Approach to a conceptual model of indexation in a weak ontology described in RDFS

Traoré Issa[1], Souleymane Oumtanaga[1], Babri Michel[1] and Claude Lischou[2]

**[1] Laboratory for Informatics and Telecommunications Research (LARIT), INPHB,
08 BP 475 Abidjan 08 (225), Cote d'Ivoire.**

[2]University Cheikh Anta Diop (UCAD), BP 5353 Dakar-Fann-Senegal, Dakar (205), Senegal.

Abstract

In this paper, we describe a model of conceptual indexation adapted to the RDFS structured resources. The process of the structured IR (Information Research) in XML corpus, based on the balancing of a document terms improves the IR. However, silence in the resources return remains a concern. Operating a weak ontology described by RDFS can improve the percentage of relevant return of documents. The use of semantic relationship between pairs of synonymous modeled concept by graphs enables to take into account the meanings of the terms in a RDFS corpus. This model allows us to consider some types of structural information by exploiting the semantic properties, particularly the synonyms and abbreviations. What considerably improves the performance of the SIR (Systems of Information Research). Experiments on a collection of documents were used to evaluate our model.

***Keywords**: structured IR, SIR, ontology, RDFS, BM25, synonyms.*

1. Introduction

The explosion of data volume and the improvement of the storage capacity of databases were not supported by the development of analytical tools and the IR to exploit this mass of information. The achievement of intelligent research systems has become a matter of urgency or even a necessity. Research methods based on the key-words are not accurate enough for the description of the text content. Given the increasing number of documents on the web, several models of SIR are created. Indeed, the main issue of the SIR is to find the ten or thousand odd relevant documents among millions of documents. This utmost makes this task even more difficult. There are some classical models of information research such as the Boolean model [18], the matching score [21], the vector model [28], [5], the probabilistic model [6], [22], and so on. These models are based on the IR of structured or non-structured documents. The main drawback of these models is that they generate a significant silence [1], [27]. Indeed, two words with different syntaxes may have the same meaning. For instance: "father" and "dad" or "web 3.0" and "semantic Web". A user seeking information on a "semantic Web" will be unlucky to be informed of the "Web 3.0".

The structure and the content processing are the main issues of the structured information research. It is essential to obtain highly relevant scores in a graduated scale. In the purpose of increasing the research models, various strategies are implemented so as to be grafted. These strategies use various sources of evidence: semantic relationship defined in the thesaurus , classes and contextual use of the concepts, research results, relevant judgments of the users, element of the information theory, heuristics [3], [16], and [14].

The semantic relations provides a great advantage in IR, it allows exploiting the meaning of the concepts in the ontology, these latter being constructed with structured languages such as the XML (Extensible Markup Language). Several works have been done in IR with the format XML documents. The well-known are the works of TREC [1], [4] and [12], INEX [27], [13], then project WorldNet [7], [29] and MESH [31], they are ontologies to facilitate IR [26] and [33], by using the similarity distance between concepts. However, XML documents cannot efficiently operate on the meaning of terms.

Concerning the approach presented in this paper, the consideration of semantic in a weak ontology is exploited. These resources are described in the RDFS language (Resource Description Framework Schema). RDF is a markup language which has the advantage of describing objects. It was standardized in 1999 (by a W3C recommendation) [10]. It enables human beings to read multiple metadata schemes as well as their analysis by machines. It uses XML to express a structure allowing the metadata communities to define the real semantics. The RDFS is an extendable language knowledge representation

[11]. The RDFS resource is characterized by a content (of the text) and a structure (tags). However such documents cannot be efficiently exploited by the classical methods of IR. Indeed these latter treat a document only in terms of its content (key words), whereas XML and RDF are used to add structural constraints (tags). From time to time this requires to adapt the classical methods of IR or introduce new mechanisms in order to well exploit the meaning of the terms of the document.

This paper is organized as follows: in addition to the introduction, the first part develops the problematic and gives a state of the art XML documents in IR. The first contribution of this paper is the suggestion of a formal operational, semantics in a light ontology described in RDFS, presented in the second section. Then, considering explicitly the semantic of the terms of documents, we calculate the relevance score of a term paper and the document containing it. In part 4, we evaluate our model by an experiment on its corpus.

2. State of the art.

2.1. problematic

The basic question is: how must the information of the structured XLM or RDFS documents be indexed in order to respond efficiently to a request?

By focusing on the structure, we can identify some specific issues related to XML. It doesn't consider the semantics in terms of the documents [27]. Indeed, following a request from a user, only documents containing the query terms will be ranked and then returned. The problem is then: how to define a formalism which will reduce the silence by returning the documents containing the demanded requests and their synonyms? Can we combine the silence reduction and the improvement of the documents returned by the system? It's all these questions we will try to answer in this article.

2.2. Some research work

The research of a document on the web is based on the key words. In fact, the relevance of a web resource for a given query depends on one hand, on the weight of the query words, and on the frequency of the occurrence of words on the other hand.

Generally speaking, the web resources are not structured, the most existing used methods in IR are: Boolean method, vector method, Page Rank method, BM25, and so on.

The IR process role is to establish a correspondence between the relevant information sought by the user, usually represented through a query, and the set of available documents. It is based on two essential steps: the indexing phase and the research phase.

