Ontology Enrichment based on Generic Basis of Association Rules for Conceptual Document Indexing

Lamia Ben Ghezaiel[1], Chiraz Latiri[2] and Mohamed Ben Ahmed[1]
[1]*Computer Sciences School*, RIADI-GDL *Research Laboratory, Manouba University, 2010, Tunis, Tunisia*
[2]LIPAH *Research Laboratory, Computer Sciences Department, Faculty of Sciences of Tunis, El Manar University, 1068, Tunis, Tunisia*

Keywords: Text Mining, Ontology Enrichment, Information Retrieval, Association Rule, Generic Basis, Distance Measures, Conceptual Indexing.

Abstract: In this paper, we propose the use of a minimal generic basis of association rules (ARs) between terms, in order to automatically enrich an initial domain ontology. For this purpose, three distance measures are defined to link the candidate terms identified by ARs, to the initial concepts in the ontology. The final result is a proxemic conceptual network which contains additional implicit knowledge. Therefore, to evaluate our ontology enrichment approach, we propose a novel document indexing approach based on this proxemic network. The experiments carried out on the OHSUMED document collection of the TREC 9 filtring track and MeSH ontology showed that our conceptual indexing approach could considerably enhance information retrieval effectiveness.

1 INTRODUCTION

Recently, several research communities in text mining and semantic web spent a determined efforts to conceptualize competencies of a given domain through the definition of a domain ontology. However, in order to make that ontology actually of use in applications, it is of paramount importance to enrich its structure with concepts as well as instances identifying the domain.

Many contributions in the literature related to Information Retrieval (IR) and text mining fields proved that domain ontologies are very useful to improve several applications such as ontology-based IR models (Song et al., 2007). While several ontology learning approaches extract concepts and relation instances directly from unstructured texts, in this paper, we show how an initial ontology can be automatically enriched by the use of text mining techniques. Especially, we are interested in mining a specific domain document collections in order to extract valid association rules (Agrawal and Skirant, 1994) between concepts/terms. Thus, we propose to use a minimal generic basis of association rules, called $\mathcal{MGB}$, proposed in (Latiri et al., 2012), to detect additional concepts for expanding ontologies. The result of our enrichment process is a proxemic conceptual network, denoted $\mathcal{O}_{\mathcal{MGB}}$, which unveils the semantic content of a document. To show the benefits of this proxemic conceptual network in the IR field, we propose to integrate it in a document conceptual indexing approach.

The remainder of the paper is organized as follows: Section 2 recalls paradigms for mining generic basis of association rules between terms. In Section 3, we briefly present related works dedicated to the enrichment of ontology. Section 4 introduces a novel automatic approach of ontology enrichment based on the generic basis $\mathcal{MGB}$. Section 5 presents a document conceptual indexing approach based on the enriched ontology $\mathcal{O}_{\mathcal{MGB}}$. Section 6 is devoted to the experimental evaluation, in which the results of the carried out experiments on OHSUMED collection and MeSH ontology are discussed. The conclusion and work in progress are finally presented in Section 7.

2 GENERIC BASIS OF ASSOCIATION RULES

In a previous work (Latiri et al., 2012), we used in the text mining field, the theoretical framework of Formal Concept Analysis (FCA), presented in (Ganter and

Ben Ghezaiel L., Latiri C. and Ben Ahmed M..
Ontology Enrichment based on Generic Basis of Association Rules for Conceptual Document Indexing.
DOI: 10.5220/0004131600530065
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2012), pages 53-65
ISBN: 978-989-8565-30-3

Wille, 1999), in order to propose the extraction of a minimal generic basis of irredundant association rules between terms, named $\mathcal{MGB}$ (Latiri et al., 2012).

2.1 Mathematical Foundations and Basic Definitions

First, we formalize an extraction context made up of documents and index terms, called *textual context*.

2.1.1 Textual Context

Definition 1. *A textual context is a triplet* $\mathfrak{M} = (\mathcal{C}, \mathcal{T}, I)$ *where:*

- $\mathcal{C} = \{d_1, d_2, \ldots, d_n\}$ *is a finite set of n documents of a collection.*
- $\mathcal{T} = \{t_1, t_2, \cdots, t_m\}$ *is a finite set of m distinct terms in the collection. The set* $\mathcal{T}$ *then gathers without duplication the terms of the different documents which constitute the collection.*
- $I \subseteq \mathcal{C} \times \mathcal{T}$ *is a binary (incidence) relation. Each couple* $(d, t) \in I$ *indicates that the document* $d \in \mathcal{C}$ *has the term* $t \in \mathcal{T}$.

Table 1: An example of textual context.

I	A	C	D	T	W
d_1	×	×		×	×
d_2		×	×		×
d_3	×	×		×	×
d_4	×	×	×		×
d_5	×	×	×	×	×
d_6		×	×	×	

Example 1. *Consider the context given in Table 1, used as a running example through this paper and taken from (Zaki, 2004). Here,* $\mathcal{C}$ = {d_1, d_2, d_3, d_4, d_5, d_6} *and* $\mathcal{T}$:= {*A, C, D, T, W*}. *The couple* (d_2, *C*) $\in I$ *since it is crossed in the matrix. This denotes that the document* d_2 *contains the term C.*

Each document $d \in \mathcal{C}$ is represented by a binary vector of length *m*. A *termset T* can be interpreted as a set of *m* terms $T \in \mathcal{T}$, that occur together in the document. For example, *ACW* is a termset composed by the terms *A*, *C* and *W*. The support of a termset is defined as follows:

Definition 2. *Let* $T \subseteq \mathcal{T}$. *The support of T in* $\mathfrak{M}$ *is equal to the number of documents in* $\mathcal{C}$ *containing all the terms of T. The support is formally defined as follows*[1]*:*

$$Supp(T) = |\{d \mid d \in \mathcal{C} \ \wedge \ \forall\, t \in T : (d, t) \in I\}| \quad (1)$$

A termset is said *frequent* (*aka large* or *covering*) if its terms co-occur in the collection a number of times greater than or equal to a user-defined support threshold, denoted *minsupp*.

2.1.2 Galois Closure Operator

Two functions are defined in order to map sets of documents to sets of terms and *vice versa*. Thus, for $T \subseteq \mathcal{T}$, we defined (Ganter and Wille, 1999):

$$\Psi(T) = \{d | d \in \mathcal{C} \ \wedge \ \forall\, t \in T : (d, t) \in I\} \quad (2)$$

$\Psi(T)$ is equal to the set of documents containing all the terms of T. Its cardinality is then equal to $Supp(T)$.

For a set $D \subseteq \mathcal{C}$, we define:

$$\Phi(D) = \{t | t \in \mathcal{T} \ \wedge \ \forall\, d \in D : (d, t) \in I\} \quad (3)$$

$\Phi(D)$ is equal to the set of terms appearing in all the documents of D.

Both functions Ψ and Φ constitute *Galois operators* between the sets $\mathcal{P}(\mathcal{T})$ and $\mathcal{P}(\mathcal{C})$. Consequently, the compound operator $\Omega = \Phi \circ \Psi$ is a *Galois closure operator* which associates to a termset T the whole set of terms which appear in *all* documents where the terms of T co-occur. This set of terms is equal to $\Omega(T)$. In fact, $\Omega(T) = \Phi \circ \Psi(T) = \Phi(\Psi(T))$. If $\Psi(T) = D$, then $\Omega(T) = \Phi(D)$.

