing statistical methods [13], [14] which are based only on the contents of the databases.

In [13], the authors present how to calculate the similarity of two terms by collecting snippets containing the first term from a Web search engine, retrieving the context around it, replacing it with the second term and checking whether the context is modified or not.

Another statistical method was presented by [14], which proposes a new way for calculating semantic similarity. Authors collect snippets from the returned results and present each of them as a vector. The similarity is calculated as the inner product between the centroids of the vectors corresponding to a pair of terms.

A further measure of relatedness between terms is term proximity which is the co-occurrences of terms within a specified distance [15]. Particularly, the distance is the number of intermediate terms in a document [5] [15] [16].

Several works have been done to integrate term proximity into both probabilistic and language models.

In PRF model, the authors in [5] have studied how to adapt the traditional rocchio's model [8] for proximity information, and propose a proximity-based feedback model, called PRoc, in which the traditional statistics of expansion terms and the proximity relationship between expansion terms and the query terms are taken into account.

We notice from this overview that few works included relationships between terms which are incorporated in PRF models specially in rocchio's feedback model. These relations play an important role in information retrieval field and are usually defined between pairs of terms.

In this paper, we propose, in the first time, a new statistical method to calculate and define semantic relationships between terms in the top k returned documents by the first search process. This method called "least square method" and arisen in machine learning applications provides linear relationships between a set of terms and not only between pairs of terms [17][18][19] [20].

In the second time, defined relations will be studied and used in rocchio's model in order to take into account the relationships between query terms and expanded terms.

III. Proposed Method

In this section, we present our proposed method based on a local automatic document-analysis to define semantic relationships between terms of the top k returned documents, and terms of the original user's query. Thereafter, we explain how to revisit rocchio's model in order to take into account these defined relations when we extend automatically the original query. The process of our search system is illustrated in (figure 2).

When the user describes his need, our information retrieval system retrieves documents that meet this latter by calculating similarity between the query and the documents in the database. We assume that top k documents are relevant. These documents will be analyzed by the proposed statistical least square method that allows to define relationships that may exist between terms appearing in these documents [17] and terms in the submitted query.

Afterwards, we apply revisited rocchio's model for pseudo relevance feedback technique to automatically expand the original query with terms that are in strong relationships with terms in the original query. At the end, information retrieval system will restart to find similar and relevant documents to this need.

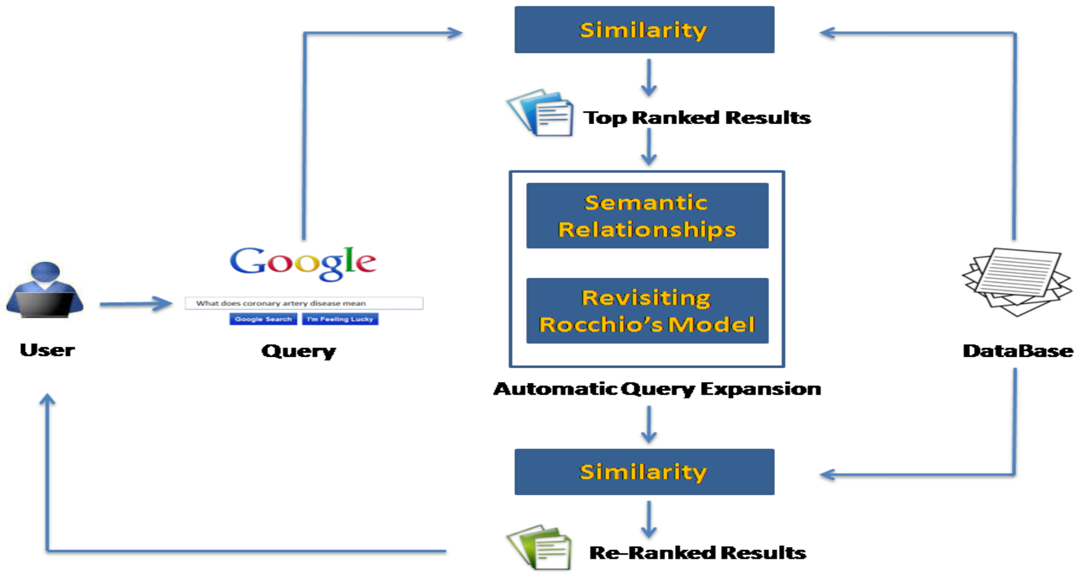


Figure. 2: Process of our search system

A. Semantic Relationships

Our aim is how to define relationships between terms $(t_1, t_2, \cdots, t_n)$ that exist in the top k returned documents from the search system and query terms.