The indexing phase focuses on the documents and requests analysis in order to create a representation of their textual content which should be usable by SIR. Each document (and research) is then associated with a descriptor, represented by a set of the extracted indexing terms.

The research phase aims at connecting the indexing of the documents with the user's request. It is based on a precise formality defined by a model of IR. The documents presented as a result to the user, and considered as the more relevant, are the ones which indexing terms are the closest to those of the request.

Many works have been done in the finding structured or semi-structured resources [3], [17]. XML documents are used in the background (words) and forms (tags). These principles have already been used separately in research INEX [14] and [11] and in other research work using the balancing of the tags [8]. Therefore, the XML research works done have sought not only to identify some more concise information units but also to use these markers to detect the information in a more appropriate response to a need of information. As a result, the relevance of a resource (Fig.1) not only depends on the frequency with which the request terms appear, but also on the weight of the tags which mark these words in the document.

Generally speaking, any SIR has two basic purposes: find all the relevant documents, and reject all the irrelevant ones. These purposes are evaluated by the measures of recall and precision.

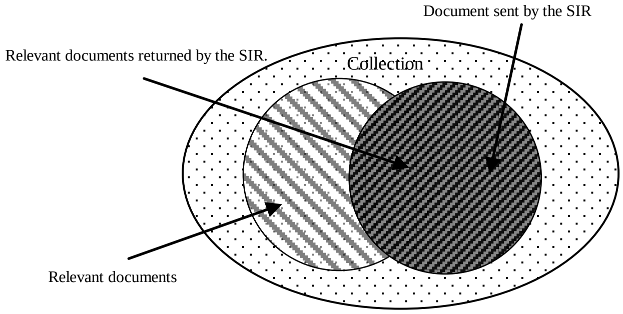

Fig. 1 Precision and Recall

In works [15] and [12], the tags judged important as well as their weights are selected manually. For instance in work [12], the weight of the "title" tag is set to 2 and the one of the "abstract" tag to 1.5. Another method in [9], consists in automatically taking these weights. For instance [12] and [2] use the genetic algorithm, whereas [32] uses techniques based on simulated annealing. In the case of a set of structured documents, the weight of resources is obtained by combining the weight of words and tags.

Let N_r be the number of documents re turned by the system for a given query, N_p, the number of relevant

documents in the collection for this query and N_{pr}, the number of relevant documents returned by the system . Precision and recall are given by the following formulas:
Precision measures the proportion of the relevant documents restored by the system. It is expressed by:

$$\text{Precision} = \frac{N_{pr}}{N_r} \qquad (1)$$

The recall measures the proportion of relevant documents restored by the system related to the set of relevant documents in the basic information. It is expressed by:

$$\text{Recall} = \frac{N_{pr}}{N_p} \qquad (2)$$

By these formulas, silence and noise rates are given by the following formulas:

$$\text{Silence} = 1 - \text{Recall} \quad \text{and} \quad \text{Noise} = 1 - \text{precision} \qquad (3)$$

In the case of an ideal system, the precision rate is equal to the recall rate that is all the relevant documents are selected.
In the model we present, the structure and the meaning of the documents are operated at two levels:

- Logical structure and layout: the tags are used to determine the granularity of the index, so the granularity of the elements the system will likely return. The weight of each of the tags is expressed by learning.
- Formalization of the semantics: the terms with the same meaning are linked by a relation R. they are measured weighted in accordance with the R relation.

At the query step, the probability for a resource to be relevant is estimated by combining the weights of the words it contains with the weight of the tags which label them.
In the next section, we present a light ontology described in RDFS, then we will find the formality enabling to consider the meaning of the terms by the SIR.

3. The ontology built with lightweight RDFS

3.1. The RDFS language

Most of the standardized languages by the W3C as parts of the semantic web are WML dialects; this is the case of RDF and RDFS we are going to deal with in this article. RDFS provides some basic elements for defining ontologies and vocabularies for structuring the RDF resources [10]. The language SPARQL (SPARQL Protocol and RDF Query language) is a query language for the RDF. Just like SQL for relational databases or XQuery for XML documents, this language is used to extract information from the RDF documents.

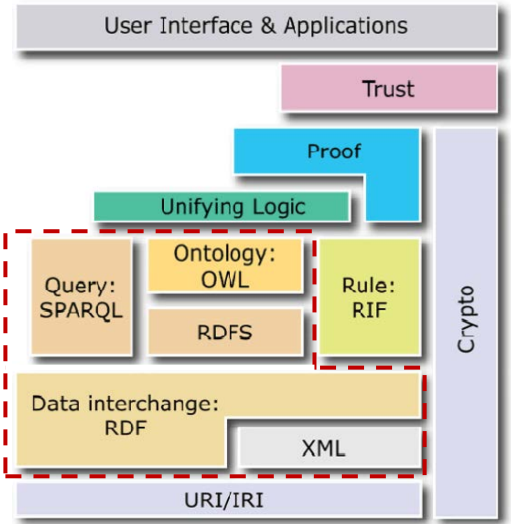

Fig. 2 Architecture of the semantic web

More specifically, the semantic web is based on an architecture called the Semantic Web Stack (see Fig.2). It essentially deals with an illustration of the languages hierarchy (XML, RDF, RDFS, OWL, and so on) used in the semantic web. It's important to notice that all these technologies have been standardized in the last decades by the W3C consortium to permit the passage across the semantic web. We can also notice that from the bottom of the stack down to this ontology, all the documents have been accepted and standardized.
The ontology construction requires a consensus in order to avoid the lexical disambiguation, due to hyperonymies and polysemies. The light ontology constructed in RDFS is a graph in which inferences can be exploited.
RDFS descriptions are presented as oriented graph (see Fig.3 and Fig.4). There are various notations of RDFS graphs. The simplest one is the N-Triples, according to which the RDFS graph is represented by a collection of triplets [11] with the following abbreviated form:

ex : Patient rdf : SubClassOf ex : Person

This graph means that: “a patient” is described as a “sub class” of “persons”.