Example 2. *Consider the context given in Table 1. Since both terms A and C simultaneously appear in the documents* d_1, d_3, d_4, *and* d_5, *we have:* $\Psi(AC)$ = {d_1, d_3, d_4, d_5}. *On the other hand, since the documents* d_1, d_3, d_4, *and* d_5 *share the terms A, C, and W, we have:* $\Phi(\{d_1, d_3, d_4, d_5\}) = ACW$. *It results that* $\Omega(AC) = \Phi \circ \Psi(AC) = \Phi(\Psi(AC)) = \Phi(\{d_1, d_3, d_4, d_5\}) = ACW$. *Thus,* $\Omega(AC) = ACW$. *In other words, the term W appears in all documents where A and C co-occur.*

2.1.3 Frequent Closed Termset

A termset $T \subseteq \mathcal{T}$ is said to be *closed* if $\Omega(T) = T$. A *closed termset* is then the maximal set of terms common to a given set of document. A closed termset is said to be *frequent* w.r.t. the *minsupp* threshold if $Supp(T) = |\Psi(T)| \geq minsupp$ (Bastide et al., 2000). Hereafter, we denote by FCT a frequent closed termset.

Example 3. *With respect to the previous example, ACW is a closed termset since there is not another*

[1] In this paper, we denote by $|X|$ the cardinality of the set X.

term appearing in all documents containing ACW. ACW is then the maximal set of terms common to the documents $\{d_1, d_3, d_4, d_5\}$. *We then have:* $\Omega(ACW) = ACW$. *If minsupp is set to 3, ACW is also frequent since* $|\Psi(ACW)| = |\{d_1, d_3, d_4, d_5\}| = 4 \geq 3$.

The next property states the relation between the support of a termset and that of its closure.

Property 1. *The support of a termset* T *is equal to the support of its closure* $\Omega(T)$, *which is the smallest* FCT *containing* T, *i.e.,* $Supp(T) = Supp(\Omega(T))$ *(Bastide et al., 2000).*

2.1.4 Minimal Generator

A termset $g \subseteq \mathcal{T}$ is a *minimal generator* of a closed termset T, if and only if $\Omega(g) = T$ and $\nexists\, g' \subset g$: $\Omega(g') = T$ (Bastide et al., 2000).

Example 4. *The termset DW is a minimal generator of CDW since* $\Omega(DW) = CDW$ *and none of its proper subsets has CDW for closure.*

Corollary 1. *Let g be a minimal generator of a frequent closed termset T. According to Property 1, the support of g is equal to the support of its closure, i.e.,* $Supp(g) = Supp(T)$.

2.1.5 Iceberg Lattice

Let $\mathcal{FCT}$ be the set of frequent closed termsets of a given context. When the set $\mathcal{FCT}$ is partially ordered w.r.t. set inclusion, the resulting structure only preserves the *Join* operator (Ganter and Wille, 1999). This structure is called a *join semi-lattice* or an *upper semi-lattice*, and is hereafter referred to as *Iceberg lattice* (Stumme et al., 2002).

In (Latiri et al., 2012), we presented an approach that relies on irredundant association rules mining starting from the *augmented Iceberg lattice*, denoted by $\mathcal{AL} = (\mathcal{FCT}, \subseteq)$, which is the standard Iceberg lattice where each FCT is associated to its minimal generators.

Example 5. *Consider the context given in Table 1. The minsupp threshold is set to 3. The associated augmented Iceberg lattice is depicted in Figure 1, in which the minimal generators associated to each* FCT *are given between brackets.*

Each frequent closed termset T in the Iceberg lattice has an *upper cover* which consists of the closed termsets that immediately cover T in the Iceberg lattice. This set is formally defined as follows:

$$Cov^u(T) = \{T_1 \in \mathcal{FCT} \mid T \subset T_1 \text{ and } \nexists\, T_2 \in \mathcal{FCT}: T \subset T_2 \subset T_1\}$$

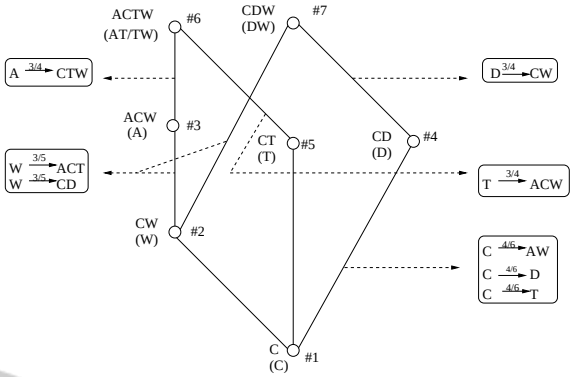

Figure 1: The augmented Iceberg lattice.

Example 6. *Let us consider the frequent closed termset CW of the Iceberg lattice depicted by Figure 1. Then, we have:* $Cov^u(CW) = \{ACW, CDW\}$.

2.2 Association Rules Mining

An association rule R is an implication of the form R: $T_1 \Rightarrow T_2$, where T_1 and T_2 are subsets of $\mathcal{T}$, and $T_1 \cap T_2 = \emptyset$. The termsets T_1 and T_2 are, respectively, called the *premise* and the *conclusion* of R. The rule R is said to be based on the termset T equal to $T_1 \cup T_2$. The *support* of a rule R: $T_1 \Rightarrow T_2$ is then defined as:

$$Supp(R) = Supp(T) \tag{4}$$

while its *confidence* is computed as:

$$Conf(R) = \frac{Supp(T_1)}{Supp(T)} \tag{5}$$

An association R is said to be *valid* if its confidence value, *i.e.*, $Conf(R)$, is greater than or equal to a user-defined threshold denoted $minconf$[2]. This confidence threshold is used to exclude non valid rules.

Example 7. *Starting from the context depicted in Table 1, the association rule R:* $W \Rightarrow CD$ *can be derived. In this case,* $Supp(R) = Supp(CDW) = 3$, *while* $Conf(R) = \frac{Supp(CDW)}{Supp(W)} = \frac{3}{5}$. *If we consider the minsupp and minconf thresholds respectively equal to 3 and 0.5, the considered rule R is valid since* $Supp(R) = 3 \geq 3$ *and* $Conf(R) = \frac{3}{5} \geq 0.5$.

2.3 Minimal Generic Basis of Association Rules

Given a document collection, the problem of mining association rules between terms consists in gen-

[2] In the remainder, $T_1 \overset{c}{\Rightarrow} T_2$ indicates that the rule $T_1 \Rightarrow T_2$ has a value of confidence equal to c.

erating all association rules given user-defined *minsupp* and *minconf* thresholds. Several approaches in the literature deal with the redundancy problem. More advanced techniques that produce only a limited number of rules rely on Galois closure (Ganter and Wille, 1999). These techniques focus on extracting irreducible nuclei of all association rules, called *generic basis*, from which the remaining association rules can be derived (Bastide et al., 2000; Latiri et al., 2012). An interesting discussion about the main generic bases of association rules is proposed in (Ben Yahia et al., 2009; Balcázar, 2010; Latiri et al., 2012).

However, the huge number of irredundant association rules constitutes a real hamper in several applications related to text mining. To overcome this problem, we proposed in (Latiri et al., 2012) the use of a *minimal generic basis*, called $\mathcal{MGB}$, based on the extraction of the *augmented Iceberg lattice*. This basis involves rules that maximize the number of terms in the conclusion. We distinguish two types of association rules: *exact association rules* (with confidence equal to 1) and *approximate association rules* (with confidence less than 1)(Zaki, 2004).

