Journal of Theoretical and Applied Information Technology
31st January 2014. Vol. 59 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SEMANTIC ENRICHMENT OF QUERIES WITH GENERIC AND SPECIFIC TERMS IN THE DEFINITION SENTENCES

[1]MOHAMED RACHDI, [2]EL HABIB BEN LAHMAR, [3]EL HOUSSINE LABRIJI
[1,2,3] Faculty of Computing Ben M'sik, Casablanca, Morocco
E-mail :[1] mohamed.rachdi@yahoo.fr ,[2]h.benlahmer@gmail.com,[3] labriji@yahoo.fr

ABSTRACT

Increasing the relevance of the results of research tools remains a real challenge for users and researchers in the field of information retrieval. To overcome this problem, several studies have been conducted, mainly in the methods and techniques that focus on the treatment and the reformulation of queries, to meet the needs of the users. The purpose of this paper is to contribute to the improvement of approaches to reformulate queries by semantic enrichment based on definition sentences.
Definitions and terminology have several features and components that can be used in information retrieval, especially in the enrichment of queries. The essential components of the definition sentences are generic and specific terms. Our approach consists of exploiting the definition sentences making use of their generic and specific terms.
Keywords: *Semantic Enrichment, Relevance, Queries, Generic, Spécific*

1. INTRODUCTION

An information search system must provide the user with the most relevant documents that respond to his needs that are expressed in the entered query. Before displaying the results to the user query, a search system has to go through several steps. Among these is the query processing step which is of paraamount importance as it has a major impact on the degree of relevance of the results returned to the user. One of the most challenges that are faced in this stage is query reformulation.

Several solutions are suggested to address this problem, such as the reformulation feedback, semantic enrichment by using a custom ontology, etc. This paper proposes a new solution using definition sentences in order to semantically enrich the user query. It exploits the richness of these sentences in terms of words and represents a remarkable similarity between them.

This approach is an extension of the work presented in [1] in that it exploits the results of the study with an upgrading in the components used. Definition sentences have two components: Generic and Specific. In this work, we use specific terms the definition sentences to increase relevance and reduce noise and silence of the returned results.

In this paper, we begin with a state of the art of the various approaches of query reformulation. Then, we move to the presentation of our approach. After that, we draw a comparison between our approach in query enrichment and that used in search engines in order to evaluate our system. Finally, the paper ends with a conclusion and outlook.

2. STATE OF THE ART

The main goal of an information retrieval system (IRS) is finding the correspondence between the needs of the user (the query) and the information contained in the resource. The system should provide the maximum number of documents relevant to the query. One of the most prominent features of modern IRS is enrichment. As mentioned earlier, this work is the continuation of earlier one in which we presented an approach for enrichment using generics with a bibliography that has tested the most popular enrichment approaches. The approaches used in the previous study are ontology profile, user and reinjection enrichment. In the literature, we find that some studies have used approaches that are based on external resources (WordNet Arabic and Arabic dictionary) for the enrichment of queries [2]. Other approaches are based on the definition sentences but they do not take into consideration the semantic proximity between their components [1] [3] [4].
A state of rich and detailed art is presented in [1], [3] and [4].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3. CONTRIBUTION

The purpose of an information retrieval system is to satisfy the user by selecting the relevant documents. This objective cannot be achieved in the absence of a good strategy of query reformulation and enhancement, which is being increasingly covered in the academia but not completely solved. There are two criteria used to determine the performance of an IRS: the research model used and the user query. The user uses personal terms that have nothing to do with the terms used to index the relevant documents that correspond to his query. So, it is necessary that the user chooses terms that are used in indexation, a scenario that is theoretically impossible.

To tackle this problem, we make use of query reformulation and expansion techniques. The idea is to formulate a query from the user's request in order to coordinate it with the indexing language. This aim of this process is to eliminate irrelevant documents. The major question that should be answered in this step is what terms should be added and how to integrate them.

The enrichment technique we propose in this paper lies at the heart of reformulation approaches. It allows the reformulation of the user query based on the specific components of definition sentences, namely Generic and Specific. A definition (according to Larousse dictionary) uncovers the set of the essential properties of a concept, word, object, etc. It establishes an equivalent relationship between a term (signifier) and meaning (signified). A definition provides a delimitation of the concept by describing its characteristics and relationship building between its different defining elements. It determines the position that the concept occupies in relation to other concepts in the field [5]. The definition makes it possible to situate the concept in a system and its hierarchy in a specific environment. A definition fulfills the following functions:

Describe, explain, explain and/ or define a concept, distinguish concepts from each other , recognize the defined, attest to the existence of a concept, a concept set, the link between linguistic unity, concept and referent structure or reflect a conceptual system, establish equivalence and synonymy between linguistic units, complete a didactic function and / or normalizing .