3.2. The light weight ontology in RDFS semantics

RDFS extends RDF language to describe more precisely the resources used to label the graph. For that, it provides a mechanism to specify the classes whose resources will be instances like properties. The use of this set of theories to describe these models has two interests: the generic of set notion based on mathematical basis and its universality that is the semantic domain is for experts in this field.

RDFS resource can also be translated into a formula of the positive logic (without negation), conjunctive, and existential of the first order (without functional symbols), which models are the same with those defined by the direct semantic in the theory of models (see [19]). However, the extension to RDFS only provides primitive mechanism for specifying these classes.

In our approach, RDFS is enough to warrant the construction of a lightweight ontology describing some synonymous concepts. In this work, the literals will have a particular importance since they contain words whose scores are used by the SIR.

Let's consider the two RDFS below graphs representing the resources describing some diseases in the medical field.

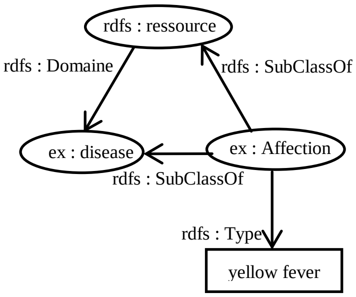

Fig.3 Resource 1

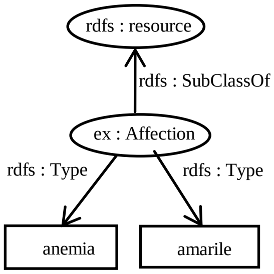

Fig.4 Resource 2

The abbreviated N-triple entry of the two graphs, an inference on the concepts see [29] and [30]. So the two examples bellow:

ex: Affection rdf :type "Yellow Fever"
ex: Affection rdf: type "Amarile''

Let's suppose that these two RDFS graphs in are from the same ontology of a medical research. At the request of a user, who seeks documentation of Yellow fever, document n° 1 will be returned because it has a literal containing the term "Yellow fever". Document 2 will not be returned because it no longer contains the term "Yellow fever". Our approach consists in reducing the silence during the return of the relevant documents by the SIR. Knowing that " Amarile" stands for "Yellow fever" the following equivalence can be written:

IF rdf : hasSynonym rdf : type rdf : SymmetricProperty
AND ex : Amarile rdfs : hasSynonym ex: yellow_fever
THEN ex : yellow_fever rdfs : hasSynonym ex : Amarile

Thus, according to this algorithm, it is possible to infer the terms of a document. Although the vocabulary with which to build ontology in RDFS is limited, it is possible to infer the concepts [19] and [20], and formally justify that two concepts are equivalent. The query language SPARQL is the RDFS whose syntax is:

```
SELECT ? Subject ? Object
       WHERE {
              ? subject rdfs: subClassOf? object.
               }
```

4. The database of synonyms

4.1. Properties of the relation of synonymy.

The database of synonyms will be composed of a set of couple of words. These couples will be composed of words and their synonyms or their abbreviations such as: (Web_3.0, Web semantic),
(Amarile,Yellow fever) or even (PC, Personal Computer), and so on. In some situations, a term may have various synonyms, so the exploitation of the synonym relationship noted *syn ( )* will solve this problem. In our approach we use the pure synonymy defined by:

Two lexical units are in a pure synonym relation if any occurrence of one can be replaced by the occurrence of the other in any environment without changing significantly the meaning of the statement in which it is. Let S be the set of synonym terms, $\forall$ A, B $\in$ S *syn* (A, B) means that the term B is the synonym of term A. The synonym relation thus defined checks if:

reflexivity : $syn(A,A) \ \forall \ A \in S$;

symetry : $syn(A,B) \Rightarrow syn(B,A) \ \forall \ A, B \in S$;

transitivity: $syn(A,B), syn(B,C) \Rightarrow syn(A,C) \ \forall \ A,B, C \in S$.

So the relation of pure synonym is a relation of equivalence. It allows the construction of classes of lexical units taking into account all the word which are ynonyms. These synonyms will be grouped in the database of synonyms.

The synonym relation enables to select in the database of synonym all the synonyms of a term. When the system receives an event (query), it searches simultaneously the "key" terms in the database of synonym. As soon as a term is found, then the other element of the couple is considered as the synonym of the term query. Thus, the synonym terms enrich the query in order to take into account the semantics.