Hence, in the following, we will adapt the Minimal Generic Basis $\mathcal{MGB}$ of association rules, defined in (Latiri et al., 2012), to enrichment ontology issue. When considering a textual context $\mathfrak{M} := (\mathcal{C}, \mathcal{T}, I)$, the minimal generic basis $\mathcal{MGB}$ is defined as follows:

Definition 3. *Given $\mathcal{AL}$ an Iceberg Galois lattice augmented by minimal generators and their supports, T_i a frequent closed termset, $Cov^u(T_i)$ its upper cover and $\mathcal{G}_{T_i}$ the list of its minimal generators of the frequent closed termset T_i, we have:*

$$\mathcal{MGB} = \left\{ \begin{array}{l} R : g \rightarrow (T_i - g) \mid g \in \mathcal{G}_{T_i} \wedge T_i \in \mathcal{AL}_c \wedge \\ Conf(R) \geq minconf \wedge \nexists\, s \in Cov^u(T_i) \mid \\ \frac{support(s)}{support(g)} \geq minconf \end{array} \right. \quad (6)$$

In our approach, the augmented Iceberg lattice $\mathcal{AL}$ supports the irredundant association rules discovery between terms. The main advantage brought by this partially ordered structure is the efficiency. In fact, by using such a precedence order, irredundant exact and approximate association rules are directly derived, without additional confidence measure computations.

The GEN-MGB algorithm which allows the construction of the $\mathcal{MGB}$ generic basis is detailed in (Latiri et al., 2012). It iterates on the set of frequent closed termsets $\mathcal{FCT}$ of the augmented Iceberg lattice $\mathcal{AL}$, starting from larger FCTs and sweeping downwardly w.r.t. set inclusion $\subseteq$. The algorithm takes the augmented Iceberg lattice $\mathcal{AL}$ as input and gives as output the irredundant approximate and exact association rules (*i.e.*, IARs and IERs). With respect to Equation (6) and considering a given node in the Iceberg lattice, we consider that IARs represent implications that involve the minimal generators of the sub-closed-termset, associated to the considered node, and a super-closed-termset. On the other hand, IERs are implications extracted using minimal generators and their respective closures, belonging to the same node in $\mathcal{AL}$ (Latiri et al., 2012).

Example 8. *Consider the augmented Iceberg lattice depicted in Figure 1 for minconf = 0.6. Let us recall that the set value of minsupp is equal 3. All irredundant approximate association rules are depicted in Figure 1. In this case, none irredundant exact rule is mine since all of them are redundant w.r.t. irredundant approximate rules belonging to $\mathcal{MGB}$. For example, starting from the node having CDW for frequent closed termset, the exact rule $DW \stackrel{1}{\Rightarrow} C$ is not generated since it is considered as redundant w.r.t. the approximate association rule $D \stackrel{0.75}{\Rightarrow} CW$.*

In this regard, we propose in this paper to disclose how that can be achieved when a domain ontology is enriched using irredundant association rules belonging to $\mathcal{MGB}$.

3 RELATED WORKS TO ENRICHMENT ONTOLOGY

In the literature, there is no common formal definition of what an ontology is. However, most approaches share a few core items: concepts, a hierarchical *is-a*-relation, and further relations. For sake of generality, we formalize an ontology in the following way (Cimiano et al., 2004):

Definition 4. *An ontology is a tuple $O = \langle \mathcal{C}_\mathcal{D}, \leq_\mathcal{C}, \mathcal{R}, \leq_\mathcal{R} \rangle$, where $\mathcal{C}_\mathcal{D}$ is a set whose elements are called concepts of a specific domain, $\leq_\mathcal{C}$ is a partial order on $\mathcal{C}_\mathcal{D}$ (*i.e.*, a binary relation* is-a $\subseteq \mathcal{C}_\mathcal{D} \times \mathcal{C}_\mathcal{D}$*), $\mathcal{R}$ is a set whose elements are called relations, and $\leq_\mathcal{R}$ is a function which assigns to each relation name its arity.*

Throughout this paper, we will consider the Definition 4 to designate a specific domain ontology. It is thus considered as a structured network of concepts extracted from a specific domain and interconnects the concepts by semantic relations. Naturally, the construction of an ontology is hard and constitutes an expensive task, as one has to train domain experts in formal knowledge representation. This knowledge is usually evolvable and therefore an ontology maintenance process is required (Valarakos et al., 2004; Di-Jorio et al., 2008) and plays a main role as ontolo-

gies may be misleading if they are not up to date. In the context of ontology maintenance, we tackle in this paper, the problem of enrichment of an initial ontology with additional concept derived from a compact set of irredundant association rules.

Roughly speaking, ontology enrichment process is performed in two main steps namely: a *learning step* to detect new concepts and relations and a *placement step* which aims to find the appropriate place of these concepts and relations in the original domain ontology. Several works in literature were proposed to handle the two previous steps. In this work, we focus on methods dedicated to the discovery of new candidate terms from text and their relation with initial concepts in a domain ontology.

Thereby, two main classes of methods have been explored for detecting candidate terms, namely (DiJorio et al., 2008):

- *Statistical based Methods.* They consist in counting the number of occurrences of a given term in the corpus. Most of the statistical methods select candidate terms with respect to their distribution in the corpus (Faatz and Steinmetz, 2002; Parekh et al., 2004), as using other measures such as test mutual information, *tf-idf* (Neshatian and Hejazi, 2004), T-test or statistic distribution laws. These methods allow to identify the new concepts but they are not able to add them into the original ontology without the help of the domain expert (DiJorio et al., 2008).
- *Syntactic based Methods.* They require a grammatical sentence analysis. Most of these methods include upstream a part of speech tagging process. They assume that grammatical dependencies reflect semantic dependencies (Bendaoud et al., 2008). In (Maedche et al., 2002; Navigli and Velardi, 2006), authors introduced the use of lexico-syntactic patterns to detect ontological relations. However, to overcome the problem of huge number of related terms extracted, data mining techniques are applied in some approaches such as association rules discovery from the syntactic dependencies (Benz et al., 2010). So, association rules based approaches allow strong correlations detection. They highlight frequent grammatical dependencies and thus are a good way to prune many insignificant dependencies through the metrics of support and confidence and pruning irredundant associations rules (Balcázar, 2010). The first advantage of the syntactic based methods compared to statistical based ones is that they allow to put automatically new terms into the initial ontology. Nevertheless, they do not label new relations.

4 ONTOLOGY ENRICHMENT BASED ON A GENERIC BASIS OF ASSOCIATION RULES

We propose in this paper a fully automatic process to expand a given ontology, based on the minimal generic basis association rules $\mathcal{M}\,\mathcal{G}\,\mathcal{B}$, defined in Subsection 2.3. Indeed, we propose to use association rules between terms to discover new concepts and relations which link them to other concepts. We aim to enhance the knowledge captured in a domain ontology, leading to a proxemic conceptual network to perform then conceptual indexing in IR. The main motivation behind the idea, that for a given domain ontology, we focus on finding out automatically uniquely relevant concepts for enrichment by using irredundant association rules between terms. This allows to reduce the huge number of related terms extracted by removing redundant association rules during the derivation process. For this, we propose to:

1. Generate a minimal generic basis of irredundant rules $\mathcal{M}\,\mathcal{G}\,\mathcal{B}$ from a specific document collection to the domain;
2. Detect a set of candidate concepts from the basis $\mathcal{M}\,\mathcal{G}\,\mathcal{B}$. This implies that an ontology-based approach is needed to calculate the semantic distances between the candidate concepts;
3. Select a subset of those candidate concepts with respect to their neighborhood to concepts already existing in the original domain ontology;
4. Add new concepts to the ontology;
5. Build a proxemic conceptual network from the enriched ontology in order to perform then conceptual document indexing in IR.

