The Query Expansion Method "QUEXME" in an application environment

Guillermo Valente Gomez Carpio, Lylia Abrouk, Nadine Cullot

The Query Expansion Method "QUEXME" in an application environment

Guillermo Valente Gómez Carpio, Lylia Abrouk, Nadine Cullot
LE2I, UMR CNRS 5158
University of Burgundy
Dijon, France
gomezcarpio@yahoo.com
{lylia.abrouk, nadine.cullot}@u-bourgogne.fr

ABSTRACT: *The aim of the paper is to present and apply a QUery EXpansion MEthod called QUEXME while querying the Euro-Mediterranean Information System (EMWIS) on know-how in the Water sector. EMWIS provides a strategic tool for exchanging information and knowledge in the water sector between and within the Euro Mediterranean partnership countries (www.emwis.net). Information retrieval on the web or through some cooperation of information sources or some general knowledge bases is a complex process and a great challenge with the emergence of the semantic web. The aim of the query expansion method is to help and guide users to build their requests giving them some usually related terms close to their queries. Information retrieval in EMWIS is based on the use of a thesaurus to query the information system and to find relevant documents on some specific topics in the water sector. This thesaurus can be viewed as a light-weight web ontology. It is multilingual. This paper proposes an experimentation of our query expansion method within the framework of the EMWIS information system.*

Keywords: Query Expansion, Enrichment, Ontology, Thesaurus, Method

Received: 2 November 2009, Revised 29 December 2009, Accepted 2 Janurary 2010

1. Introduction

Since Semantic Web was envisioned, one of the motivations has been the semantic search, but also has led to an increase complexity of the information retrieval process. Search engines allow to find resources generally associated with keywords resulting from annotations done on the available resources. Indexing documents with keywords is a difficult task to realize manually because the results can change from one expert to another. The use of ontologies in the annotation process offers several advantages such as helping in solving information retrieval problem with a "free" vocabulary.

A lot of applications need to access and share more and more sophisticated information. Considering the domain of the information management of the water sector, a large volume of information has to be taken into account. However, at both international and national levels, the know-how in the water sector remains fragmented, dispersed and heterogeneous. Therefore, there is a real need to rationalize the information and to make it both understandable and easily available. The EMWIS (Euro-Mediterranean Information System on Know -how in the Water Sector) project has been initiated in order to facilitate access to existing information on know-how in the water sector and develop the sharing of information.

Searching and retrieving information is not enough if the results do not meet the user needs. However, searching for the most relevant information is still a challenging problem especially when information is coming from heterogeneous sources. To help in tackling this problem, we present in this paper an expansion query approach based on concept popularity we have developed and we apply this method in the context of the Euro-Mediterranean Information System.

The rest of the paper is organized as follows: In section II are presented some related works on query expansion. A brief description of the EMWIS (Euro-Mediterranean Information System on Know-How in the Water Sector) environment and the

context of our work are developed in Section III. Section IV shows the experimentation that reminds our method QUEXME of query expansion based on the notion of Concept Rank. Section V explains the validation of QUEXME of this experimentation realized in the EMWIS context, and finally the conclusion and some perspectives are given in section VI.

2. Related Works

This section explains the concept of query expansion and presents some works that have developed this subject.
The objective of query expansion is to enrich the initial query of the user to obtain as relevant as possible answers. The aim is to help the user to query a system or the Web. Two steps are necessary to carry out the enrichment process:

1. The formulation of the initial query by the user.
2. The enrichment of his query.

For the user, the enrichment process can be more or less visible and it can be completely enclosed into the query process or more loosely coupled requiring the help of the user.

Among the authors who have worked to improve the query expansion, we can mention E. Efthimiadis [7], who was the first to propose a classification of the query expansion methods. We agree his definition that "query expansion (or the term expansion) is the process of supplementing the original query with additional terms, and it can be considered as a method for improving retrieval performance". He performs a classification of methods according to these following criteria:

- The interaction of the user in the process to help in selecting additional terms.
- The kinds of resources which are used to find the additional terms of a query.
- The methods used to select the terms to be added.

On this occasion to support our work, we will focus on interactive query expansion methodologies and mention the different possible resources which can be used for the expansion process. We can note that the resources can i) relate to the relevant feedback process (to identify pertinent resources), ii) be based on the profile of the user or iii) on other type of knowledge (corpus-based or not).

Some related works found are as follows:

A widget proposed by Touminen et al. [16] uses two ways for query expansion: general and domain-specific ontologies and the Spatio-Temporal Ontology SAPO. It also uses related terms to enrich the query. In the first case, the query is expanded with subclasses of the query concepts using the transitive is-a relation. In the second case, if the query is a spatial query, the query expansion uses the SAPO ontology to expand the spatial terms including possible temporal information. The query expansion widget supports semantic web and legacy systems and the interface created integrates ONKI Ontology Browsers widgets.

The system proposed by Gong et al. [10] applies the query expansion from several semantic relations between words, uses also an average precision to determine in quantity how to expand the query. It enriches the query along three dimensions including hypernym, hyponym and synonym relations. The system is a method of Web query expansion that uses the WordNet as lexical dictionary and Term Semantic Network TSN as a filter and a supplement. The algorithm proposed processes the keyword expansion using two functions “confidence and support” in describing word relations and some matrix to store these information. The system works only with single word queries.

Hust et al. [11] proposes a system based on query similarities and relevant documents (QSD) to make the query expansion. It uses feedback information and the ground truth that provides a list of relevant documents for each existing query. It computes a document vector from each set of relevant documents. To enrich the new query the system uses the document vectors and a weighting scheme.

Another system is proposed by N. Messai [12]. It is dedicated to available bioinformatic data sources. The lattice and also the generalization/specialization relations of a domain’s ontology are used in the query expansion process. The system allows

the construction of a lattice of concepts including the terms used in the users' queries. Each concept represents a pair: (resources, metadata) where metadata are also some sets of terms.