The particularity of our approach is that it takes into account the various synonyms of the query key terms, so the database of synonyms will have a great importance. For the enrichment, two basic methods are considered:

- First, it is manually created by experts is this domaine, preferably those that will create the knowledge base (see Fig.5).
- Then a learning method that allows the automatic addition of new synonyms for user input. This can offer synonyms for certain query terms, which will be subject to registered and subsequently used by the SIR.

It is important to have a minimum consensus on the definitions of the field in order to avoid problems of polysemy and hypernyms. Since the performance of the SIR depends on the one hand, on the response time of the system and also the quality of answers that is to say, the relevance of the documents that will be returned.

Definition 1: Let I and J be two sets of the finite integers, $\rho_i = \bigcup_{i \in I} < s_i, P_i, o_i >$ is a collection of triplets of a resource, and $O = \bigcup_{k \in J} \rho_{k_i}$, a set of all the resources describing an ontology of documents in RDFS field.

Remark 1: The o_i objects are subjects or literals. In our approach we look for terms in the literals. As a result, the literals play an important role in our approach to SIR.
There, $\rho_1 = <s_1, P_1, o_1 = \{4t_1, 2t_2, t_3, t_4\}>$, a triplet of a resource which object is the literal with terms. The literal shows that the resource contains four times term t_1, twice term t_2, and once terms t_3 and t_4.Without losing general information, we will simplify the writing of the resources by the following:

$$\nabla_1 = \{4t_1, 2t_2, t_3, t_4\}$$

To show that document ∇_1 contains four times term t_1, twice term t_2, and once the terms t_3 and t_4.

Definition 2: Let I ∈ {1, 2, …, n} and J ∈ {1, 2, …, m}. $\rho_j = \bigcup_{i \in I} < s_i, P_{i,}, o_i >$ be a document in the format of a collection of RDFS triplet. We call representative of ρ_j the set ∇_j of all the terms of the documents.

4.2. Approach of the system model of information research.

The search of relevant information which responds to the need of a user consists in corresponding the representation of the information contained in a document collection with those of the user's need. In the figure below (Fig.5), we present an architecture of our SIR model in the form of the "U" process. This process manipulates some documents collection, a database of synonyms and some user's needs for search from a collection of documents, those which best meet a user's need.

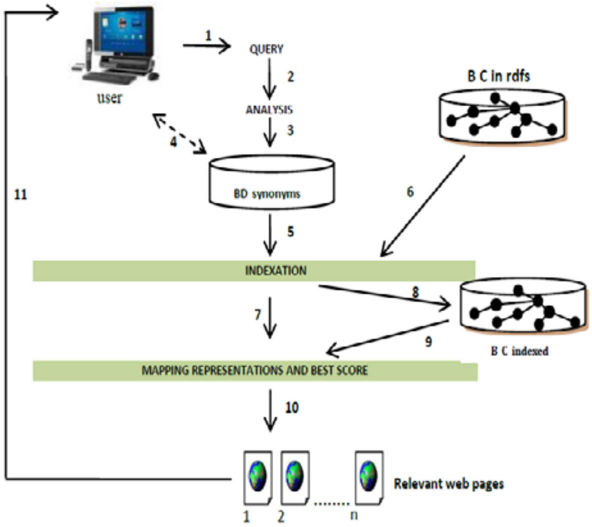

Fig.5. Conceptual architecture of the SIR.

The process of SIR is shown in Fig.5: (1) a user formulates his information need as a query, (2) this query is analyzed and then put in a RDFS format (triplet <s, P, o>). (3) Then the obtained description enables to look for synonyms in a database of synonyms, which enriches the query with the terms of the same meaning. (4) The enrichment of the synonym base by experts in the field or by users. (5) After identifying the synonyms, the indexing is done. At the same time or previously, (6) the RDFS collection documents in the knowledge base is also indexed. Thinks to the index (indexed collection and query), (8) the system can build the representations, then (7 and 9) mapping of the representation of the query with the representations of the document in the collection. (10) The calculation of the documents scores being done, (11) it returns a list of documents considered by the IR engine as relevant compared with the user request.
It is important to notice that in our model, the SIR proceeds to the removal of the tool-words (stop words). It deals with suppression the words of the common language which do not contain more semantic information (example: "a", "the", "of").

In the following paragraph, an evaluation of our approach allows checking the influence and advantage of the consideration of the meaning of documents.

4.3. A probabilistic approach of the resources representation.

The relevance of document ∇_j relatively to a query Q depends on the weight of the terms which appear in the document and in the request. We note w_{ji} the weight of term t_i in the document ∇_j. Let suppose that the weight of the t_i term in Q equals to 1 (the t_i term appears once in the query).

We define as X_j a vector of random variables and, $x_j = (x_{j1}, ..., x_{ji}, ..., x_{jn})$ a realization of this vector X_j, with $x_{ji} = 1$ (resp .0). If the term: t_i appears (resp. doesn't appear) in the document ∇_j.

Given these notations, let's consider π as the relevance of X_j based on the weights of the terms and U the universe of representative terms collection. π is given by the score :

$$\pi(x_j) = \sum_{t_i \in U \cap Q} x_{ji} \times w_{ji} \tag{4}$$

The BM25 weighting scheme (BM stands for "Best Match") which has been developed by [22], and then BM25 OKAPI in [28] and [23], is a weighting scheme based on the probabilistic model. They use the probability distribution of 2-Poisson.