We assume that we have, for a given domain, a document collection denoted $\mathcal{C}$. Before mining the generic basis $\mathcal{M}\,\mathcal{G}\,\mathcal{B}$, we need to generate the textual context $\mathfrak{M} = (\mathcal{C}, \mathcal{T}, I)$ from the collection $\mathcal{C}$. Hence, in order to derive the generic basis of association rules between terms $\mathcal{M}\,\mathcal{G}\,\mathcal{B}$ from our textual context $\mathfrak{M}$, we used the algorithm GEN-MGB to get out irredundant associations between terms (*i.e.,* approximative and exact ones) (Latiri et al., 2012).

Furthermore, we consider an initial domain ontology denoted by $\mathcal{O}$, such as the medical ontology MeSH (Díaz-Galiano et al., 2008). A such ontology includes the basic primitives of an ontology which are concepts and taxonomic relations such as the subsomption link *is-a*. Then, to evaluate the strength of the semantic link between two concepts inside the ontology $\mathcal{O}$, we use *Wu and Palmer's* similarity measure (Wu and Palmer, 1994). It is a measure between

concepts in an ontology restricted to taxonomic links. Given two concepts C_1 and C_2, *Wu and Palmer's* similarity measure is defined as follow (Wu and Palmer, 1994):

$$Sim_{WP}(C_1,C_2) = \frac{2 \times depth(C)}{depth(C_1) + depth(C_2)} \quad (7)$$

where $depth(C_i)$ is the distance which separates the concept C_i to the ontology root and C is the common concept ancestor of C_1 and C_2 in the given ontology.

4.1 Enrichment Ontology Process

Our enrichment process aims to bring closer the original ontology O of the terms contained in the premises of $\mathcal{MGB}$ association rules. Once the new concepts placed in the ontology, we calculate the different distance measures which evaluate the semantic links existing between the concepts of enriched ontology, denoted in the sequel by $O_{\mathcal{MGB}}$.

The $\mathcal{MGB}$-based enrichment process iterates through the following steps.

4.1.1 Step 1: Detecting Candidate Concepts for Enrichment

We calculate for each concept C_O in the initial ontology O, the set of candidate concepts to be connected to C_O. This set includes the terms in the conclusion parts of valid association rules in $\mathcal{MGB}$, whose premise is C_O.

In the example depicted in Figure 2, the candidate concepts for the enrichment related to the concept C_1 are $\{C_{10}, C_{12}, C_5, C_{15}\}$.

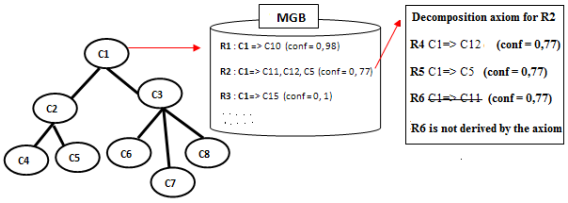

Figure 2: Example of detecting candidate concepts.

4.1.2 Step 2: Placement of New Concepts

In this step, we add the candidate concepts to the initial ontology O, while maintaining existing relations. This allows to avoid adding redundant links in the case of a concept candidate to be linked to several concepts in the initial ontology O. In other words, given a valid association rule $R : C_O \Rightarrow C_j$ in $\mathcal{MGB}$, we select the candidate concept C_j in the the association rule R related to the initial concept $C_O \in O$ where

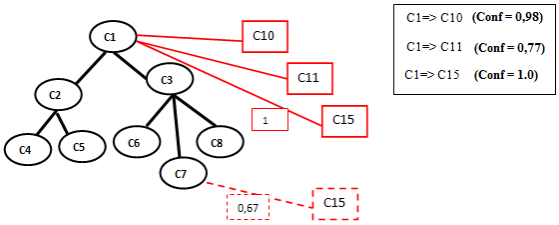

Figure 3: Example of candidate concepts placement.

R has the greater confidence among those in $\mathcal{MGB}$ and having C_O as premise.

The example depicted in Figure 3 illustrates the placement of the new concepts C_{10} et C_{11} where concept C_{15} is removed since $Conf(C_1 \Rightarrow C_{15}) = 1$ is greater than $Conf(C_7 \Rightarrow C_{15}) = 0.67$.

4.1.3 Step 3: Computing of C_i Neighborhood and Distance Measures

Among extracted terms as conclusions of valid association rules in $\mathcal{MGB}$, there are some already known terms, as they are already referenced as concepts by the initial ontology O. In order to link only new terms extracted with existing concepts, we propose to define the neighborhood of these concepts. Given a concept C_O in O, its neighborhood is defined as follows:

Definition 5. *Let C_O be a concept, the neighborhood $\mathcal{N}_{C_O}$ of C_O is the set of concepts in the ontology O that can be accessed from C_O by using the hierarchical link or by using one or more valid irredundant associations rules in $\mathcal{MGB}$.*

The relations between a concept C_i in O and its neighborhood, *i.e.*, each candidate concept $C_k \in \mathcal{N}_{C_i}$, are evaluated through a statistical measure called *distance measure* between C_i and its neighborhood, denoted $Dist_{O_{\mathcal{MGB}}}$. It is calculated based on: (*i*) the confidence values of the association rules in $\mathcal{MGB}$ selected for the ontology enrichment; and, (*ii*) similarities between concepts in the original ontology O assessed using *Wu and Palmer's* similarity measure (Wu and Palmer, 1994) (*cf.* Equation (7)).

In our enrichment approach, three configurations are possible to evaluate the distance measure between two concepts C_i and C_j in the enriched ontology $O_{\mathcal{MGB}}$. For this, we present the following propositions.

Proposition 1. *Given C_i a concept of the initial ontology O. If it exists a link between C_i and a concept C_j derived from an association rule in the generic basis $\mathcal{MGB}$, then:*

$$Dist_{O_{\mathcal{MGB}}}(C_i,C_j) = Conf(R : C_i \Rightarrow C_j) \quad (8)$$

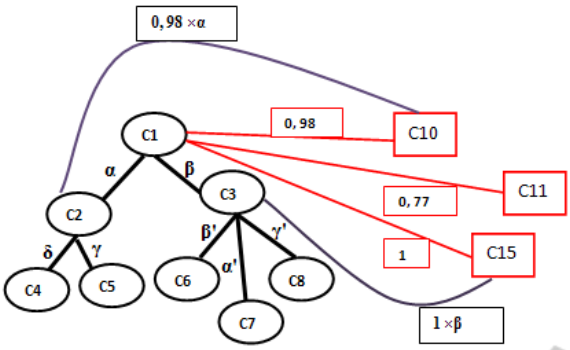

Figure 4: Examples of distance evaluation between concepts in $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$.