Ontologies and relevance feedback are two kinds of resources used by the system proposed by E. Schweighofer [15]. It uses the ontology to extract related terms occurring in the query such as synonyms, sub-terms, etc. These terms provided in a query are searched in the knowledge base (ontology) and weighted. After, the Boolean retrieval makes the selection.

Languages and tools have been developed with the emergence of the semantic web. One of them is SPARQL. This is a language dedicated to query RDFS sources (www.w3.org/TR/2000/ CR-rdf-schema-20000327/). Q. Zhou [18] proposes a system called SPARK. It is a tool which allows the translation of a query consisting in a set of terms into a SPARQL query. Term mapping, query graph construction and query ranking are the tree major steps that composed the translation.

Proposed by C. Raymond [14], the SIAC system is based on an unsupervised classification method and a query reformulation. The query expansion method uses decision trees and a Boolean expression is associated to each tree node. The terms used to enrich the query are extracted from the previous retrieved documents.

A system based on a probabilistic method is proposed by H. Cui [5]. The terms are analyzed in the system assuming that they are corralled to the terms in the documents that the user clicked on. It extracts correlations between query terms and document terms by analyzing query logs. To expand the query, the best terms from documents called high quality are used.

Y. Qiu [12] also proposes a method which takes into account the whole query and considers terms similar to the "query concept" instead of working on each term of the query. This method uses a probabilistic query expansion and a weighting model based on a similarity thesaurus.

The work of M. Bertier [3] is an applied method in a system, called Gossple, which aims at improving the exploration of internet using social networks. It is a query expansion method based on the use of users' tags which builds a matrix, called TagMap which contains the relationships and the distances between the tags. It uses an algorithm based on the PageRank algorithm to find the terms "tags" close to the user query to expand it.

This system query expansion proposed by J-C. Bottraud [4] is done in two phases: vocabulary learning and queries learning. According to his search the user profile is updated. To enrich the queries, the system builds a user profile in order. It identifies a context based on profile and query's user.

Another that uses different resources such as the users' profile is proposed by C.A. Zayani [16]. To make the query expansion, this system compares the similarity between query elements and user profile elements in order.

To sum up, the systems presented are based on different query expansion methods. They use individually resources to compute new terms to enrich users' queries. The resources used are: search results (feedback, user's profile or user's behavioral) and "external" resources as knowledge structures (thesaurus or ontologies).

We propose an expansion query method which is based on the usual behavior of the users and a knowledge structure. We can say that 1) our method is interactive: additional terms are proposed to the user who can select the most adapted to his needs, 2) the process uses the usual behavior of the users to find relevant terms, so a learning phase is necessary and 3) our method is part of information retrieval systems that have evolved together with the emergence of the semantic web. In conclusion, we can consider that our proposed method mixes the main characteristics of the mentioned systems. The next sections detail the context and the algorithm.

3. EMWIS Context

This section presents an overview of the project EMWIS (Euro-Mediterranean Information System on Know-How in the Water Sector) with its objectives and the actors of the project. Also, we give an overview of the architecture, and then a special focus is put on the thesaurus of the project which allows to query the cooperation. We have applied our query expansion method "QUEXME" in this thesaurus which can be viewed as light-weighted ontology. And finally, the general process of querying including the enrichment phase is depicted.

A. EMWIS Architecture

EMWIS[1] (Euro-Mediterranean Information System on Know-How in the Water Sector) is an information system that is intended to allow to share and approve information on know-how in the water sector covering the Euro-Mediterranean countries. It is an initiative of the Euro-Mediterranean Partnership.

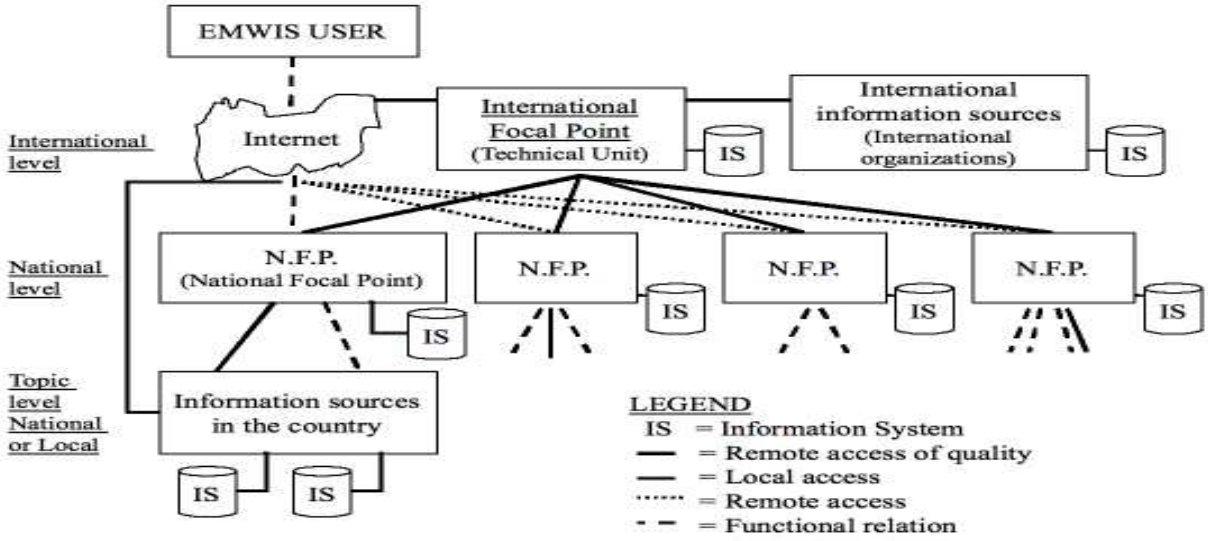

Figure 1. EMWIS architecture

The architecture has three levels: International level, national level and topic level. EMWIS architecture [1] is illustrated in figure 1. At international level, the interest of such system basically comes from the aggregation of individual pieces of national information. The production of national content is linked to the availability of National Water Information Systems offering interoperable services for real-time integration.