Several versions of this formula have been suggested, and BM25 is a compromise on the nature and the length of the documents and the queries. This formula has been set on the corpus of the TREC competition. The apparent complexity of the formula (compared to the vector model), and its high efficiency (this formula is currently one of the most successful) well shows the advantages of the probabilistic model. Below is the BM25 formula which enables to calculate the weight w_{ji} of a term t_i in a document ∇_j.

$\forall\ k_1, b \in \mathbb{R}$

$$w_{ji} = \frac{tf_{ji}(k_1+1)}{tf_{ji} + k_1 \times ((1-b) + b \times ndl} \times \log\frac{(N - dl_i + 0{,}5)}{(dl_i + 0{,}5)} \tag{5}$$

With:

tf_{ji} : the frequency of t_i in the document ∇_j.

- N : the number of documents in the collection
- dl_i : the number of documents with the term t_i
- $ndl = \dfrac{dl_i}{avg - dl_i}$: the ratio between the size of the document and the avg (average) size of the elements (in terms number)
- k_1 and b : the classical parameters of BM25

Parameter k_i permits to adjust the overloading of tf_{ji} , and parameter b is used to adjust the focus on ndl, that is the importance of standardization of the items size. We will take k_j=2 and b=0,75.

Let's notice that we will choose k_j and b so as BM25 will be a non-linear function compared to the terms frequency as stipulated by [9].

The weight of the t_i term of the request, called w_{Qi} takes into account the appearance number of term t_i in the request tf_{Qi}, and a parameter k_2, of value equals to 8 (found in experiments).

$$w_{Qi} = \frac{tf_{Qi} \times (k_2 + 1)}{k_2 \times tf_{Qi}} \tag{6}$$

Let $\nabla_1 = \{4t_1, 2t_2, t_3, t_4\}$ be the representation of a document ∇_1 and Q = <s ; P; { t_1, t_2}> a user request. We consider $\nabla_Q = \{t_1, t_2\}$ as a representative of the request.

In our approach, we also consider the meaning of the query terms. Let t_i be a term representing ∇_Q t^*_{ik} the t_k term representing ∇_j as the synonym of t_i. The SIR will look for the concepts t_i and $t^*_{2,4}$ (*syn*(t_i, t^*_{ik})) in the different documents. Indexing will be done with the BM25 formula with a change in the calculation of the tf_{ji} frequencies of the t_i term in a representative ∇_j.

$$tf_{ji} = \frac{n_i + \sum_{t_k \in \nabla_j} n^*_{ik}}{\left|\nabla_j\right|} \tag{7}$$

With:

- n_i : the number of terms in the representative ∇_j ;
- n^*_{ik} : The number of synonyms of the t_i term in the representative ∇_j ;
- $\left|\nabla_j\right|$: The sum of all the terms of the representative ∇_j.

So, tf_{ji} is the arithmetical average of the number of words having the same meaning in the resource ∇_j. We can deduce from it the following remark:

Remark 2: Let's focus on n^*_{ik} , the number of synonyms of the term t_i in the representative ∇_j .

- When $n^*_{ik} = 0$, we have the formula for BM25.

- When $n^{*}_{ik} \neq 0$, the representatives with t_i and t_k as terms will have their scores improved.
- The representatives with t_k term and not the t_i term will no longer have their nil score.

5. Evaluation of our approach

5.1. Score concepts

To evaluate our approach, we'll consider the
universe representatives collection terms:
U = $\{t_1, t_2, t_3, t_4, t_5, t_6\}$. In this universe it is assumed that t_2 and t_4 have the same meaning doing . Learning will allow the system to identify the literal meaning like the same. The query representative $\nabla Q = \{t_1, t_2\}$.
Let the following representatives be ones of a corpus of 12 documents in RDFS:

$$\nabla_1 = \{t_2, 5t_3, 7\,t^{*}_{2,4}, 2t_5\},\ \nabla_2 = \{5t_1, 3t_2, 7t_5, 2t_6\},$$

$$\nabla_3 = \{t_1, t_2, 3\,t^{*}_{2,4}, 2t_5, 5t_6\},\ \nabla_4 = \{2t_3, 8\,t^{*}_{2,4}, 2t_5, 8t_6\},$$

$$\nabla_5 = \{10t_2, t_3, 2t_5\},\ \nabla_6 = \{2t_1, 7t_3, t^{*}_{2,4}, t_6\},$$

$$\nabla_7 = \{5t_1, 3t_3, 2t_5\},\ \ \nabla_8 = \{t_1, t_2, 4t_5\},$$

$$\nabla_9 = \{t_1, t_2, t_3, t^{*}_{2,4}, t_5, t_6\},\ \nabla_{10} = \{t_1, 4t_5, t_6\},$$

$$\nabla_{11} = \{3t_1, 5\,t^{*}_{2,4}\} \text{ et } \ \nabla_{12} = \{2t_2, 7\,t^{*}_{2,4}, 3t_6\}.$$

The calculation of the representative scores terms lead to this document, we have the graphs below for the query Q. This score gives the first performance of the various documents by the BM25 method. From Fig.6, the relevant documents are respectively: ∇_2, ∇_9, ∇_8, and ∇_3.