Indeed, we attempt to find relations that may exist between each term t_j of the query (for $j \in 1, \cdots, m$, with m is the total number of query terms), which exists in the keyword set of the returned documents, and each term t_i of this last set (for $i \in 1, \cdots, n$, with n is the total number of returned terms) with the following form:

$$t_j = f(t_1, t_2, \cdots, t_{j-1}, t_{j+1}, \cdots, t_n) \tag{2}$$

The least square method [17], [18], [20] is a frequently used method to solve approximately this kind of problems.

Indeed, this method, known as linear regression, is the oldest and most widely used predictive model in the field of machine learning [21][22]. It tries to find the connection that may exist between an explained variable (y) and explanatory variables (x). It is a procedure to find the best regression line $(y = ax + b)$ to (x_i, y_i) data observed for $i \in 1, \cdots, n$, where a represents the coefficient of relations between variables x and y, and b represents the residual or the error which represents the disruption of the regression model.

The objective is to find, for the given data, the values of a that minimize the error (Err) given by:

$$Err = \sum_{i=1}^{n} (b_i)^2 = \sum_{i=1}^{n} (y_i - ax_i)^2 \tag{3}$$

In our case, let query term (t_j), that exists in the keyword set of the documents assumed as relevant, be the explained variable and the remaining terms (keywords) of the top k returned documents $(t_1, t_2, \cdots, t_{j-1}, t_{j+1}, \cdots, t_n)$ the explanatory variables (see figure 3).

The goal is to find the relation between these variables as follows:

$$\begin{aligned} t_j &\approx \alpha_1 t_1 + \alpha_2 t_2 + \cdots + \alpha_{j-1} t_{j-1} + \alpha_{j+1} t_{j+1} \\ &+ \cdots + \alpha_n t_n + \epsilon = \textstyle\sum_{i=1}^{j-1} (\alpha_i t_i) + \sum_{i=j+1}^{n} (\alpha_i t_i) + \epsilon \end{aligned} \tag{4}$$

Where α represents the real coefficients of the regression model that are the weights of relationships between terms and ϵ is the associated error.

We have k measurements for the explained and the explanatory variables which represents the TF-IDF values of these variables in the top k documents. Our goal is to calculate the appropriate $\alpha_1, \alpha_2, \cdots, \alpha_n$ for the whole set of following equations:

$$\begin{cases} t_j^1 \approx \alpha_1 . t_1^1 + \alpha_2 . t_2^1 + \cdots + \alpha_n . t_n^1 \\ t_j^2 \approx \alpha_1 . t_1^2 + \alpha_2 . t_2^2 + \cdots + \alpha_n . t_n^2 \\ \vdots \\ t_j^k \approx \alpha_1 . t_1^k + \alpha_2 . t_2^k + \cdots + \alpha_n . t_n^k \end{cases} \tag{5}$$

Where t_j^k is the TF-IDF weight of term j in document k.

Using the matrix notations the system becomes:

$$\underbrace{\begin{pmatrix} t_j^1 \\ t_j^2 \\ \vdots \\ t_j^k \end{pmatrix}}_{T} \approx \underbrace{\begin{pmatrix} t_1^1 & t_2^1 & \cdots & t_n^1 \\ t_1^2 & t_2^2 & \cdots & t_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ t_1^k & t_2^k & \cdots & t_n^k \end{pmatrix}}_{X} \times \underbrace{\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}}_{A} \tag{6}$$

where vector T represent TF-IDF values of term (t_j) and X present a TF-IDF matrix whose columns represent the keyword set and rows represent the returned documents.