Proposition 2. *If C_i and C_j belong to the initial ontology O then:*

$$Dist_{O_{\mathcal{M}\mathcal{G}\mathcal{B}}}(C_i,C_j) = Sim_{WP}(C_i,C_j) \quad (9)$$

Proposition 3. *If C_i is a concept added to the enriched ontology $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$ and linked to C_O where $Dist_{O_{\mathcal{M}\mathcal{G}\mathcal{B}}}(C_O,C_i) = Conf(R : C_O \Rightarrow C_i) = \beta$, then each concept C_j in O in relation with C_O where $Sim_{WP}(C_O,C_j) = \alpha$, is also in relation with C_i. In this case, the distance measure is a mixte one and it is calculated as follows:*

$$Dist_{O_{\mathcal{M}\mathcal{G}\mathcal{B}}}(C_i,C_j) = \alpha \times \beta. \quad (10)$$

Thereby, we consider that the neighborhood $\mathcal{N}(C_i)$ of a concept C_i involves the set of k concepts belonging to the conceptual proximic network $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$, in relation with the concept C_i, where the semantic distance between them is greater than or equal to a user-defined θ. Formally, we have:

$$\mathcal{N}(C_i) = \{C_j \mid Dist_{O_{\mathcal{M}\mathcal{G}\mathcal{B}}}(C_i,C_j) \geq \theta,\ j \in [1..k]\} \quad (11)$$

Example 9. *The three configurations of the distance measure evaluation are depicted in Figure 4, namely:*

- ***Case 1** (Proposition 1): The concept C_{10} is selected from an association rule in $\mathcal{M}\mathcal{G}\mathcal{B}$, so: $Dist_{O_{\mathcal{M}\mathcal{G}\mathcal{B}}}(C_1,C_{10}) = Conf(C_1 \Rightarrow C_{10}) = 0.98$.*
- ***Case 2** (Proposition 2): The two concepts C_1 and C_2 belong to the initial ontology O and $Dist_{O_{\mathcal{M}\mathcal{G}\mathcal{B}}}(C_1,C_2) = Sim_{WP}(C_1,C_2) = \alpha$.*
- ***Case 3** (Proposition 3): A mixte distance is computed between C_2 and C_{10} since C_1 is linked to C_{10} thanks to the valid association rule $R : C_1 \overset{0.98}{\Rightarrow} C_{10}$ and an initial relation exits in O between C_1 and C_2, so: $Dist_{O_{\mathcal{M}\mathcal{G}\mathcal{B}}}(C_2,C_{10}) = 0.98 \times \alpha$*

The generated result, *i.e.*, the enriched ontology $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$, is then explored as a proxemic conceptuel network to represent the domain knowledge.

4.2 $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$: A Proxemic Conceptuel Network for Knowledge Representation

In what follows, we describe an original proposal for knowledge representation, namely a proxemic conceptual network resulting from the enriched ontology $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$. The relationships between the concepts of the conceptual network is quantified by the distance measures introduced in Propositions 8, 9 and 10. The originality of our proxemic conceptual network is its completeness thanks to the combination, on the one hand, of knowledge stemming from the initial ontology, *i.e.*, concepts and semantic relations, and, on the other hand, implicit knowledge extracted as association rules between terms.

Thus, a concept in our proxemic conceptual network has three levels of semantic proximities, namely:

1. **A Referential Semantic:** assignes to each concept of $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$ an intensional reference, *i.e.*, its best concept sense through a disambiguation process.
2. **A Differential Semantic:** associates to each concept its neighbors concepts, *i.e.*, those correlated with it in the local context according to the Equation (11).
3. **An Inferential Semantic:** induced by irredundant association rules between terms that associate to each concept a inferential potential. In our case, it will link the initial ontology concepts to concepts contained in valid association rules of $\mathcal{M}\mathcal{G}\mathcal{B}$ with respect to a minimum threshold of confidence *minconf* and the proposed distance measure.

Indeed, around each concept C_i in $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$, there is a semantic proxemic subspace which represents the different relations by computing distance measure $Dist_{O_{\mathcal{M}\mathcal{G}\mathcal{B}}}$ between its different neighbors concepts, and between concepts and their extensions. This results an enriched knowledge representation.

In order to prove that the proxemic conceptual network can have a great interest in IR and can contribute to improve the retrieval effectiveness, we propose a document conceptual indexing approach based on $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$.

5 EVALUATION OF $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$ IN INFORMATION RETRIEVAL

Several ways of introducing additional knowledge

into Information Retrieval (IR) process have been proposed in recent literature. In the last decade, ontologies have been widely used in IR to improve retrieval effectiveness (Vallet et al., 2005). Interestingly enough, such ontology-based formalisms allow a much more structured and sophisticated representation of knowledge than classical thesauri or taxonomy (Andreasen et al., 2009). They represent knowledge on the semantic level thanks to concepts and relations instead of simple words.

Indeed, main contributions in *document conceptual indexing* issue are based on detecting new concepts from ontologies and taxonomies and use them to index documents instead of using simple lists of words (Baziz et al., 2005; Andreasen et al., 2009; Dinh and Tamine, 2011). Roughly, the indexing process is performed in two steps, namely (*i*) first detecting terms in the document text by mapping document representations to concepts in the ontology; then, (*ii*) disambiguating the selected terms.

In the following, we introduce a novel conceptual documents indexing approach in IR, based on the proxemic network $O_{\mathcal{MGB}}$ which is the result of the enrichment of an initial ontology O with the generic basis $\mathcal{MGB}$. Our strategy involves three steps detailed below: (1) Identification and weighting representative concepts in the document through the conceptual network $O_{\mathcal{MGB}}$; (2) Concepts disambiguation using the enriched ontology; and, (3) Building the document semantic kernel, denoted *Doc*-$O_{\mathcal{MGB}}$, by selecting the best concepts-senses.

5.1 Identification and Weighting of Representative Concepts in the Document

We assume that a document d is represented by a set of terms, denoted by $d = \{t_1, \ldots, t_m\}$ and resulting from the terms extraction stage. A term t_i of a document d, denoted $t = \{w_1, \ldots, w_n\}$ is composed of one or more words and its length $|t|$ is defined as the number of words in t.

This step aims to identify and weight, for each index term of the document, the corresponding concept in the proxemic conceptual proxemic $O_{\mathcal{MGB}}$. Thus, the process of identification and weighting of representative concepts in a document d proceeds as follows:

1. **Projection of the Document Index on $O_{\mathcal{MGB}}$.** It allows to identify mono-words or multi-words concepts which correspond to index terms with respect to their occurrence frequencies (Amirouche et al., 2008). We select then the set of terms t_i characterizing the document d, denoted by $T(d)$, namely:

$$T(d) = \{(t_1, f(t_1)), \ldots, (t_n, f(t_n))\} \text{ such that } t_i \in d, \quad (12)$$

where $f(t_i)$ means the occurrence frequency of t_i in d.

2. **Concepts Weighting.** A widely used strategy in IR for weighting terms is $tf \times idf$ and its variants, expressed as $W(t,d) = tf(t) \times idf(t,d)$ (Salton and Buckely, 1988). Il this formula, tf represents term frequency and idf is the inverse document frequency. In (Baziz et al., 2005), authors proposed, for the case of multi-word terms, a statistical weighting method named $cf \times idf$ and based both on classical $tf \times idf$ measure and the length of the terms. So, for an extracted concept t composed of n words, its frequency in a document d is equal to the number of occurrences of the concept itself t, and the one of all its sub-concepts. The frequency is calculated as follows (Baziz et al., 2005):

$$cf(t) = f(t) + \sum_{t_i \in sub(t)} \left(\frac{|t_i|}{|t|} \times f(t_i)\right) \quad (13)$$

where $sub(x)$ is the set of all possible sub-sets which can be derived from a term x, $|x|$ represents the number of words in x and $f(t)$ is the occurrence frequency of t in d.

In this step of our conceptual indexing process, concepts weighting assignes to each concept a weight that reflects its importance and representativity in a document. We propose a new weighting measure which considers both *statistical* and *semantic* representativities of concepts in a given document.

The key feature of the weighting way that we propose is to consider for a term t, in addition to its weight given by $cf(t) \times idf(t,d)$, the weights of concepts C_i belonging to its neighborhood.