At national level, the information is provided by the various "National Focal Points (NFP)" or by international organizations which participate in the system. The information remains managed by these "local" providers. As far as possible, the information is not centralized in a unique database located in one server but is distributed in the various organizations involved and acting as information nodes.

And at local level, the Technical Unit acts as a facilitator in helping each National Focal Point to set up their information system and ensuring the coordination among all the NFPs.

In EMWIS, needs for content organization and structuring are generated by the rapidity of the evolution of the information in all domains. To represent the semantic and the organization of the data exchanged in a particular domain, it can use ontologies, in order to facilitate the internal and external communication between the different actors of the domain.

B. EMWIS Ontology

A light-weighted ontology has been designed in the project in order to share information and resources. This ontology is multilingual and provides a hierarchical description of the concepts of the domain with some semantic links such as synonyms, related to, etc. To each concept of the ontology are associated some resources. EMWIS ontology is a thesaurus which defines terms with semantic relations. These terms are used to build user's queries. Figure 2 gives the hierarchical structure of the themes and some of the concepts in the thesaurus.

[1] *www.semide.net*

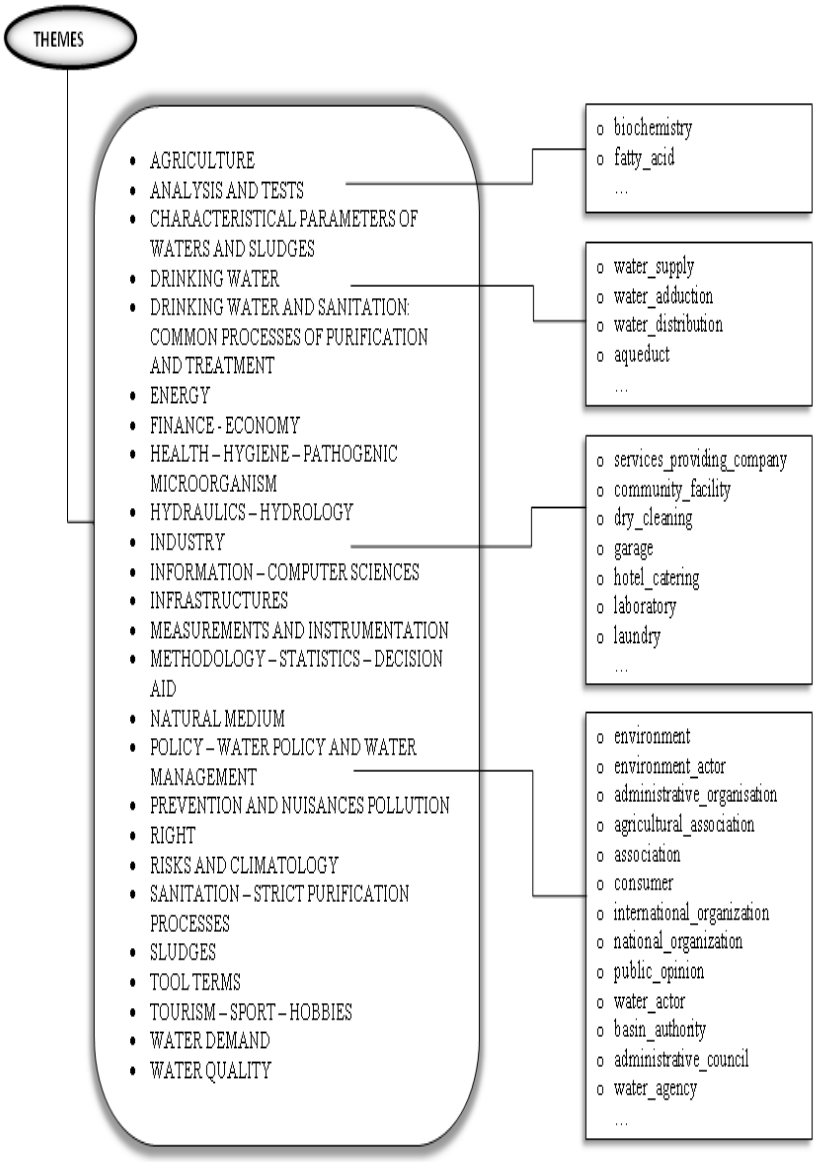

Figure 2. EMWIS themes and some concepts

We can list as characteristics of EMWIS ontology:

- It is based on the thesaurus of the International Office for Water (OIEau[2]).
- It was developed to share common understanding of the structure of information among people or software agents.
- It contains more than 1400 concepts.
- The associated terms are in three languages (English, French and Arabic).
- It can be enriched by user queries [2] and documents corpus [6].

In EMWIS, the information research process is based on the ontology. And to construct a query, the users pick ontology concepts. In this paper we present an expansion query method in order to enrich the query results in the EMWIS system.

[2] *www.oieau.fr*

C. Query Expansion Process

The query process, including the query expansion phase in the EMWIS search engine, can be illustrated as shown in figure 3.