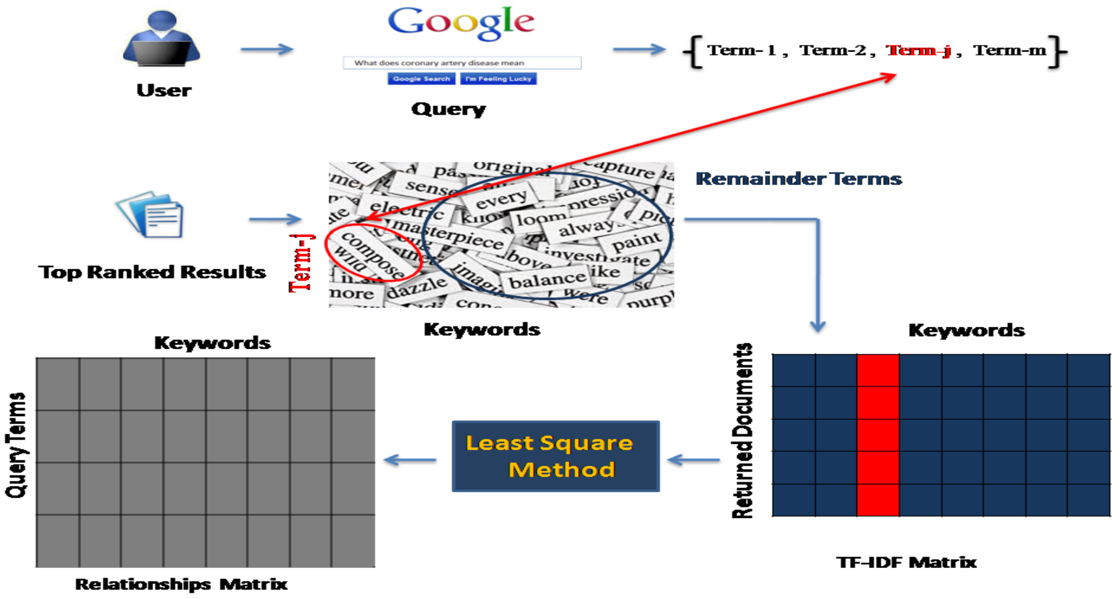


Figure. 3: The steps for defining semantic relationships between terms

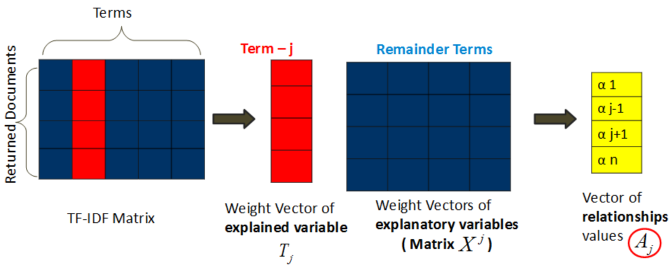

Terms
Returned Documents
TF-IDF Matrix
Term – j
Weight Vector of explained variable T_j
Remainder Terms
Weight Vectors of explanatory variables (Matrix X^j)
α 1
α j-1
α j+1
α n
Vector of relationships values A_j

Figure. 4: Process of the proposed least square method

Therefore, we search the weight vector $A = (\alpha_1, \cdots, \alpha_n)$ of relationships between terms such as $X \times A$ is more nearer to T. Least Square Method gives the solution to find this vector in an approximate way:

$$A = (X^T \times X)^{-1} \times X^T \times T \tag{7}$$

To determine the vector A_j for each term (t_j) (see figure 4), we applied this regression model on the matrix X.