According to ISO 704 [2000] standard, the two main functions of a definition are to identify the concept and clearly differentiate it from other concepts.

The terminological definition consists of the domain corresponding to a certain way to an orientation that is attributed to the definition and initial definer. The latter represents one or more lexical elements that introduce the definition, used to situate the concept in relation to others in a conceptual system.

At this point we have defined the definition and its utilities the question that arises then is how to detect definition sentences. [6] have distinguished for types of markers in identifying definitional sentences:

- ✓ Metalinguistic markers
- ✓ Nominal metalinguistic markers
- ✓ Lexical markers
- ✓ Punctuation

Regarding the components of the definition, two parts are identified: the generic and specific [5]

The generic GEN is the element linking the set to a more general concept. The specific elements SPE (also called characters lines) are the ones that define the conceptual scope of the generic element and allow in particular the distinction of concepts from each other. The purpose of this work is to propose an enrichment approach to queries based on the GEN and SPE of definition sentences. The GEN and SPE are selected by using measures of association between the words that make up the definition sentences.

The general process of the approach is as follows:

Journal of Theoretical and Applied Information Technology
31st January 2014. Vol. 59 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

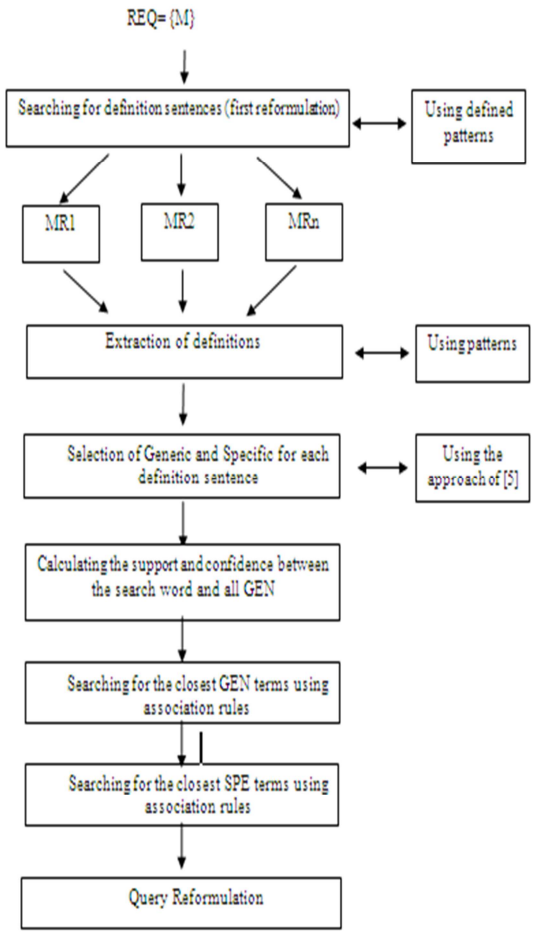

Figure 1: The General Process Of The Approach

3.1. Step 1: searching for definition sentences (first reformulation)

In this step, instead of launching the query as formulated by the user, we look for the possible definition sentences forms of the entered words.

To do this, upon the receipt of the query, we divide it into several words. Then, we proceed to reformulation by reinjection for each of the words. This type of reformulation consists of using markers used in [6] for the identification of definition sentences. Four types of markers, defining four types of patterns, are distinguished.

Metalinguistic markers to use independently (number 9): call, baptize, defined as, denominate, denote, designate, appoint, mean, want to say;

Nominal metalinguistic markers (11): naming, meaning, concept, designation, expression, Word, noun, concept, term, Word to associate with a supporting verb like: apply, give, employ, take, carry, receive, refer, reserve, use.

Lexical markers are not explicitly metalinguistic, or those of reformulation (21):, that is to say, in other words, either, namely, in some sorts, a kind of, finally, it is, by agreement, , mean, indicate, as said, for example, put differently, same as, equal to, used to, mark, explain, clarify;

Punctuation: parentheses, quotation marks and dashes are also mentioned in the literature. Because paradigms (help in horizontal ontology modeling) and hypernyms can also be extracted, these four schemes have been included in the evaluation of the method even if they are not exclusively targeted on the extraction of the definition.

In our case the first three markers are used to reformulate the initial query.

The received query will therefore have initial reformulation for each word. We get then several reformulated requests that will be sent to the search engine. The obtained result will be retrieved as a set of documents

Let us consider the initial query:

$$Q = \{M\} \quad (1)$$

Where MI is the i word of the query

On using the reformulation by reinjection of relevance, we obtain a set of queries with each consisting of a set of forms of the definition sentences for a word.