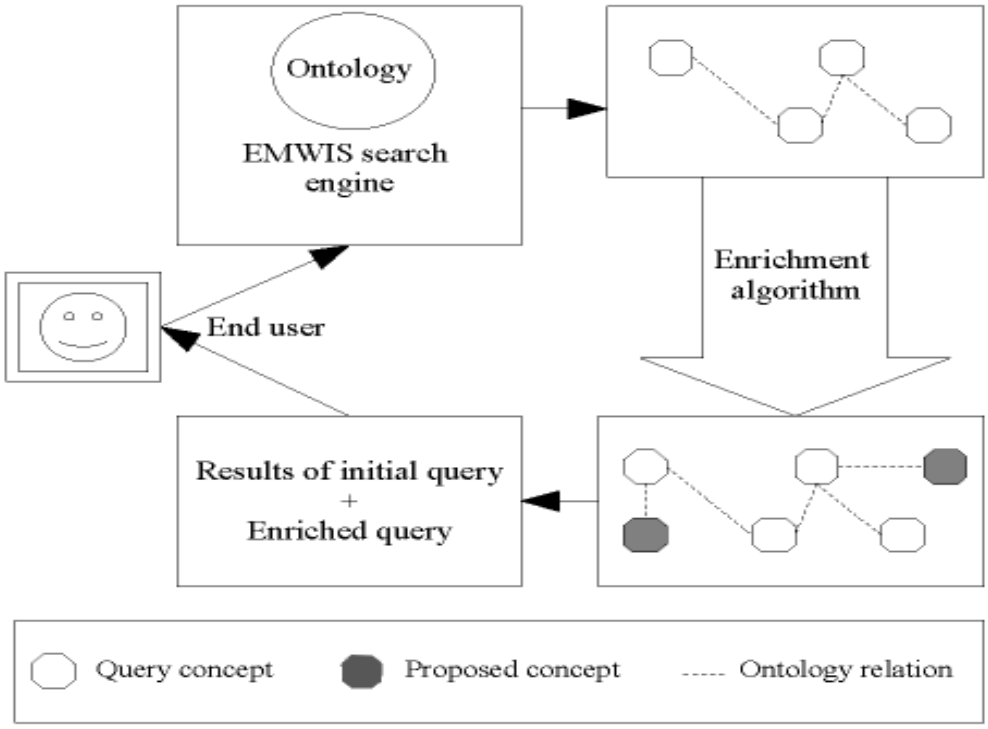

Figure 3. EMWIS query expansion process

The query expansion process aims at proposing to the user relevant terms to enrich his query. First, the user submits his query based on terms of the ontology. To facilitate user choices, concepts are organized in topics. Then, the enrichment algorithm is applied and the new query with the additional terms is proposed to the user with the results of the initial query. Last, the user can refine his request using these new terms or selecting some of them. It should be noted that in QUEXME, the query expansion process can be decomposed in two phases as shown in figure 4. A learning phase is the first phase. The "usual" behavior of users is the basis of our query enrichment method. So, it is necessary to have a learning phase to initialize the matrix used to compute the concept rank by the algorithm. Then, the second phase consists in the query enrichment, strictly speaking.

The next section is dedicated the detailed presentation of the query expansion method and its experimentation in the environment EMWIS.

QUEXME Experimentation

This section is dedicated to remind our proposed method [8] and how we have experimented our algorithm on the EMWIS ontology which is, as we discussed before (section III), a light-weighted ontology on the water sector. This experimentation is an extended presentation of [9].

Our approach is inspired from the PageRank algorithm but it is applied in a different context. We define the notion of ConceptRank value which expresses the popularity of a concept and the notion of importance of concept relatively to another

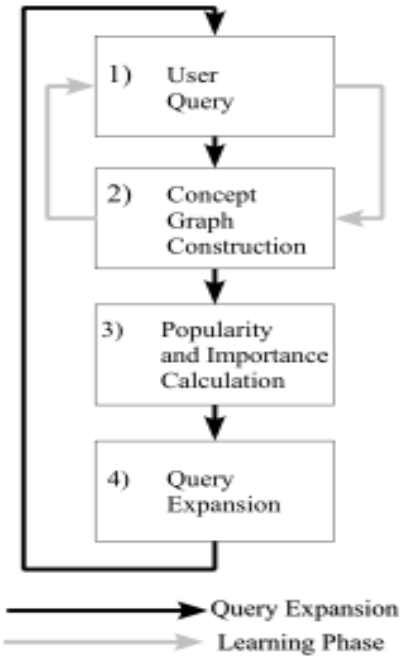

Figure 4. Two phases in the query expansion process.

one or to a whole query. At the end of the enrichment process, the additional terms and the results of the initial query are sent to the user, who validates these results and chooses one or several concepts of the enriched query.

A. Learning phase

We must remember that during the learning phase no calculation of popularity or importance takes place and neither proposition of enrichment is given to the user, as we can see in figure 4.

In this phase, a concept graph is built on the basis of the users' queries. Intuitively, the nodes of the graph are the concepts and an edge from a node N1 to a node N2 expresses that the concept of the node N2 has been searched after the concept of the node N1 by a user.

The graph is a directed and weighted. First, we introduce the definitions of query sequence and sub-sequence and then we define the notions of concept graph for a sub-sequence and sequence of queries.

Definition: Let C be a set of concepts of a domain, *Q* a query which is a subset of concepts of *C*, a query sequence is a n-uplet *S(Q1,Q2,...,Qn)* where *Qi* is a query.

A query sub-sequence of a sequence *S(Q1,Q2,...,Qn)* is a 2-uplet *Si(Qi,Qi+1)* where *i=1...n-1*.

The cardinality |*Q*| of a query *Q*, is the number of concepts belonging to this query.

Figure 5 illustrates three queries sequences *S(Q1,Q2,), S(Q3,Q4,Q5)* and *S(Q6)*. Each sub-sequence Si of a sequence S is represented by a concept graph *Gi*.

Definition: A concept graph *Gi(Oi,Vi)* is a directed weighted graph, for a sub-sequence *Si(Qi,Qi+1)* of a sequence *S(Q1,...,Qn)* where:

Vi is the set of edges belonging to $(Qi \cap Qi+1) \times (Qi+1 - Qi)$.

Oi is the set of the concepts belonging to $Qi \cup Qi+1$.

The weight *wi(vi)* associated to an edge $vi \in Vi$ is defined by:

$$wi(vi) \frac{1}{\left|(Qi \cup Qi+1)\right|} \quad (1)$$

It is based on the number of concepts that a user has kept between two successive queries of a sub-sequence *Si(Qi,Qi+1)*.