$\forall j = 1, \ldots, m,$

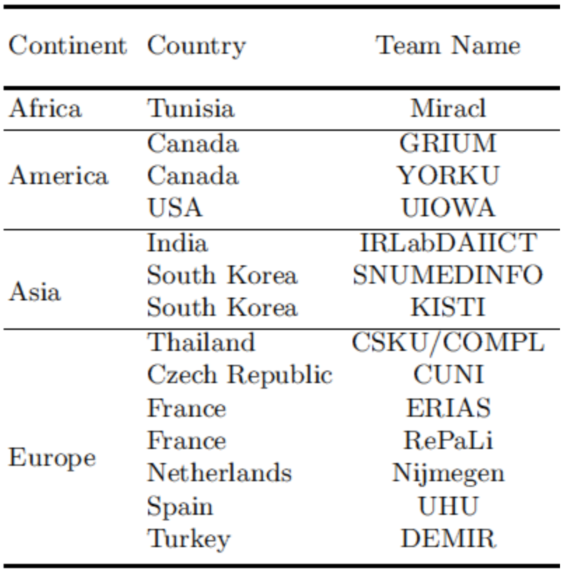
$$A_j = (X^{jT} \times X^j)^{-1} \times X^T[., j] \times T_j \tag{8}$$

Where:
X^j is obtained by removing the column of the term t_j in matrix X.
$X^T[j, .]$ represents the transpose of the weight vector of the term t_j in all documents. At the end of this process, we obtain terms by terms matrix (see figure 3) (Query terms × keyword set matrix) which contains the relation values founded for each query term with the remainder terms.
Once the relations are defined, we study how to include them in rocchio's model. Thus, it will take into account semantic relationships in the automatic query expansion process.

B. Revisited Rocchio's Model

The rocchio's model is a classic framework for implementing PRF to improve the query representation. It incorporates relevance feedback information in the first-pass retrieval into the VSM in IR.
Generally, In PRF works, irrelevant documents are ignored in the Rocchio's model.
Indeed, the new representation of the query is finally refined by taking a linear combination of the initial query vector with the vectors of returned relevant documents (see equation 1) but without taking into account relations existing between these vectors.
We study in this section, how to incorporate defined relations in the rocchio's model.
After applying least square method between the original query and the returned documents, we had as a result semantic relationships between the terms of these.
This result is shown in a matrix form (see Figure 3) where rows represent the query terms, columns represent the keyword set of the relevant documents and each cell indicates the found degree or weight of the relationship.
For each term $(term_i)$ in the keywords set, we check if it is very related to all terms in the query (all $\alpha_i > threshold$ with $i \in 1, \cdots, n$). If it is the case, we expand the original query by the $term_i$ with a weight which can be the maximum of the different values of α or the mean of α values.
We notice from this assumption, that relationships between query terms and keywords set are taken in consideration in the rocchio's model with different values of beta and not with a constant value.
Indeed, we expand automatically the original query with relevant and related terms.

IV. Results and Discussion



Continent	Country	Team Name
Africa	Tunisia	Miracl
America	Canada	GRIUM
	Canada	YORKU
	USA	UIOWA
Asia	India	IRLabDAIICT
	South Korea	SNUMEDINFO
	South Korea	KISTI
Europe	Thailand	CSKU/COMPL
	Czech Republic	CUNI
	France	ERIAS
	France	RePaLi
	Netherlands	Nijmegen
	Spain	UHU
	Turkey	DEMIR

Figure. 5: Participants for CLEF-eHealth Competition [23]

In order to check the performance and validity of our proposed method, an experimental procedure was set up. This evaluation was carried out on a large collection of documents from the CLEF company after the participation in CLEF-eHealth competition in 2014 [23][24] (with Miracl team name).

Indeed, 14 groups participated in this competition and submitted runs to information retrieval task (see figure 5).

A. Database

The document collection used for the experimental study is composed of a set of medical documents covering a wide set of medical topics and does not contain information about patients. This collection is about of one million documents provided by the Khresmoi project [23] which come from different online sources such as known databases and medical sites (e.g. Genetics Home Reference, ClinicalTrial.gov, the health certified websites).

These crawled documents are given with their raw HTML (Hyper Text Markup Language) format and their uniform resource locators (URL).

The test set comprises 50 professional and medical queries provided by experts in the domain (doctors for example). Given queries present different cases of patient diseases [23].

1) The Indexation Process

After extracting HTML documents from the raw files, the indexation process of the crawled documents was carried out with the terrier platform developed at the School of Computing Science, University of Glasgow [25]. It is a clear, flexible and efficient open source written in java language and easy to aplly it on a large collection of documents.