In the same figure, the synonyms of query terms Q are taken into account. Indeed, terms t_2 and t_4 have the same meaning with the document containing them, they are calculated accordingly. We obtain the following graph below:

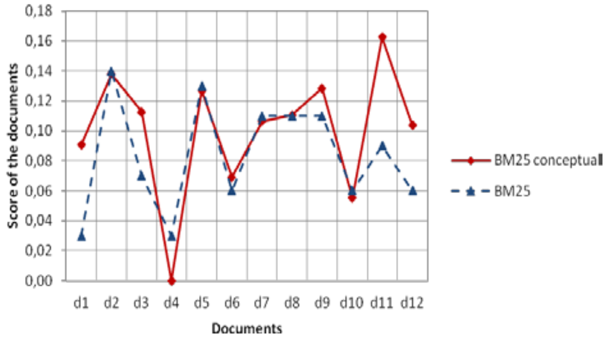

Fig.6 Score of the documents according to BM25 method and our approach.

The observation of this curve suggests that some documents with nil scores now have good scores. The relevant documents are in the following order: ∇_{11}, ∇_2, ∇_9, ∇_3, ∇_8 et ∇_6.

The consideration of the meaning of the terms increases the number of relevant documents and reduce the silence.

5.2. Probability of relevance of a concept

The SIR evaluation is a main issue on which the IR community has worked a lot. INEX (Initiative of the evaluation of XML retrieval), TREC (Text Retrieval Conference) and CLEF (Cross Language Evaluation Forum) designed for flat documents (see [4], [12], [24] and [25]). These evaluation campaigns of the SIR enabled a comparison as strict as possible of the systems of information research in the XML collections-oriented documents.
The return of the documents as an answer to a user need is based on the "Probability Ranking Principle (PRP)", stimulating a better performance of the system when the documents are returned in a decreasing number of their relevance.
In our approach, we will use the evaluation methods advised by the INEX, in which measures based on the recall-precision curves are widely used.

The probability for a document ∇ to be relevant for a given query Q , for measuring the performance of the SIR. For this purpose, we have two events:

- R , the event, the document is relevant for Q;
- NR, the event, the document is not relevant for the query Q.

The recall also measures the probability for document ∇_k to be selected, knowing that it is relevant:

$$\text{Recall l} = P(t_{ik}=1|R) \qquad (8)$$

The precision also measures the probability for a document ∇_k to be pertinent knowing that it is selected:

$$\text{Precision} = P(R|\, t_{ik}=1) \qquad (9)$$

In practice, we must estimate the probability p_{ik} and q_{ik} , $i \in \{1, .., n\}$, $k \in \{1, .., m\}$ where:

- p_{ik} : the probability for t_i to appear in the document ∇_k, knowing that a relevant document.
- q_{ik} : the probability for t_i to appear in the document ∇_k, knowing that a non relevant document.

To make the evaluation of the ∇_k documents possible, p_{ik} and q_{ik} must be estimated on a set of pre-definite queries.

Given the set R (resp. NR) which has the relevant documents, and (resp. irrelevant), (see Table 1) the contingency can be constructed for each t_i term of the document we have . Given the consideration of the synonyms, we have the equalities below:
We write:

$\mu_{ik} = \sum_{i \in I} n^*_{ik}$ the sum of the various synonyms of t_i in the document ∇_k .

$$p_{ik} = \frac{r_{ik}}{R} \quad \text{and} \quad 1 - p_{ik} = \frac{R - r_{ik}}{R} \qquad (10)$$

$$q_{ik} = \frac{n_i + \mu_{ik} - r_{ki}}{N - R} \qquad (11)$$

$$\text{and} \quad 1 - q_{ik} = \frac{N - R - n_i - \mu_{ik} + r_{ik}}{N - R} \qquad (12)$$

Hence the following table.

Table 1. Under the contingency documents

	Relevant (R)	Non relevant (NR)	Total
$t_i \in D_k$	r_{ik}	$n_{i+}\mu_{ik}$ $-r_{ik}$	$n_i+\mu_{ik}$
$t_i \notin D_k$	R-r_{ik}	N-n_i-μ_{ik}-R+r_{ik}	N-n_i+ μ_{ik}
Total	R	N-R	N

Table 1 enables to construct the recall/precision curve. The recall measures the capability of the system to find all the relevant documents and the precision measures its ability to find only relevant documents. The measure of the silence evaluation is a complementary notion to the recall, it is defined by;

$$S = 1 - \text{Recall and } S = 1 - \frac{r_{ik}}{R} \qquad (13)$$

- Where S is the silence.

This measure is in the interval [0,1]. It also evolves toward to 0 (for r_{ik} evolves toward to R) for the best SIR. To examine efficiently the results, we calculate the pair of measures (recall rate, precision rate) for each returned document. Table 1 illustrates some calculations of precision and recall for the first ten documents returned by a system for a different query, for which the collection contains 4 relevant documents (in first BM25model) and 5 relevant documents (in our approach, BM25 conceptual). The associated recall-precision graphs are drawn on Fig.7. As we can notice it in Table 1, several precision values can correspond to the same point of each recall. In order to have curves which are easy to read, we generally present the interpolated precision calculated at each point of the recall (which means that each relevant document is returned).
The perfect system will only find the relevant documents with a precision and a recall of 100%. In practice, the precision and recall measures increase inversely, which means that the interpolated curve according to the recall decreases. The higher the curve is, more performing is the system. We consider the 10 first referred documents.