Definition: A graph G for a sequence *S(Q1,..,Qn)* is the union of the Gi graphs of the sub-sequences *Si(Qi,Qi+1)*.

$$G = \bigcup_{i=1}^{n-1} Gi \quad (2)$$

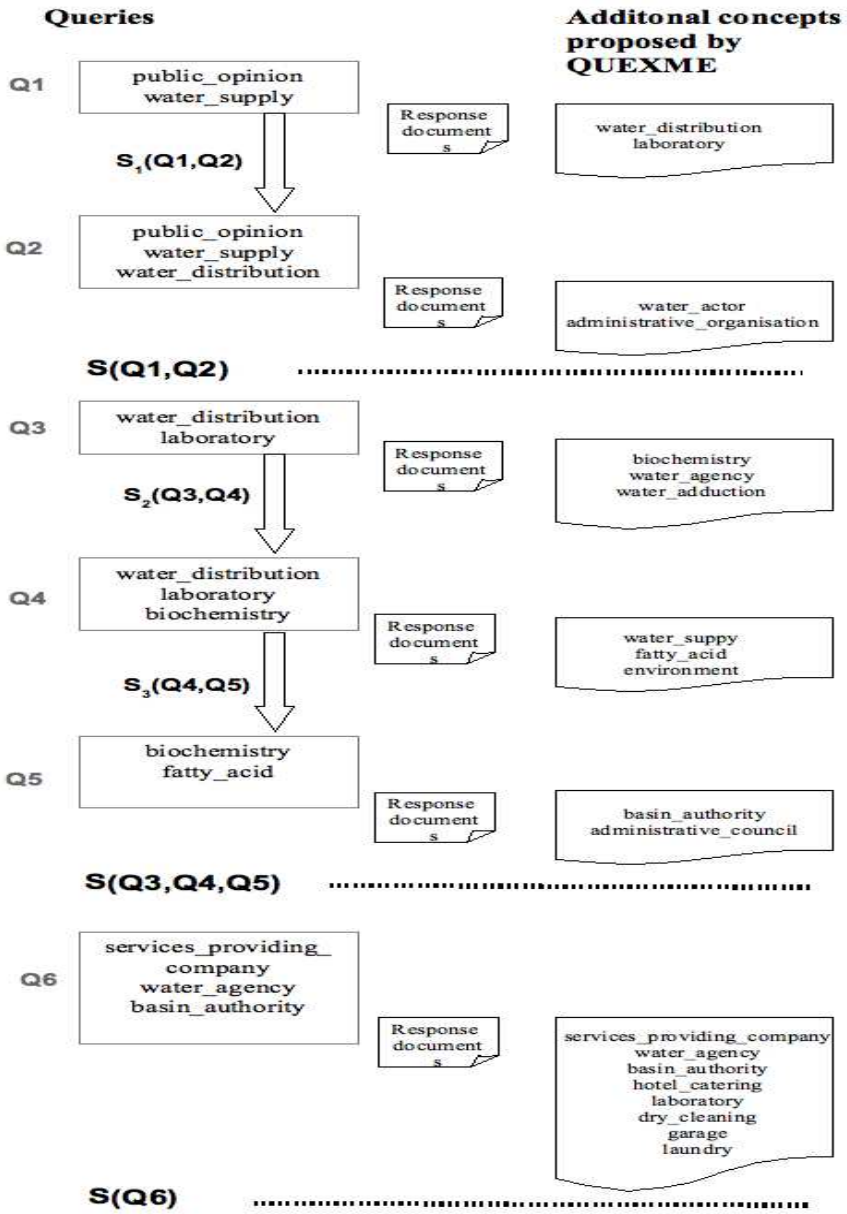

Figure 5. Example of three queries sequences

Figure 6 illustrates the concept graph G1 for the sub-sequence S1(Q1,Q2) shown in figure 5 where we put the concepts of the intersection of the queries Q1 and Q2: (public_opinion, water_supply), the concept belonging to the difference between the two queries (water_distribution) and the associated weights.

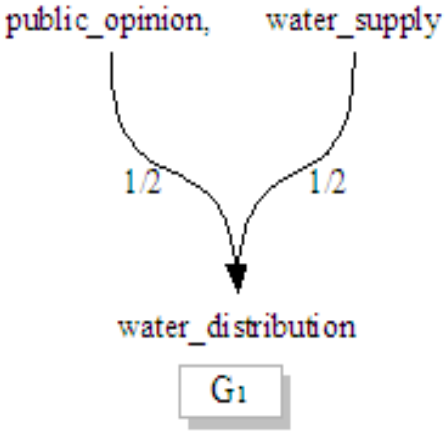

Figure 6. Concept graph of the sub-sequence S1

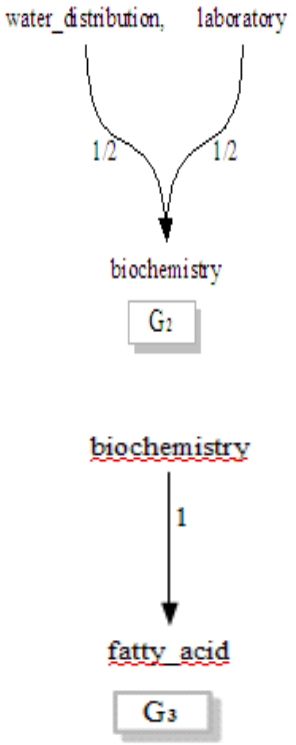

Figure 7. Concept graphs of the sub-sequences S2 and S3

The next sequence S(Q3,Q4,Q5) of the example in figure 5 consists of two sub-sequences S2(Q3,Q4) and S3(Q4,Q5). We put in G2, the concepts of the intersection of the queries Q3 and Q4 (water_distribution, laboratory), the concept of the difference between the two queries (biochemistry) and the associated weights. In the same way, we construct G3 with (biochemistry) and (fatty_acid). Figure 7 illustrates the concept graphs G2 and G3.