2) The Search Process

Our system identifies the most relevant documents for each provided query taking into account our proposed method for pseudo relevance feedback.

Indeed for each query, the system searches at the beginning the top k ranked documents from the one million documents in the database using the vector space model to calculate the similarity between the topic (query) and the documents in the CLEF collection with the cosine measure.

We seek, for these top k documents, terms or tokens that they form and we prepare the TF-IDF matrix.

After that, we check out for each term in the query whether it exists in the token list of the most relevant returned documents or not. If it is the case, we calculated the coefficients of relationships (α) that may exist between this term and terms of the token list with the least square method defined above.

Calculated values of α were stored in another matrix called the relationship matrix with the rows represent the query terms and the columns represent token list terms.

Once all terms in the query are all checked, we apply rocchio's model to expand the original query with related terms based on the defined relations with the least square method.

Indeed, for each term in the token list, if it is very related to query terms, then we add it. In fact, we add to the original query only terms which have α values that are above a certain threshold and with weight which can be the maximum of the different values of α or the mean of α values.

Finally, we restart the search with the new query and we display the relevant results to the user.

B. The Results

In our participation last year to CLEF-eHealth 2014 competition [24], we submitted one run which consists of the baseline system based only on the vector space model (VSM). We have obtained results with Mean Average Precision (MAP) equal to 0.17 and p@10=0.5460 (which represents the precision of the search system on the first 10 results).

After the integration, this year, of the proposed method in our research system, but without modifying the weights of added terms (we let weights as constant) [20], we have obtained the first results that range between 0.20 and 0.23 for MAP measure when adding terms which are related to all terms in the query (see table 1) and between 0.13 and 0.18 when adding terms which are related to at least one term in the query (see table 2).

In the first case (table 1), values of chosen k are higher than values chosen in the second case (table 2), because we want to have more terms and more defined relationships since we add those terms that are related to all the terms of the initial query.

But in case 2, the values of α are higher than those chosen in case 1. In fact, we do not add the terms with only positive values of relationship, but terms that are highly related with the terms of initial query since we add those terms that are related to at least one query term.

Table 1: Experimental results obtained by adding terms related to all terms in the query

k	50			100		
	$\alpha > 0$	$\alpha > 0.3$	$\alpha > 0.5$	$\alpha > 0$	$\alpha > 0.3$	$\alpha > 0.5$
MAP	0.2099	0.23	0.23	0.2204	0.2289	0.2297
P@5	0.46	0.49	0.49	0.49	0.50	0.50
P@10	0.46	0.50	0.50	0.48	0.50	0.50

Table 2: Experimental Results obtained by adding terms related to at least one term in the query

k	10		25		50	
	$\alpha > 0.5$	$\alpha > 0.7$	$\alpha > 0.5$	$\alpha > 0.7$	$\alpha > 0.5$	$\alpha > 0.7$
MAP	0.17	0.18	0.15	0.18	0.13	0.14
P@5	0.46	0.43	0.39	0.44	0.33	0.37
P@10	0.40	0.39	0.35	0.38	0.29	0.34

Take as examples these topics (queries) in CLEF2014:

$< topic >$
$< id > qtest2014.1 < /id >$
$< title >$ Coronary artery disease $< /title >$
$< desc >$
What does coronary artery disease mean
$< /desc >$
$< /topic >$

We took in this example the top 100 documents returned by the baseline system as relevant. This set form a list of 4377 terms. Then we calculate the relations between terms in the query with this set of terms and we expanded the original query by adding terms which have values of $\alpha > 0$ (positive relationships) and which are related to all terms in the initial query.

We have obtained this new query:

$< topic >$
$< id > qtest2014.1 < /id >$
$< title >$ Coronary artery disease$< /title >$
$< desc >$
coronari arteri diseas mean myocardi bypass angiographi nstemi nospac tvr aortic charlson clot andrew unstabl antiplatelet vein ptca cholesterol fogoro blocker nitroglycerin atherosclerosi mmhg bytreat pravastatin mmol dissect coronarographi intracoronari heart-attack linhartov angioplastyst ticagrelor chest-pain-angina tnt toclevel interven
$< /desc >$
$< /topic >$

Where the four first underlined terms present the original query and the other terms present the added terms in their root form.