Hence, the *statistical representativity*, denoted W_{Stat}, is computed by using Equation (13), namely:

$$W_{Stat}(t,d) = cf(t) \times idf(t,d) \quad (14)$$

Moreover, while considering the proxemic conceptuel network $O_{\mathcal{MGB}}$, we propose that the *semantic representativity* of a term t in a document d, denoted $W_{Sem}(t,d)$, takes into account the different links in $O_{\mathcal{MGB}}$ between each occurrence of t and other concepts in its neighborhood. This semantic representativity is computed by using the semantic distance measure $Dist_{O_{\mathcal{MGB}}}$ as defined in

Propositions 1, 2 and 3, between each occurrence C_i of a term t in $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$ and the concepts in its neighborhood $\mathcal{N}(C_i)$. It is computed as follows:

$$W_{Sem}(t,d) = \sum_{C_i \in S_t} \sum_{C_j \in \mathcal{N}(C_i)} Dist_{O_{\mathcal{M}\mathcal{G}\mathcal{B}}}(C_i, C_j) \times f(C_j) \quad (15)$$

such that $S_t = \{C_1, C_2, \ldots, C_n\}$ is the set of all concepts linked to the term t, *i.e.*, occurrences of t.

The underlying idea is that the global representativity of a term t in a document d, *i.e.*, its weight, further denoted W_{Doc}, is formulated as the combination between the statistical representativity and the semantic one, and is computed as:

$$W_{Doc}(t,d) = W_{Stat}(t,d) + W_{Sem}(t,d) \quad (16)$$

The document index, denoted $Index(D)$, is then generated by selecting only terms whose global representativity, *i.e.*, $W_{Doc}(t,D)$, is greater than or equal to a minimal representativity threshold.

5.2 Concepts Disambiguation

We assume that each term t_i in a document d can have multiple senses, denoted by $S_i = \{C_1^i, \ldots, C_n^i\}$ and are represented by corresponding concepts in $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$. Thus, a term $t_i \in Index(d)$ has $|S_i|$=n senses, *i.e.*, it represents n concepts in $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$.

Given a term t in the document index $Index(d) = \{t_1, \ldots, t_m\}$, the disambiguation aims to identify and to assign it the appropriate sense with respect to its context. We propose in the following an $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$-based disambiguation approach. In this regard, we consider that each index term in $Index(d)$ contributes to the semantic representation of d with only one sense even if a term can have different senses in the same document (Amirouche et al., 2008). Hence, disambiguation task consists to select for each index term in $Index(d)$, its best sense in d, with respect to a computed score for each concept-sense in $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$.

In the literature, various methods and metrics have been proposed for disambiguating words in text (Navigli, 2009). In our case, we got inspired by the approach described in (Baziz et al., 2005) which is based on the computation of a score for every concept-sense linked to an index term and using WordNet ontology.

Our disambiguation approach differs from that one proposed in (Baziz et al., 2005) in the way of calculating the score. Indeed, we believe that only considering the semantic proximity between concepts is insufficient to detect the best sense of a term. In (Baziz et al., 2005), the authors do not take into account the representativity of the terms in the document context. Besides, they do not consider local context of the word in the document, *i.e.*, the correlation of the senses of neighbors terms, and in the concept hierarchy.

To overcome these limits, we suggest that the best sense to be assigned to a term t_i in the document d shall be strongly correlated with other elements, namely:

1. *The local context of the term t_i in the document d*: This means that the disambiguation of t_i considers its neighbors terms in the document d. We define the local context of a term t_i as follows:

   Definition 6. *The local context of a term t_i in a document d, denoted $Context_d(t_i)$, is a termset in $T(d)$ belonging to the same sentence where appears t_i.*

2. *The context of each sense in $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$*: The disambiguation of a concept C_i considers its neighborhood, *i.e.*, $\mathcal{N}_{C_i}$.
3. *Term representativity in the document context*: The best sense for a term t_i in d is that which is highly correlated with the most important sense in d. To do this, we integrate the term weight in the document, *i.e.*, its global representativity, computed w.r.t Equation (16).

In our disambiguation approach, we firstly define the weight of a concept-sense C_j^i in S_i as the weight of its associated index term t_i. That is, for a term t_i, the score of its j^{th} sense, denoted by C_j^i, is computed as:

$$C\_Score(C_j^i) = \sum_{C_v \in \mathcal{N}(C_j^i) \cup C_j^i} \sum_{t_l \in Context_d(t_i), l \neq i} Score_{Doc}(t_i, t_l) \times Dist(C_v, t_l) \quad (17)$$

where:

$$Score_{Doc}(t_i, t_l) = W_{Doc}(t_i, d) \times W_{Doc}(t_l, d) \quad (18)$$

and

$$Dist(C_v, t_l) = \sum_{k \in [1..n_l]} Dist_{O_{\mathcal{M}\mathcal{G}\mathcal{B}}}(C_v, C_k^l) \quad (19)$$

such that n_l represents the number of senses in $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$ which is proper to each term t_l, $W_{Doc}(t_i,d)$ and $W_{Doc}(t_l,d)$ are the weights associated to t_i and t_l in the document d.

The concept-sense C_i which maximizes the score $C\_Score(C_j^i)$ is then retained as the best sense of the term t_i. Formally, we have:

$$C_i = \underset{j \in [1..n_i]}{\arg\max} \{C\_Score(C_j^i)\} \quad (20)$$

Indeed, by performing the different formulas (17), (18), (19) and (20), we have disambiguated the concept C_i which will be a node in the proxemic semantic network of the document d.

5.3 Building the Proxemic Semantic Network *Doc-*$O_{\mathcal{MGB}}$

The last step of our document conceptual indexing process is the building of the the proxemic network representing a document content, denoted in the sequel *Doc-*$O_{\mathcal{MGB}}$. To do so, we select its nodes, *i.e.*, concept senses, by computing for each of them the best score *C_Score*. The nodes of *Doc-*$O_{\mathcal{MGB}}$ are initialized with the selected concepts in the disambiguation step. Then, each concept of *Doc-*$O_{\mathcal{MGB}}$ is declined in more intensions and extensions thanks to the structure of $O_{\mathcal{MGB}}$, which are themselves linked to other concepts of the generic basis of association rules $\mathcal{MGB}$.

Thus, around each node in *Doc-*$O_{\mathcal{MGB}}$ gravitates a three-dimensional proxemic field synthesizing three types of semantic, namely the referential semantic, the differential semantic and the inferential one as explained in Sub-section 4.2.

Therefore, thanks to the obtained structure, *i.e.*, *Doc-*$O_{\mathcal{MGB}}$, we move from a simple index containing single index terms to a proxemic three-dimensional indexing space. The expected advantage on this novel document representation is to get a richer and more precise meaning representation in order to obtain a more powerful identification of relevant documents matching the query in an IR system.

6 EXPERIMENTAL EVALUATION

In order to evaluate our ontology enrichment approach based on the generic basis $\mathcal{MGB}$, we propose to incorporate the generated conceptual proxemic network *Doc-*$O_{\mathcal{MGB}}$ into a conceptual document indexing framework. For this purpose, we consider the well known medical ontology MeSH as domain ontology (Díaz-Galiano et al., 2008). Indeed, in MeSH, each concept is described by a main heading (preferred term), one or many concept entries (non-preferred terms), qualifiers, scope notes, etc. Thus, we used main headings and qualifiers as indexing features in our evaluation.