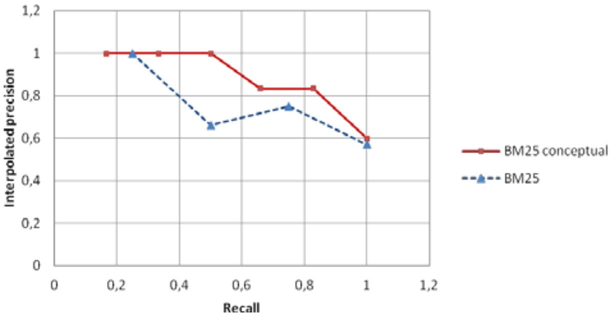

Fig.7 Curve of Recall / Precision

By comparing the rates of precision and recall of the suggested collection, we can notice that by integrating the meaning of the terms in our indexing system, we have better performances compared to the best results obtained by using BM25. In fact, the integration of semantics into the indexing system has shown a real influence on the performances of a SIR.

6. Conclusion

The web is considered as the first source of information all over the world, and the search for relevant information is viewed as one of the new needs of the information society. The interest in consulting this media is linked to the efficiency of the information research engines.
In this article, we have presented a new approach of the consideration of the RDFS structure for IR. We not only consider the structure so as to define the types of RDFS elements indexed by the system, but also the semantic of the field concepts. In fact, a consideration of the semantic permits to diminish the silence in the system while increasing the recall. It allows improving the precision of the SIR through the addition of a data base of the synonym terms.
During the questioning phase, the calculation of term relevance for a query is a combination of the weight of the required terms with their synonyms in the resource.
The main contribution of this work consists in modeling the capability of the system to identify the synonym terms and highlighting them, in accordance with the reviewed probability BM25 model. In doing so, the regulation of the score of the documents is done almost automatically.

The enrichment of the query enables to understand the research well and have results coming from the research of the synonym terms.

The second contribution of this work is the experience of the BM25 models in the context of the IR conceptual described in the RDFS language.

First of all, we have evaluated our model with a corpus of 12 documents. Our experiment consists in comparing the classical BM25 model and the BM25 adapted to the conceptual model.

Our approach gives the best results compared to the recall and the precision. Therefore it is more performing in a particular context and it will be interesting to take great advantages of the documents it offers.

In addition, the positive results of our SIR model pave the way for interesting prospects as the presentation of the result to a user is concerned. Our model brings a partial solution to the problems arisen into 2.1. This model could be improved by optimizing b and k parameters b and k_i of the BM25 formula of our conceptual approach, or to use the weight of the markup and the weight of the subjects, predicates in triplets < s, P, o > of RDFS. Then, it will be interesting to automate the construction process of the database of synonyms entirely.

The research system based on a conceptual representation of documents and queries are promising since they go from the symbol level ("characters chain ") to the conceptual one. What will enable them to raise (or slash) the morphological constraint of the synonymy and the polysemy which are known since a long time in IR as generative of silence and noise.

References

[1] G. Amati, E. Ambrosi, M. Bianchi, C. Gaibisso, and G. Gambosi. FUB, IASI-CNR and university of Tor Vergata at trec 2007 blog track. In Proc. TREC, volume Special Publication 500-274. 2007.

[2] Boyan J., Freitag D., Joachims T.. A Machine Learning Architecture for Optimizing Web Search Engines. AAAI Workshop on Internet-Based Info. Systems, 1996.

[3] Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic dynamics and collaborative tagging, Proceedings of the National Academy of Sciences (PNAS), 104, 1461–1464, 2007.

[4] C. Clarke, N. Craswell, and I. Soboro_. Preliminary report on the TREC 2009 Web track. 2009.

[5] E. Fox. Extending the Boolean and Vector Space Models of Information Retrieval with P-Norm Queries and Multiple Concept Types. PhD thesis, Cornell University,University Microfilms, Ann Arbor, Michagan, 1983.

[6] Fuhr, N.. Probabilistic models in information retrieval. The Computer Journal 35, 3, 243–255. 1992.

[7] George A. Miller. WordNet: a lexical database for English. Commun. ACM, 38:39–41, November 1995.

[8] Géry M., Largeron C., Thollard F.. Integrating structure in the probabilistic model for Information Retrieval, Web Intelligenc. p. 763-769, 2008.

[9] Géry M., Largeron C., Thollard F., UJM at INEX 2008 : pre-impacting of tags weights , Proc. of INitiative for the Evaluation of XML Retrieval (INEX), Dagstuhl. 2009.

[10] http://www.w3.org/TR/rdf-schema/Langage de description de vocabulaire RDF 1.0 : RDF Schema. Recommandation du W3C du 10 février 2004

[11] http://www.w3.org/TR/rdf-mt/RDF semantic; Recommandation du W3C du 10 février 2004;

[12] Kim Y.-H., Kim S., Eom J.-H., Zhang B.-T..SCAI Experiments on TREC-9, Text Retrieval Conference (TREC-9), p. 392-399, 2000.

[13] J. Kim, X. Xue, and W. Croft. A probabilistic retrieval model for semistructured data. Advances in Information Retrieval, pages 228–239, 2009.