We notice that almost all added terms are in strong relationship with the terms of the initial query. For example, terms like myocardi, aortic, clot, vein, atherosclerosi and chest-pain-angina present the the sugnes of heart disease such as myocardite, atherosclerosis, clot and chest-pain-angina. Terms like angioplastyst, dissect, interven, ptca (percutaneous transluminal coronary angioplasty) describe the diagnostic and operations that can be performed on patients affected by this disease. Some other terms present some medication names for this disease such as nitroglycerin and ticagrelor.

We can conclude from the obtained results that our proposed method in [20] can improve search results when adding terms which are related to all terms in the query.

We notice that the added terms appear in the same context of the query and do not make the context drift to the initial query (illustrated example above).

But, in the second type of the evaluation when adding terms that are related, at least one term in the query, we notice that the results does not improve the values of MAP but in some cases they are slightly affected.

For example, when we took 50 top documents and we expanded the original query by adding terms which have values of $\alpha > 0.5$, we obtain better results in the first case (MAP=0.23, p@10=0.50) than the second case (MAP=0.13, p@10=0.29).

The obtained results from the proposed method in this paper are also motivating. Indeed, when we expand the original user's query by the most related terms and with their calculated weights (revisited rocchio's model), we obtain for MAP measure values nearly to 0.30. Experiments are based on the different values of k and on how to calculate the final weight of the term that will be added to the query. In fact, we obtain these results when setting k at 10 and the final weight represents the mean of the different values of α found between this term and query terms.

Take as examples these topics (queries) in CLEF2014:

$< topic >$
$< id > qtest2014.35 < /id >$
$< title >$ Peptic Ulcer disease $< /title >$
$< desc >$ What kind of food is recommended after being diagnosed with peptic ulcer$< /desc >$
$< /topic >$

$< topic >$
$< id > qtest2014.49 < /id >$
$< title >$ Chronic lymphocytic leukemia and hereditarity
$< /title >$
$< desc >$
Is chronic lymphocytic leukemia hereditary
$< /desc >$
$< /topic >$

We took for each query the top 10 documents returned by the baseline system as relevant. This set form a list of 973 and 808 terms respectively. Then we calculate the relations between terms in the query with each set of terms and we expanded each original query by adding terms which have the mean of α values > 0 (positive relationships).

We have obtained this new query:

$< topic >$
$< id > qtest2014.35 < /id >$
$< title >$ Peptic Ulcer disease $< /title >$
$< desc >$
kind food recommend be diagnos peptic ulcer antacid duodenum
$< /desc >$
$< /topic >$

The added terms in this query are (antacid) which presents a kind of recommended medicament and (duodenum) which presents the C-shaped or horseshoe-shaped structure that lies in the upper abdomen near the midline.

$< topic >$
$< id > qtest2014.49 < /id >$
$< title >$ Chronic lymphocytic leukemia and hereditarity $< /title >$
$< desc >$
chronic lymphocyt leukemia hereditari pathobiolog asco
$< /desc >$
$< /topic >$

The added terms in the query number 49 are (pathobiolog) which presents the term pathobiology that is very related to the context query and (asco) which presents the term ascot, this is a medical center specific for this disease.
In comparison with other works which participated at echealthCLEF 2014 [23] and used the pseudo relevance feedback in their proposed methods with the same database, we find Team CSKU-COMPL [23] which used vector space retrieval model of Lucene as baseline.
As improvement, they proposed a simple pseudo-relevance feedback method which used the Genomic collection as external resource to perform query expansion. The expansion terms selection is based on the Rocchio's formula with dynamic tunable parameter of Pseudo-relevance feedback. Their run obtained a Map=0.20 after query expansion.
We can conclude after this discussion, that proposed method can improve user's query. Thus, results obtained from search engine system are also improved.