The concept graph of a query sequence is translated into a matrix. Intuitively, this matrix contains for each couple of concepts (C1, C2), the sum of the weights of the edges existing between the nodes C1 and C2 in the concept graph.

Below, we first define the notion of matrix for a sub-sequence and then the notion of matrix for a whole sequence.

Definition: A graph Gi for a sub-sequence *Si(Qi,Qi+1)*, can be represented by a matrix MCi, such as:

$MCi : C \rightarrow C$, when C is the set of concepts of the domain.

$$MCi\ (ci,cj) = \begin{cases} wi\ (ci,cj)\ \text{if there is edge}\ vi : ci \rightarrow cj \\ 0\ \text{else} \end{cases} \quad (3)$$

Definition: The matrix *MC* of a graph G of a sequence *S(Q1,..,Qn)* is the sum of the matrix *MCi* of the sub-sequences *Si(Qi,Qi+1)*.

$$MC = \sum_{i=1}^{n-1} MCi \quad (4)$$

For each query sequence, a concept graph is computed and the matrix is incremented.

In the first phase of the learning process, we have computed 50 queries to initialize the matrix *MC*. Table I shows some snatches of the sequences of queries applied during the learning phase. It presents two successive queries submitted by a user, the kept edges of the graph used to update the matrix and the associated weights used to increment the suitable values of the matrix. For example, the two first queries *Q1(environnement,consummer)* and *Q2(environment, water_actor)* lead to the incrementation of the matrix value *MC(environment,water_actor)* of 1.

B. Query expansion phase

In the second phase, the query enrichment process can be applied. All the steps shown in figure 4 take place with each user's question. Here, we detail an example of our query expansion method as shown in figure 5.

Query1	Query2	Edge of the Concept Graph	Weight
environment, consumer	environment, water_actor	environment → water_actor	1
environment, environment _actor	environment _actor, water_actor	environment _actor → water_actor	1
consummer, association	consummer, water_actor	consummer → water_actor	1
association, administrative _organization	association, public_opinion	association → public_opinion	1
public_opinion, international _organization	public_opinion, national _organization	public_opinion → national _organization	1
international _organization, national _organization	administrative _ organization, international_ organization, national _ organization	international _organization → administrative _organization, national _organization → administrative _organization	1/2

Table 1. Example of some query sub-sequences of learning phase

Based on the matrix defined in the previous section, we can now introduce a new measure of popularity called the ConceptRank measure. As in the PageRank algorithm, the computation of ConceptRank measure needs a number of iterations to fix the value. The ConceptRank measure expresses the popularity of a concept according to the usual behavioral of the users during their querying process. The ConceptRank measure is defined as follows.

Definition: Let *ci* be a concept of *C, B(c$_i$)* the set of concepts, such as *MC(c$_j$,c$_i$) ≠ 0, d* is a normalization factor.

The Concept Rank measure of a concept *ci, CR(ci)* is defined by:

$$CR(c_i) = 1-d + d \sum_{cj \in Bci} \left[\frac{CR(c_j)}{CR(c_j)} \right] \quad (5)$$

where we note *N(c$_j$)* the number of the edges of G, going from c_j to another concept.

$$N(c_j) \sum_{k=0}^{n} vk \begin{cases} vk = 1, \text{ if } MC(cj,ck) \neq 0 \\ vk = 0 \text{ else,} \end{cases} \quad (6)$$

We complete this notion of popularity of a concept with the notion of importance of a concept relatively to another concept or to a query.

The notion of importance of a concept relatively to another concept and more precisely the importance of a concept relatively to a whole query is the measure that we use to choose the concepts to add to enrich the user's query. These measures are based on the popularity of the concepts as defined previously with the ConceptRank measure.

Definition: Let *MC* be the matrix representing some graph *G*, and *Q* a query.

CI(c$_i$,c$_j$) is the Concept Importance of the concept c_i relatively to the concept c_j which is defined by:

$$CI(c_i,c_j) = MC(c_i,c_j) \; X \; CR(c_i) \quad (7)$$

where *MC(c$_i$,c$_j$)* is the value of the matrix MC for the couple *(C$_i$,C$_j$)* et *CR(c$_i$)* is the ConceptRank of c_i.

Definition: $I(c_i,Q_i)$ is the Importance of a concept c_i relatively to a query Q_i which is defined by:

$$I(c_i,Q_i) \sum_{c_j \in Q_i} CI(c_i,c_j) \quad (8)$$

where *CI(ci,cj)* is the Concept Importance of c_i relatively to a concept c_j as previously defined.

Minimum Importance Factor Threshold

To choose the concepts to add to expand a query considering their importance relatively to this query, we define a threshold value called the minimum factor importance. A query Qi is expanded with a concept ci if its concept importance *I(ci,Qi)* is greater than this minimum importance factor.

Definition: Let *IQi* be a set of concepts importance values for a query Q_i, $IQ_i=I(c_1,Q_i),..., I(c_n,Q_i)$. The minimum importance factor *Min* (FI/Q_i) for a query Q_i is defined by:

$$Min\,(FI/Q_i) \quad \frac{Max\ (I/Q_i) + Min\ (I/Q_i)}{2} \quad (9)$$

This factor represents the average of the maximum and the minimum values of the importances.

In the second phase of query expansion, for example, we consider a new user's query which is *Q(public_opinion, water_supply)*. The terms of the query are extracted: *public_opinion and water_supply*. These two concepts are then treated by the system generating the respective response documents. This process takes place after each user's question. Then according to the previous built matrix (*MC*), the concept rank of each concept is calculated following the formula (5).

In order to expand the query and select the additional terms, it is necessary to compute the measure of the importance of a concept relative to the query. This measure is done in two steps:

1) First, we compute the importance of this concept relative to another concept of the query. In order to find these values the method applies the formula (7) described above.

2) Then, for each concept of the query, we compute the Importance of this concept relatively to the whole query. The measures of the computation are calculated following the formula (8) given in this section. These measures are computed for each concept.