$Q0=\{Q1,\ldots Qp/\ Qp$ is the p sent query$\}$ (2)

where

Qi is i sent query,

As a result of this step, we obtain a set Ωi of elements for each word i of the query such as:

$$\Omega i = \{d1,..., dn\} \quad (3)$$

where n is the number of documents found

DJ is the j found document

3.2. Step 2: extraction of definitions

Every document dj found in a set Ωi contains a set of sentences. The next step is then to extract the definition sentences and the rest of the document will be ignored for the time being.

Let us consider ∑k the set of definition sentences for a Word

$$\sum k = \{pd1, \ldots\ldots, pdm\} \quad (4)$$

Where m is the total number of the sentences found

Pdm is the phrase m of the ∑k set.

3.3. Step 3: Selection of Generic and Specific for each definition sentence

To extract the generic and specific terminology from the definition sentences, an automatic identification approach is used [5]. This approach allows the identification and marking of the generic elements from the definition sentences

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

exploiting the formal characteristics of the definitional sub-language. The task is to identify the generic terms GEN and mark it with the following XML tag : <GEN> xxxx </ gen >. The issue that arises at this stage is how to detect it.

This author believes that the GEN often consists of a single word. Otherwise, it is of two words where one is a false intruder starting with a relational marker (the set of…). The difficulty that is faced is the identification of the right border of the GEN. For that matter, the author exploits the definitional speech patterns to identify these boundaries using morphosyntactic markers, which can then be translated into rules of identification in the form of a regular expression.
In the terminological definition, the GEN is always followed by a SPE. This brings to surface another which how to detect the right border of GEN (which corresponds to the end of the GEN and the beginning of SPE).

A detailed study of the main characteristics of GEN, the anteposed of SPE and the beginning of SPE coming immediately after GEN helped build a list of morphosyntactic markers [5]. This list is translated into rules of identification in the form of regular expressions.
For the identification of SPE, let us consider that the rest of the terms in the definition sentence are SPE since it consists of one GEN and one or more SPE.
Let us consider phdi the definition sentence i
At first it is of the following form phdi = { Mot1 , ..., Motn }

With the application of the detection approach of SPE and GEN, we get the following definition sentence phdi':
Phdi'={GEN,SPE1,SPE2,SPE3,SPE4,SPE5} (5)

As long as the terminological definition consists of a maximum of five SPE [5].

3.4. Step 4: Search for similar words using association rules

3.4.1. Searching for the closest GEN termes

An association rule between terms as defined by [7] is an implication of the form:
R : Ti➔Tj, where Ti,Tj ∈ T are termsets and Ti ∩ Tj = ∅

According to the same author, the validity of this association rule is evaluated using the two measures used in data mining, namely support and confidence.

$$Support(R) = \frac{||D\{TiUTj\}||}{||D||} \quad (6)$$

$$confidence(R) = \frac{\text{Support(TiUTj)}}{\text{Support(Ti)}} \quad (7)$$

In this step, we look for the possible links between the search word and all GEN so as to find the closest GEN (in other words, the domains that the search term belongs to)
Searching for the closest GEN is done with the help of support and confidence operations. We calculate the support and confidence between the search word and all GEN. After that, we select all the words whose support and confidence is greater than minsup and minconf. The main objective is to segment all the found GEN and select only those which are closer to the searched word.
Let us consider SupGENi the support of GENi
We look for the GEN terms that satisfy the following formula:

SupGENi >=minsup (8)

and

confGENi >=minconf (9)

The GEN terms that meet the criteria in the formula are then selected:
Therefore, LG list of the close GEN is built.

LG={GEN1,..GENk} (10)

3.4.2. Searching for the closest SPE terms

In a definition sentence, there is one GEN and several SPE (5). The SPE terms do not necessarily belong to the same domain of the GEN term as they may appear with different GEN in different definitions and therefore have a different degree of closeness with the GEN (support and confidence).

In this step, we look for relationships in each definition sentence between each GEN term and all the SPE ones. We calculate the support and confidence between each close GEN and all SPE terms which appears in the same definition. It is a process that allows us to know the groups of SPE that are closer to each close GEN. Search for SPE is accomplished through support (6) and confidence (7). We calculate the support and confidence between GEN and its SPE. We then select the SPE with a sup and conf exceeding minsup and minconf,
Let us consider SupSpei the support of SPEi
We look for the SPE terms that satisfy the following formula:

SupSpei>=minsup (11)

and

ConfSpei>=minconf (12)

The SPE terms that meet the criteria of the formula are then selected:
Therefore, LS list of the close SPE is built.