V. Conclusions and Future Work

In this paper, a revisited rocchio's model is proposed by incorporating semantic relationships between terms into the classic rocchio's model. Specifically, We expand automatically the original query, without any user interaction, by related terms with their similarity weight values. These values of relationships between expansion terms and query terms calculated by a proposed statistical method in an information retrieval task called least square method. This latter allows to define the relations that may exist between terms in the top ranked documents based on their frequency in the corpus and query terms.
Our purpose is to improve the user's need in order to get better search results in a large database. Obtained results from this method are motivating and show the originality of the proposed method to define semantic relationships.
As future work, firstly, we will look for using techniques to reduce the dimensionality like the matrix decomposition technique to refine the term-document matrix for a better representation (more abstract level: conceptual level) and to facilitate the search process and we will try to use the proposed method in this paper for the new generated matrix to define the relations on the concept level.
Secondly, we will try to evaluate the proposed method with other database such as CLEF-eHealth-2015 and to participate at the CLEF-eHealth competition in 2016.

References

[1] Carpineto, C. and Romano, G. ”A survey of automatic query expansion in information retrieval”. ACM Computing Surveys (CSUR),vol. 44, no 1, p. 1, 2012.

[2] Vaidyanathan, R., Das, S., and Srivastava, N. ”Query Expansion Strategy based on Pseudo Relevance Feedback and Term Weight Scheme for Monolingual Retrieval”. arXiv preprint arXiv:1502.05168. 2015.

[3] Aly, A. A. ”Using a query expansion technique to improve document retrieval”. ,International Journal Information Technologies and Knowledge”, 2(4), 343-348. 2008.

[4] Singh, J., and Sharan, A. ”A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach”. Neural Computing and Applications, 1-24. 2016.

[5] Miao, J., Huang, J. X., and Ye, Z. ”Proximity-based rocchio's model for pseudo relevance”. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. p. 535-544. ACM. 2012.

[6] Singh, J., and Sharan, A. ”Relevance Feedback Based Query Expansion Model Using Borda Count and Semantic Similarity Approach”. Computational intelligence and neuroscience, 2015.

[7] Rocchio, J. ”Relevance feedback in information retrieval.”,313-323, 1971.

[8] Salton, G., and Buckley, C. ”Improving retrieval performance by relevance feedback”. Readings in information retrieval, 24(5), p.355-363. 1997.

[9] Pal, D., Mitra, M., and Datta, K. ”Improving query expansion using WordNet. Journal of the Association for Information Science and Technology”, 65(12), 2469-2478, 2014.

[10] Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paca, M., and Soroa, A. ”A study on similarity and relatedness using distributional and WordNet-based approaches”. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 19-27). Association for Computational Linguistics, 2009, May.

[11] Manning, Christopher D., Raghavan, P., and Schutze, H. ”Introduction to information retrieval”. Cambridge : Cambridge university press, 2008.

[12] Jordan, C., and Watters, C. ”Extending the rocchio relevance feedback algorithm to provide contextual retrieval.” In AWIC, 135-144, 2004.

[13] Ruiz, C., Maria, Enrique, A., and Pablo, C. ”Using context-window overlapping in synonym discovery and ontology extension.” Proceedings of RANLP-2005, Borovets, Bulgaria 39, 2005.

[14] Sahami, M., and Heilman, T. ”A web-based kernel function for measuring the similarity of short text snippets”. In : Proceedings of the 15th international conference on World Wide Web. AcM,. p. 377-386, 2006.

[15] Tao, T., and Zhai, C. ”An exploration of proximity measures in information retrieval”. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. p. 295-302. ACM. 2007.

[16] Vechtomova, O., and Wang, Y. ”A study of the effect of term proximity on query expansion”. Journal of Information Science, 32(4), p.324-333. 2006.

[17] Abdi, H. ”The method of least squares.” Encyclopedia of Measurement and Statistics. CA, USA: Thousand Oaks, 2007.

[18] Miller, S.J. The method of least squares. Mathematics Department Brown University, p. 1-7, 2006.

[19] Ksentini, N., Tmar, M., and Gargouri, F. ”Detection of semantic relationships between terms with a new statistical method”, Proceedings of the 10th International Conference on Web Information Systems and Technologies, Barcelona, Spain, p.340-343, 2014.