Table II Overview of some results of the computation of the importance of a concept relative to the query.

Query	**I(c,Q)**
biochemistry	0.8417629
laboratory	2.5507965
environment	1.2141806
national_organization	1.2141806
services_providing_company	0.40067962
agricultural_association	0.0000000
water_distribution	2.5507965

Table 2. Some results of I(C,Q)

According to these results, a threshold, called the minimum importance Min(FI/Qi) is determined to choose which concepts would be selected to enrich the query. Taking into account the values described in the table II, the retained value is given by the formula (10).

$$Min\,(FI/Q_i)\left[\frac{\left[\frac{2.5507 + 0.4006}{2}\right]}{\left[\frac{2.9513}{2}\right]}\right]_{1.476} \qquad (10)$$

All the concepts for which the importance relatively to the query is greater than this minimum factor are proposed to enrich the query. In the presented example, two concepts *water_distribution* and *laboratory* are retained and proposed to the user.

At this time the user's first option is to choose the concepts considered as pertinent to enrich his question and rebuild it. In this case, a new sub-sequence will be generated.

The user's second option is to change the meaning of the question selecting other concepts. In this case, a new sequence will be initiated. Both cases can be seen on figure 5.

Table III shows the incrementation of the values of the matrix of each sub-sequence of our sample sequence 1.

For example, the two first queries *Q1(public_opinion, water_supply)* and *Q2(public_opinion, water_supply, water_distribution)* lead to the incrementation of the matrix values *MC(public_opinion, water_distribution)* and *MC(water_supply, water_distribution)* of 1/2. Then, the user's choice leads to start another cycle of the question enrichment process.

Query1	Query i+1	Edge of the Concept Graph	Weight
public_opinion, water_supply	public_opinion, water_ supply, water_distribution	public_opinion → water_distribution, water_supply → water_distribution	1/2
water_ distribution, laboratory	water_distribution, laboratory, biochemistry	water_distribution → biochemistry, laboratory → biochemistry	1/2
water_distribution, laboratory, biochemistry	biochemistry, fatty_acid	biochemistry → fatty_acid	1

Table 3. Example of query sub-sequences in the query expansion phase

Next section details some results and how the validation of QUEXME has been carried out.

5. Results and validation

After the application of the enrichment method, some concepts considered as pertinent according to the computation of the minimum importance factor are proposed to the user. The user can retained or not these news concepts to rebuild his query. We call percentage of acceptance for a query, the ratio of the number of proposed concepts and the number of retained concepts by the user.

A. Results Table IV shows an example of user' behavior in 16 queries taken at random. We can see, in the first column, the concepts included in the query submitted by the user and in the second column the concepts proposed by QUEXME to enrich his query. In this second column, the concepts which were not retained by the user are shown in gray. Then in the third and fourth columns, we give the number of concepts proposed and the number of accepted concepts by the user. Finally, we present the percentage of acceptance for every query asked by the user.

Query's Concepts	Concepts proposed by QUEXME	Concepts proposed	Concepts selected	% of acceptance
Public _opinion, Water _actor	Water _supply	1	1	100%
Community _facility, Water _agency	Water _supply	1	1	100%
Environment, Biochemistry	Environment _actor	1	0	0%
Water _distribution, Water _adduction	Biochemistry, Water _agency	2	1	50%
Biochemistry, Fatty_acid	Environment _actor	1	1	100%
Consummer, Association	Water_supply, Agricultural _ association, Public_opinion, Dry_cleaning, Garage, Laundry, Water_actor	7	2	35%
Administrative _ council, organisation, basin_authority	Environment, Administrative _ Biochemistry, Fatty_acid	4	2	50%
Water _distribution, Laboratory	Biochemistry, Water_agency, Water _adduction	3	1	33%
Public_opinion, Water_supply	Laboratory, Water _distribution	2	1	50%
Water_agency, Administrative _ council, Basin_ authority	Environment, Administrative _ organisation, Hotel_catering, Biochemistry, Fatty_acid, Dry_cleaning, Garage, Laundry	8	3	27%
Water_supply, Agricultural _ association, Association	Laboratory, Water _distribution	2	1	50%
Service _providing _company, Water_ agency, Basin_ authority	Hotel_catering, Laboratory, Dry_Cleaning, Garage, Laundry	5	5	100%
Water _distribution, Water _adduction, Biochemistry	Environment _actor	1	0	0%
Water_actor, Environment _actor, Association,	Water_supply, Public _opinion	2	2	100%

International _ organization, Biochemistry, Laboratory	Environment _actor	1	1	100%
Environment, Water_ supply, Water _ distribution, Water _ adduction	Laboratory, Water_ actor, Biochemistry	3	1	0%

Table 4. Acceptance

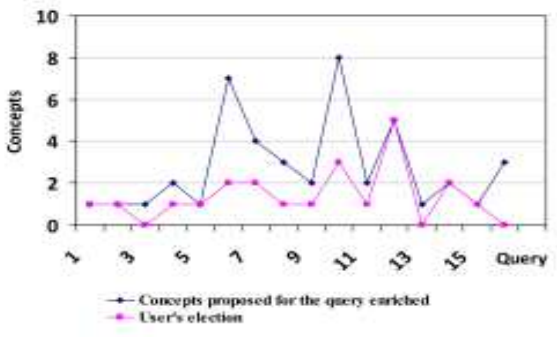

Figure 8. User behavior

With these results, we can observe user' behavior with respect to their percentage of acceptance. In figure 8, the graph shows the number of concepts proposed by QUEXME with the gray dots and the number of concepts validated by that user with the black dots.

B. Validation To validate the experimentation of QUEXME applied to EMWIS system, we have proposed to an expert to analyze about a thousand of enriched queries and to validate or not the new proposed terms for each query. For example, if a query is enriched by N terms, the expert has to validate these terms or not.