[14] M. Koolen, R. Kaptein, and J. Kamps. Focused search in books and Wikipedia. Categories, links and relevance feedback. In S. eva, J. Kamps, and A. Trotman, editors, Focused Retrieval and Evaluation: 8th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2009), volume 6203 of LNCS, pages 273–291, 2010.

[15] Lalmas M.. XML Information Retrieval, Encyclopedia of Library and Information. Sciences,J. Bates and M.N. Maack (Eds), 2009b.

[16] H. Li, T. Y. Liu, and C. Zhai, Learning to rank for information retrieval, SIGIR Forum, vol. 42, no. 2, pp. 76–79, LR4IR 2008.

[17] Louis, A., Nenkova, A.. Performance confidence estimation for automatic summarization. In: EACL, The Association for Computer Linguistics 541–548. 2009.

[18] Noreault, T., M. Koll, and M. McGill. Automatic Ranked Output from Boolean Searches in SIRE. J. American Society for Information Science, 28(6), 333-39. 1977.

[19] Pascal Hitzler, Marcus Krötzsch, Sebastian Rudolph. Foundations of Semantic Web Technologies. Chapman & hall, CRC / 2009.

[20] Pascal Hitzler · Markus Krötzsch, Sebastian Rudolph,York Sure. Semantic Web, Grundlagen, Erste Auflage springer.com; Springer-Verlag Berlin Heidelberg ISBN 978-3-540-33993-9. 2008.

[21] S. Ribaric, I.Fratric. A Matching-Score Normalization Technique for Multimodal Biometric Systems. Proc. 3rd COST 275 Workshop: Biometrics on the Internet, Hatfield, UK, 27-28 pp. 55-58. October 2005.

[22] Robertson S., and Walker, S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In Proc. of the International ACM-SIGIR Conference, pp. 232–241. 1994.

[23] Robertson S., Walker, S., and Beaulieu, M.. Experimentation as a way of life : Okapi at trec. In Information Processing and Management, 36, 95–108.808,809–840. 2000.

[24] Robertson S. and H. Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. Foundations

and Trends in Information Retrieval Vol. 3, No. 4 (2009) 333–389 ; DOI: 10.1561/1500000019. 2009

[25] G. Salton. A comparison between manual and automatic indexing. American Documentation, 20(1) :61–71. 1969.

[26] Schenkel, R., Suchanek, F.M., Kasneci, G.: Yawn. A semantically annotated wikipedia xml corpus. In Kemper, A., Schöning, H., Rose, T., Jarke, M., Seidl, T., Quix, C., Brochhaus, C., eds.: BTW. Volume 103 of LNI., p. 277–291. GI 2007.

[27] Shlomo Geva, Jaap Kampsl, Andrew Trotman. INEX 2010 Workshop Pre-proceeding. Published by: IR Publications, Amsterdam. ISBN 978-90-814485-2-9.; December 6–10,Woodlands of Marburg, Ipswich, Queensland, Australia http://www.inex.otago.ac.nz/ . 2009.

[28] R. Song, M. J. Taylor, J. R. Wen, H. W. Hon, and Y. Yu. Viewing term proximity from a different perspective.A dvances in Information Retrieval (ECIR 2008), Springer LNCS 4956, pp. 346–357, 2008.

[29] Sun, K.-T., Huang, Y.-M., & Liu, M.-C. A WordNet-Based Near-Synonyms and Similar-Looking Word Learning System. Educational Technology & Society, 14 (1), 121–134. 2011.

[30] Stevenson M., Guo Y., Gaizauskas R., Martinez D.. Knowledge sources for word sense disambiguation of biomedical text. BioNLP'08, p. 80-87. 2008.

[31] Trieschnigg D., Pezik P., Lee V., de Jong F., Kraaij W., Rebholz-Schuhmann D., MeSH Up : effective MeSH text classification for improved document retrieval , Bioinformatics, vol. 25, n° 11, p. 1412-1418, June, 2009.

[32] Trotman A.. Choosing document structure weights. Processing and Management, vol. 41, n° 2, p. 243-264, 2005.

[33] Zargayouna, H. et Salotti, S. Mesure de similarité sémantique pour l'indexation de documents semistructurés. In IC'2004 : 15e journées francophones d'Ingénierie des connaissances. Cité 1 fois, p. 11.2004.

Traoré Issa, is a student preparing a Ph.D, at the university Cheick Anta Diop (UCAD) of Dakar in Senegal. He is also engineer in Systems, Networks and telecommunication. This research focuses on semantic Web, data mining and web service discovery. He made a communication on this work during a workshop named CARI 2010. He has taken up web mining as his desired interest area of research.

Oumtanaga Souleymane, is a Professor in Data processing and telecommunication. He teaches in Inp-Hb and is the director of LARIT Laboratory, he is making a lot of research in telecommunication and web mining.

Babri Michel, is a doctor in data processing and carries out research in the field of the data bases. He teaches data processing and the telecommunication in Inp-HB.

Claude Lichou, is a Professor in data processing. He also teaches at the university Cheick Anta Diop (UCAD) of Dakar in Senegal. He supervises the students in Thesis of doctorate. He is my teacher adviser for my thesis, he is making in many fields as follows: mathematics, computing and telecommunication. He is in charge of the technological computing doctorate school of UCAD in Dakar.