[20] Ksentini, N., Tmar, M., and Gargouri, F. ”Controlled automatic query expansion based on a new method arisen in machine learning for detection of semantic relationships between terms”. In The 15 th International Conference on Intelligent Systems Design and Applications (ISDA). Morocco. 2015. in press.

[21] Andrew Ng. CS229 lecture notes, http://cs229.stanford.edu/notes/cs229-notes5.pdf, 2012.

[22] Huang, G. B., Wang, D. H., and Lan, Y. ”Extreme learning machines: a survey”. International Journal of Machine Learning and Cybernetics, 2(2), 107-122, 2011.

[23] Goeuriot, L., and al. ”Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval.”, 2014.

[24] Ksentini, N., Tmar, M., and Gargouri, F. ”Miracl at CLEF 2014: eHealth information retrieval task”. Proceedings of the ShARe/CLEF eHealth Evaluation Lab, 2014.

[25] Ounis, I. and Amati, G. and Plachouras, V. and He, B. and Macdonald, C. and Lioma, C., ”Terrier: A High Performance and Scalable Information Retrieval Platform”, 2006.

[26] A. Bonnaccorsi. On the Relationship between Firm Size and Export Intensity, *Journal of International Business Studies*, XXIII (4), pp. 605-635, 1992. (journal style)

[27] R. Caves. M*ultinational Enterprise and Economic Analysis*, Cambridge University Press, Cambridge, 1982. (book style)

[28] M. Clerc. The Swarm and the Queen: Towards a Deterministic and Adaptive Particle Swarm Optimization. In *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*, pp. 1951-1957, 1999. (conference style)

[29] H.H. Crokell. Specialization and International Competitiveness, in *Managing the Multinational Subsidiary*, H. Etemad and L. S, Sulude (eds.), Croom-Helm, London, 1986. (book chapter style)

[30] K. Deb, S. Agrawal, A. Pratab, T. Meyarivan. A Fast Elitist Non-dominated Sorting Genetic Algorithms for Multiobjective Optimization: NSGA II. *KanGAL report 200001*, Indian Institute of Technology, Kanpur, India, 2000. (technical report style)

Author Biographies

Nesrine Ksentini she recieved the Master’s degree in Computer Science from the university of Sfax, Tunisia, in 2011. She is currently a phd student from the same university.
Her main research topic is semantic in information retrieval systems (IRS). Member of IEEE since 2010.
She is a member of Multimedia Information systems and Advanced Computing Laboratory, Sfax, Tunisia. She published 6 papers in international conferences.

Mohamed Tmar holds Ph.D. in Computer Science, University of Paul Sabatier, Toulouse, France (2002). He is a member of Multimedia Information systems and Advanced Computing Laboratory, Sfax, Tunisia.
His research interests are information retrieval, information filtering, XML and multimedia retrieval, query optimization and language modeling. He published more than 50 papers in journals and conferences.

Faiez Gargouri is Professor of computer sciences at the Higher Institute of Computer science and Multimedia of Sfax, Tunisia (www.isimsf.rnu.tn), where he is the Head (since August 2014). He was the Director of the Multimedia, InfoRmation Systems and Advanced Computing Laboratory (www.miracl.rnu.tn) from 2011 to 2014. From 2007 to 2011 he was the Head of the same institute. Faiez Gargouri obtained his Master's degree in Computer Science from the Paris 6 University (1990) and a PhD from the Paris 5 University (1995). In 2002, he obtained an Habilitation Universitaire en Informatique from the Facult des Sciences de Tunis (Tunisia). His research interest focuses on different information systems fields, such as, Design, Quality Measurement, Verification, Data Warehousing, Multimedia, Knowledge Management, Ontology. He published more than 100 papers in journals and conferences as well as books (pedagogical and conference proceedings). He is member of the Scientific and Steering committees of various international conferences. He is namely one of the founding father of the JFO conference (French Workshops on Ontologies) and ASD (workshop on decisional systems). Faiez Gargouri is the founding chairman of the scientific association AIG (Association of computer management).