LS={SPE1,..SPEn} (13)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.5. Step 5: Query Reformulation

The objective of the above steps is to find semantically close words to the searched term. This proximity is measured using association rules, especially support and confidence. The use of these two values allows the selection of the close GEN and SPE from a found set of words. Thus, the results found are exploited in enriching the initial query.

The initial query Q typed in by the user will be enriched with the list of GEN (LG) and SPE (LS) found

The final query is therefore

$$Qf=\{M+LG+LS\} \quad (14)$$

4. EVALUATION

To evaluate our approach, we draw a comparison between the results with and without our enrichment approach. Queries are entered into the Google search engine. The following chart shows the result of irrelevant links:

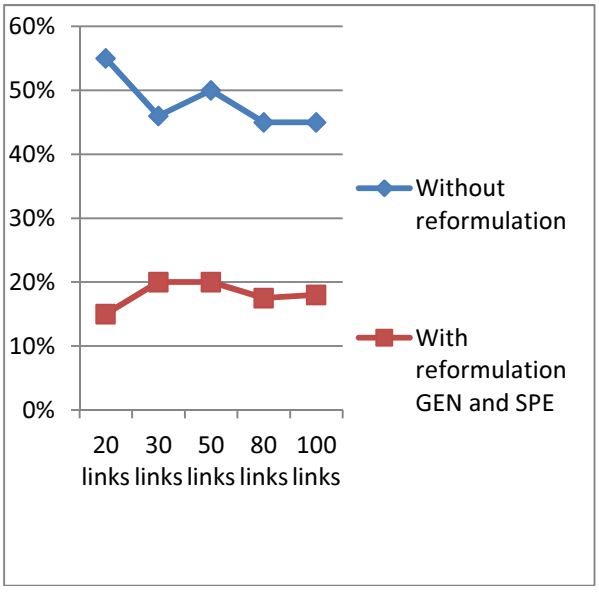

Figure2. The Rate Of Irrelevant Results With And Without Reformulation

From this graph, we see that the rate of non-relevant results without reformulation is 55% for the first 20 links while it is only 15% with the implementation our query reformulation approach. The same is true if continue looking into other links, such as 100 where we see that 45% of the results are irrelevant without reformulation while just 18% of the results are not relevant using our approach . Knowing that the majority of users look just at the first few links, our approach represents a significant advantage over sending the query without enrichment.

5. CONCLUSION AND OUTLOOK

This work is part of the development of an information retrieval system for expanding the user query with terms extracted from definition sentences. To enrich the query, we have used the concept of GEN and SPE, extracted by using association rules.

The purpose of this approach is to provide higher accuracy and reduce at maximum the noise in the returned results. The feasibility of the method was tested by running a simulation in the Google search engine. We have shown that our approach represents a contribution to the query enrichment step in information retrieval systems.

Despite the added value of the approach, an improvement is estimated to take place if we could consider the enrichment of queries consisting of multiple words. Also, using the GEN and SPE in the query reformulation taking into account the relationships that may exist between all the words will yield better results.

Our approach has shown good results in the evaluation part. However, the level of noise in the returned results started to increase when we go beyond 100 links, an issue to be considered in future studies.

REFERENCES

[1] Mohamed RACHDI, El Habib Benlahmar, EL Hassan Labriji Enrichissement des requêtes à l'aide des Génériques. La journée Sciences de l'Ingénieur. Faculté des sciences ben m'sik, Casablanca, Maroc. Juillet 2013

[2] Mohammed El Amine Abderrahim: Utilisation des ressources externes pour la reformulation des requêtes dans un système de recherche d'information. Prague Bull. Math. Linguistics 99: 87-100 (2013)

[3] Mohamed Rachdi, H. Ben Lahmer, H. Labriji(2011) . « Semantic enrichment of queries in search engines». Extraction et Gestion des Connaissances (EGC-M 2011) Novembre 2011,Tanger (Maroc)

[4] Mohamed Rachdi, H. Ben Lahmer, H. Labriji (2012). « Extentional Approach of Semantic Enrichment of Queries in Search Engines». International Journal of Information Technology & Computer Science.Vol 6,Pages 33-41, November / December 2012

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[5] Selja Seppälä (2007). « La définition en terminologie : typologies et critères définitoires ». TOTh 2007 : Terminologie et Ontologie : Théories et Applications., Anncey : France (2007)

[6] Véronique et Al (2004). Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologie.In Philippe Blache, editor, Proceedings of TALN 2004 (Traitement automatique des langues naturelles), pages 269-278, Fès, Maroc, April 2004. ATALA, LPL.

[7] Bsiri et Al 2003 : S. Bsiri, H. M. Zargayouna, Cherif Chiraz Latiri, Sadok Ben Yahia: Découverte de Règles Associatives Hiérarchiques entre Termes. INFORSID 2003: 333-347