From these results, we obtain the conclusions presented in table V, that means that, for 37.5% of the number of queries, 80 to 100% of the terms proposed were validated by the expert, etc.

% of terms accepted by the expert	**% of queries**
80 to 100%	37.5
50 to 80%	25.0
10 to 50%	18.75
0 to 10%	18.75

Table 5. Validation of the enriched queries

The more detailed analysis of these results shows that if the number of proposed terms is too great then the percentage of terms retained by the expert decrease. So, it is necessary to adjust the selection threshold to choose the terms. In this case up to 50% of queries are validated for about 70% of their terms. Finally, another observation is that the validation of the expert is quite similar to user' behavior with respect to the number of concepts.

6. Conclusion

We have presented an experimentation of our query expansion methodology called QUEXME. It has been experimented in the context of the EMWIS (Euro-Mediterranean Information System on Know-How in the Water Sector) project.

This proposed methodology is based on the usual behavior of users to select additional terms to add to the user query to enrich it.

The query expansion process is applied to the EMWIS light-weight ontology available to query the information system in the water sector (EMWIS).

The results of the experimentation have shown that the learning phase has to be carefully done as well as the updating of the historical knowledge stored during the query process. The query sequence has to be semantically coherent to the user's point of view to allow a pertinent selection of additional terms in next queries.

About 50% of the queries are agreed at 70% by an expert, shown by the analysis of the experimentation. To avoid "the corruption" of the available knowledge it is necessary to be able to filter the incoherent sequences of terms, therefore an optimization of our process is currently in progress.

References

[1] Abrouk, L., Mino, E. (2004). A framework to share water information, *In*: 1st International Conference on Information and Communication Technologies: from Theory to Applications (ICCTA'O4), Damascus, Syria, p. 155 - 156.

[2] Abrouk, L., Gouaich, A. (2007). Utilisation des ontologies pour le partage de l'information dans le domaine de l'eau", Journées Francophone sur les Ontologies (JFO'07), Tunisie. p. 393 - 404.

[3] Bertier, M., Guerraoui, R., Leroy, V., Kermarrec, A-M. (2009). Toward personalized query expansion, *In*: Second ACM Workshop on Social Network Systems (SNS'09), Nuremberg, Germany. p. 7 - 12.

[4] Bottraud, J-C., Bisson, G., Bruandet, M-F. (2004). Expansion de requêtes par apprentissage automatique dans un assistant pour la recherche d'information, Conference en Recherche Information et Applications (CORIA'04) p. 89 - 108.

[5] Cui, H., Wen, J. R., Nie, J. Y., Wei, M. (2002). Probabilistic query expansion using query logs, *In*: 11th International World Wide Web Conference (WWW'02) p. 325 - 332.

[6] Di Jorio, L., Fiot, C., Abrouk, L., Herin, D., Teisseire, M. (2007). Enrichissement d'ontologies: Quand les motifs sequentiels labellisent des relations", 23èmes Journées Bases de Données Avancées (BDA'07), Marseille, France. p. 24 - 35.

[7] Efthimiadis, E. (1996). Query Expansion, *In*: Annual Review of Information Systems and Technology (ARIST) p. 121 - 187.

[8] Gómez Carpio, G.V., Abrouk, L., Cullot, N. (2009). A query expansion methodology in a cooperation of information systems based on ontologies, *In*: 5th International Conference on Web Information Systems and Technologies (WEBIST'09), Lisbon, Portugal, p. 256 - 262.

[9] Gómez Carpio, G.V., Abrouk, L., Cullot, N (2009). QUEXME, A query expansion method applied to water information system, *In*: 5th International Conference on Signal Image Technology & Internet Based Systems (SITIS'09), Marrakesh, Morocco, December.

[10] Gong, Z., Cheang, C.W., Hou U, L. (2005).Web query expansion by WordNet", 16th International Conference on Database and Expert Systems Applications (DEXA'05), Copenhagen, Denmark, 2005, p. 166 - 175.

[11] Hust, A., Klink, S., Junker, M., Dengel, A (2002). Query Expansion for Web Information Retrieval, Informatik bewegt: Informatik 2002 - 32. Jahrestagung der Gesellschaft für Informatik e.v. (GI), Dortmund, Germany. p.176-182.

[12] Messai, N., Devignes, M.D., Napoli, A., Smail-Tabbone, M (2006). Ingénierie des systèmes d'information: systèmes d'information spécialisés (ISI'06), V. 11, p. 39 - 60.

[13] Qiu,Y., Frei, H-P. (1993). Concept based query expansion, *In*:16th Annual International Conference on Research and Development in Information Retrieval (ISIGIR'93), New York, p. 160 - 169.

[14] Raymond, C., Bellot, P., El-Bèze, M (2002). Enrichissement de requêtes pour la recherche documentaire selon une classification non-supervisée, 13ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et d'Intelligence Artificielle (RFIA'02) p. 625 - 632.

[15] Schweighofer, E., Geist, A (2007). Legal query expansion using ontologies and relevance feedback", Legal Ontologies and Artificial Intelligence Techniques (LOAIT'07) p. 149 - 160.

[16] Tuominen, J., Kauppinen, T., Viljanen, K., Hyvönen, E (2009). Ontology-Based Query Expansion Widget for Information Retrieval, 5th Workshop on Scripting and Development for the Semantic Web (SFSW 2009), and 6th European Semantic Web Conference (ESWC 2009), Heraklion, Greece.

[17] Zayani, C.A., Peninou, A., Canut, Sedes, F (2006). An adaptation approach: query enrichment by user profile, *In*: The international conference on Signal Image Technology and Internet-Based Systems (SITIS'06) p. 24 - 35.

[18] Zhou, Q., Wang, C., Xiong, M., Wang, H., Yu, Y. (2007). SPARK: Adapting keyword query to semantic search, *In*: 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference (SITIS'06, ISWC/ASWC'07), p. 694 - 707.