6.1 Test Collection

We used the OHSUMED test collection, which is a MEDLINE sub-collection used for biomedical IR in TREC9 filtering Track, under the Terrier IR platform (http://terrier.org/). Each document has been annotated by human experts with a set of MeSH concepts revealing the subject matter(s) of the document. Some statistical characteristics of the OHSUMED collection are depicted in Table 2.

Table 2: OHSUMED test collection statistics.

Number of documents	348, 566
Average document length	100 tokens
Number of queries	63
Average query length	12 terms
Average number of relevant docs/query	50

6.2 Experimental Setup

For measuring the IR effectiveness, we used exact precision measures P@10 and P@20, representing respectively the mean precision values at the top 10 and 20 returned documents, and MAP representing the *Mean Average Precision* computed over all topics. The purpose of our experimental evaluation is to determine the utility of our ontology enrichment approach on the MeSH ontology using irredundant association rules between terms which are derived from the document collection OHSUMED. Hence, we propose to assess the impact of exploiting the conceptual proxemic network $O_{\mathcal{MGB}}$ in document indexing on the retrieval effectiveness.

Therefore, we carried out two series of experiments applied on the articles titles and abstracts. The first one is based on the classical document indexing using the state-of-the-art weighting scheme OKAPI BM25 (Jones et al., 2000), as the baseline, denoted BM25. The second one concerns our conceptual indexing approach and consists of four scenarios, namely:

1. The first one concerns the document expansion using concepts manually assigned by human experts, denoted I_{Expert}.
2. The second one concerns the document expansion using only preferred concepts of the MeSH ontology before any enrichment, denoted I_{MeSH}.
3. The third one concerns the document expansion based on additional terms derived from valid association rules of the $\mathcal{MGB}$ generated from the document collection OHSUMED, denoted $I_{\mathcal{MGB}}$. Notice that we set up minimal support threshold *minsupp* and minimal confidence threshold *minconf*, respectively to, 0.05 and 0.3.
4. The last one concerns the document expansion using concepts identified from the proxemic conceptual network *Doc-*$O\mathcal{MGB}$, denoted $I_{O_{\mathcal{MGB}}}$, which is the result of the MeSH enrichment with the generic basis of irredundant association rules $\mathcal{MGB}$ derived from OHSUMED collection.

6.3 Results and Discussion

We now present the experimental results of the proposed document indexing strategies. We assess the IR effectiveness using the extracted concepts and our proposed disambiguation approach.

Table 3 depicts the IR effectiveness of the I_{Expert}, I_{MeSH} and our two semantic indexing approaches based on the generic basis of association rules $\mathcal{M}\mathcal{G}\mathcal{B}$ and the proxemic conceptuel network $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$. We observe that in an automatic setting, our best indexing method, namely $I_{O_{\mathcal{M}\mathcal{G}\mathcal{B}}}$, provides the highest improvement rate ($+17.57\%$) while the $\mathcal{M}\mathcal{G}\mathcal{B}$ based method only gives $+4.17\%$ in terms of MAP over the baseline BM25. This proves the interest to take into account both terms from association rules and the concepts selected from enriched ontology during the concept extraction process. Results highlight that using only concepts extracted from the MeSH ontology lead to a small improvement of IR effectiveness in the case of document indexing. Furthermore, we see that the I_{Expert}, $I_{\mathcal{M}\mathcal{G}\mathcal{B}}$ and $I_{O_{\mathcal{M}\mathcal{G}\mathcal{B}}}$ methods consistently outperform the baseline BM25.

Although the gain of the $I_{O_{\mathcal{M}\mathcal{G}\mathcal{B}}}$ method is smaller than the I_{Expert} method, which represents the best scenario, in terms of MAP (23.33% *vs.* 17.57%) (*cf.* Table 3), the former yields improvement in terms of P@10 and P@20, which is less significant in the other methods, namely I_{MeSH} and $I_{\mathcal{M}\mathcal{G}\mathcal{B}}$.

Figure 5 sheds light on the advantage of the insight gained through the $\mathcal{M}\mathcal{G}\mathcal{B}$ irredundant association rules and the conceptual proxemic network $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$ in the context of conceptual document indexing. We note that the increase of the precision at 11 points of recall with the $I_{\mathcal{M}\mathcal{G}\mathcal{B}}$ method is not so important with respect to the baseline BM25. This can be explained by the fact that OHSUMED is a scientific medical collection where terms have very weak distributions and marginally co-occur. Moreover, an important part of the vocabulary is not used, since it is not correctly analyzed, due to the used tagger which does not identify specific and scientific terms of OHSUMED. Moreover, we notice in our experiments that the improvement of the average precision is less significant for high support values. Indeed, extracting association rules, when considering a high support value, leads to some trivial associations between terms that are very frequent in the document collection.

In order to show how our indexing approach based on the conceptual proxemic network $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$ is statistically significant, we perform the Wilcoxon signed rank test (Smucker et al., 2007) between means of each ranking obtained by our indexing method and

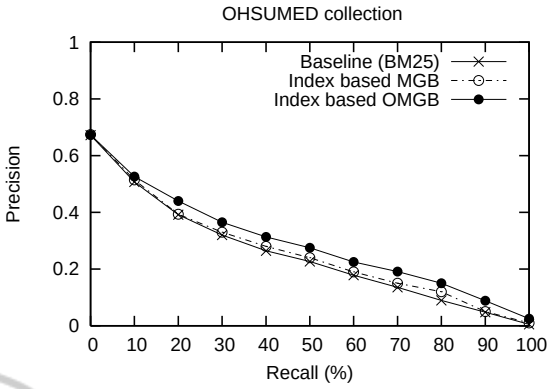

Figure 5: Precision-recall curves corresponding to the baseline *vs* index based on $\mathcal{M}\mathcal{G}\mathcal{B}$ and $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$.

the baseline BM25. The reason for choosing the Wilcoxon signed rank test is that it is more powerful and indicative test as it considers the relative magnitude in addition to the direction of the differences considered. Experimental results for a significance level $\alpha = 1\%$, show with the paired-sample Wilcoxon-test, that our based $O_{\mathcal{M}\mathcal{G}\mathcal{B}}$ document indexing approach is statistically significant (*p-value* $< 0.01\%$) compared to the baseline BM25. Thus, we conclude that conceptual indexing by an enriched ontology with irredundant association rules between terms, would significantly improve the biomedical IR performance.

7 CONCLUSIONS

The work developed in this paper lies within the scope of domain ontologies enrichment and their application in IR field. We have introduced an automatic enrichment approach based on a generic basis of association rules between terms (Latiri et al., 2012) to identify additional concepts linked to ontology concepts. Interestingly enough, these association rules are extracted from the target document collection by means of mining mechanisms which in turn rely on results from FCA field (Ganter and Wille, 1999). The placement of new concepts is carried out through the defined distance measures and the neighborhood concept. The result is a proxemic conceptual network where nodes represent disambiguated concepts and edges are materialized by the value of distance measure between concepts. Three semantic relations from this network are used, namely: a *referential semantic*, a *differential semantic* and an *inferential semantic*. To evaluate the contribution of the conceptual proxemic network in the retrieval effectiveness, we integrate it in a conceptual document indexing. In this regard, the conducted experiments using OHSUMED collection

Table 3: IR effectiveness (% change) over 63 queries.

Strategies	MAP	P@10	P@20
Baseline (BM25)	23.96	41.9	35.00
I_{Expert}	29.55 (**+23.33**)	45.08 (+7.59)	39.92 (+14.06)
I_{MeSH}	24.73 (+3.21)	41.27 (-1.50)	35.87 (+2.49)
$I_{\mathcal{MGB}}$	24.96 (**+4.17**)	42.77(+2.08)	36.08 (+3.09)
$I_{O_{\mathcal{MGB}}}$	28.17 (**+17.57**)	44.33 (+5.80)	38.17 (+10.86)

and MeSH ontology which highlighted an improvement in the performances of the information retrieval system, in terms of both recall and precision metrics. As work in progress, we focus on enrichment of multilingual ontologies by means of inter-lingual association rules between terms introduced in (Latiri et al., 2010).

ACKNOWLEDGEMENTS

This work was partially supported by the French-Tunisian project CMCU-UTIQUE 11G1417.

REFERENCES

Agrawal, R. and Skirant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the* 20th *International Conference on Very Large Databases (VLDB 1994)*, pages 478–499, Santiago, Chile.

Amirouche, F. B., Boughanem, M., and Tamine, L. (2008). Exploiting association rules and ontology for semantic document indexing. In *Proceedings of the* 12th *International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU'08)*, pages 464–472, Malaga, Espagne.

Andreasen, T., Bulskov, H., Jensen, P., and Lassen, T. (2009). Conceptual indexing of text using ontologies and lexical resources. In *Proccedings of the* 8th *International Conference on Flexible Query Answering Systems, FQAS 2009*, volume 5822 of *LNCS*, pages 323–332, Roskilde, Denmark. Springer.

Balcázar, J. L. (2010). Redundancy, deduction schemes, and minimum-size bases for association rules. *Logical Methods in Computer Science*, 6(2):1–33.

Bastide, Y., Pasquier, N., Taouil, R., Stumme, G., and Lakhal, L. (2000). Mining minimal non-redundant association rules using frequent closed itemsets. In *Proceedings of the* 1st *International Conference on Computational Logic*, volume 1861 of *LNAI*, pages 972–986, London, UK. Springer.

Baziz, M., Boughanem, M., Aussenac-Gilles, N., and Chrisment, C. (2005). Semantic cores for representing documents in IR. In *Proceedings of the 2005 ACM Symposium on Applied Computing*, SAC'05, pages 1011–1017, New York, USA. ACM Press.

Ben Yahia, S., Gasmi, G., and Nguifo, E. M. (2009). A new generic basis of factual and implicative association rules. *Intelligent Data Analysis*, 13(4):633–656.

Bendaoud, R., Napoli, A., and Toussaint, Y. (2008). Formal concept analysis: A unified framework for building and refining ontologies. In *Proceedings of* 16th *International Conference on the Knowledge Engineering: Practice and Patterns (EKAW 2008)*, volume 5268 of *LNCS*, pages 156–171, Acitrezza, Italy. Springer.

Benz, D., Hotho, A., and Stumme, G. (2010). Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge. In *Proceedings of the* 2nd *Web Science Conference (WebSci10)*, Raleigh, NC, USA.

Cimiano, P., Hotho, A., Stumme, G., and Tane, J. (2004). Conceptual knowledge processing with formal concept analysis and ontologies. In *Proceedings of the second International Conference on Formal Concept Analysis, ICFCA 2004*, pages 189–207, Sydney, Australia.

Di-Jorio, L., Bringay, S., Fiot, C., Laurent, A., and Teisseire, M. (2008). Sequential patterns for maintaining ontologies over time. In *Proceedings of the International Conference On the Move to Meaningful Internet Systems, OTM 2008*, volume 5332 of *LNCS*, pages 1385–1403, Monterrey, Mexico. Springer.

Díaz-Galiano, M. C., García-Cumbreras, M. A., Martín-Valdivia, M. T., Montejo-Ráez, A., and na López, L. A. U. (2008). Integrating MeSH Ontology to Improve Medical Information Retrieval. In *Proceedings of the* 8th *Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Advances in Multilingual and Multimodal Information Retrieval*, volume 5152 of *LNCS*, pages 601–606, Budapest, Hungary. Springer.

Dinh, D. and Tamine, L. (2011). Combining global and local semantic contexts for improving biomedical information retrieval. In *Proceedings of the* 33rd *European Conference on IR Research, ECIR 2011*, volume 6611 of *LNCS*, pages 375–386, Dublin, Ireland. Springer.

Faatz, A. and Steinmetz, R. (2002). Ontology enrichment with texts from the www. In *Proceedings of the* 2nd *ECML/PKDD-Workshop on Semantic Web Mining*, pages 20–34, Helsinki, Finland.

Ganter, B. and Wille, R. (1999). *Formal Concept Analysis*. Springer.

Jones, K. S., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management*, 36(6):779–840.

Latiri, C., Haddad, H., and Hamrouni, T. (2012). To-

wards an effective automatic query expansion process using an association rule mining approach. *Journal of Intelligent Information Systems*, pages DOI: 10.1007/s10844–011–0189–9.

Latiri, C., Smali, K., Lavecchia, C., and Langlois, D. (2010). Mining monolingual and bilingual corpora. *Intelligent Data Analysis*, 14(6):663–682.

Maedche, A., Pekar, V., and Staab, S. (2002). *Ontology Learning Part One - On Discovering Taxonomic Relations from the Web*, pages 301–322. Springer.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41:1–69.

Navigli, R. and Velardi, P. (2006). Ontology enrichment through automatic semantic annotation of online glossaries. In *Proceedings of* 15^{th} *International Conference, EKAW 2006, Podebrady, Czech Republic*, volume 4248 of *LNCS*, pages 126–140. Springer.

Neshatian, K. and Hejazi, M. R. (2004). Text categorization and classification in terms of multiattribute concepts for enriching existing ontologies. In *Proceedings of the* 2^{nd} *Workshop on Information Technology and its Disciplines, WITID'04*, pages 43–48, Kish Island, Iran.

Parekh, V., Gwo, J., and Finin, T. W. (2004). Mining domain specific texts and glossaries to evaluate and enrich domain ontologies. In *Proceedings of the International Conference on Information and Knowledge Engineering, IKE'04*, pages 533–540, Las Vegas, Nevada, USA. CSREA Press.

Salton, G. and Buckely, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

Smucker, M. D., Allan, J., and Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the* 16^{th} *International Conference on Information and Knowledge Management, CIKM 2007*, pages 623–632, Lisboa, Portugal. ACM Press,.

Song, M., Song, I., Hu, X., and Allen, R. B. (2007). Integration of association rules and ontologies for semantic query expansion. *Data and Knowledge Engineering*, 63(1):63 – 75.

Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., and Lakhal, L. (2002). Computing Iceberg Concept Lattices with Titanic. *Journal on Knowledge and Data Engineering*, 2(42):189–222.

Valarakos, A., Paliouras, G., Karkaletsis, V., and Vouros, G. (2004). A name-matching algorithm for supporting ontology enrichment. In Vouros, G. and Panayiotopoulos, T., editors, *Methods and Applications of Artificial Intelligence*, volume 3025 of *LNCS*, pages 381–389. Springer.

Vallet, D., Fernndez, M., and Castells, P. (2005). An ontology-based information retrieval model. In *The Semantic Web: Research and Applications*, volume 3532 of *LNCS*, pages 103–110. Springer.

Wu, Z. and Palmer, M. (1994). Verb semantics and lexical selection. In *Proceedings of the* 32^{nd} *annual meeting of the Association for Computational Linguistics*, pages 133–138, New Mexico, USA.

Zaki, M. J. (2004). Mining non-redundant association rules. *Data Mining and Knowledge Discovery*, 9(3):223–248.