Exploiting association rules and ontology for semantic document indexing

Fatiha Boubekeur-Amirouche, Mohand Boughanem, Lynda Tamine

Exploiting association rules and ontology for semantic document indexing

Fatiha Boubekeur
IRIT-SIG, Paul Sabatier University of Toulouse, 31062 CEDEX 9, France
Department of Computer Sciences, Mouloud Mammeri University of Tizi-Ouzou, 15000, Algeria
boubekeu@irit.fr

Mohand Boughanem
IRIT-SIG, Paul Sabatier University of Toulouse, 31062 CEDEX 9, France
boughane@irit.fr

Lynda Tamine-Lechani
IRIT-SIG, Paul Sabatier University of Toulouse, 31062 CEDEX 9, France
Lynda.Lechani@irit.fr

Abstract

This paper describes a novel approach for document indexing based on the discovery of contextual semantic relations between concepts. The concepts are first extracted from WordNet ontology. Then we propose to extend and to use the association rules technique in order to discover conditional relations between concepts. Finally, concepts and related contextual relations are organized into a conditional graph.

Keywords Information retrieval, Conceptual indexing, Association rules, WordNet.

1 Introduction

Information retrieval (IR) is concerned with selecting, from a collection of documents, those that are likely to be relevant to a user information need expressed using a query. Three basic functions are carried out in an information retrieval system (IRS): document and information needs representation and matching of these representations [6]. Document representation is usually called indexing. The main objective of indexing is to assign to each document a descriptor represented with a set of features, usually keywords, derived from the document content. Representing the user's information need involves a one step or multi-step query formulation by means of prior terms expressed by the user and/or additive information driven by iterative query improvements like relevance feedback [17]. The main goal of document-query matching, called also query evaluation, is to estimate a relevance score used as a criterion to rank the list of documents returned to the user as an answer to his query. Most of IR models handle during this step, an approximate matching process using the frequency distribution of query terms over the documents.

Numerous factors affect the IRS performances. First, information sources might contain various topics addressed with a very wide vocabulary that lead to different semantic contexts. This consequently increase the difficulty of identifying the accurate semantic context addressed by the user's query. Second, users often do not express their queries in an optimal form considering their needs. Third, in classical IR models, documents and queries are basically represented as bags of weighted terms. A key characteristic of such models is that the degree of document query matching depends on the number of the shared terms. This leads to a "lexical focused" relevance estimation which is less effective than a "conceptual focused" one [14]. This paper addresses these limitations by proposing a conceptual indexing approach based on the joint use of ontology and association

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 464–472
Torremolinos (Málaga), June 22–27, 2008

rules. We thus expect to take advantages from (1) conceptual indexing of documents and (2) from the contextual dependencies between concepts discovered by means of association rules. The paper is structured as follows: Section 2 introduces the problems we aim to tackle, namely the term mismatch and ambiguity in IR, then reports some related works and presents our motivations. The proposed semantic indexing approach is detailed in section 3. Section 4 concludes the paper.

2 Related works and motivations

Most of classical IR models are based on the well known technique of "bag of words" expressing the fact that both documents and queries are represented using basic weighted keywords. The performances of such models suffer from the so-called *keyword barrier* [20] that has serious drawbacks especially at document representation and query formulation levels that lead to bad performances. Indeed, in such IRS, a relevant document will not be retrieved in response to a query if the document and query representations do not share at least one word. This implies on one hand, that relevant documents are not retrieved if they do not share terms with the query; on the other hand, irrelevant documents that have common words with the query are retrieved even if these words are not semantically equivalent. Various approaches and techniques attempted to tackle these problems by enhancing the document representation or query formulation. Attempts in document representation improvements are related to the use of semantics in the indexing process. In this context, two main issues could be distinguished [3]: semantic indexing and conceptual indexing. Semantic indexing is based on techniques for contextual word sense disambiguation (*WSD*) [10], [13], [18], [21]. Indexing consists in associating the extracted words of document or query, to words of their own context [21]. Other disambiguation approaches [20] use hierarchical representations driven from ontologies to compute the *semantic distance* or *semantic similarity* [11], [12], [16] between words to be compared. Conceptual indexing is based on using concepts extracted from ontologies and taxonomies in order to index documents instead of using simple lists of words [3], [7], [8], [9], [19]. The indexing process runs mainly in two steps: first identifying terms in the document text by matching document phrases with concepts in the ontology, then, disambiguating the identified terms.

In [15] the FIS-CRM model focusses on the interrelations (synonymy and generality ones) among the document terms in order to build the related conceptual index. A concept is represented by the related document term, the set of its fuzzy synonymous (extracted from a fuzzy dictionary), and its more general concepts (extracted from a fuzzy ontology). A weight readjustment process allows the added terms to have fuzzy weights even if they do not occur in the document. Document clustering is then involved using concept coocurrence measures.

We aim to tackle the problems due to using basic keyword based evidence sources in IR, providing a solution at the indexing level. The proposed approach relies on the joint use of concepts and contextual relations between them in order to index the document. Both concepts and related dependencies are finally organized into a graphical representation resulting in a graphical representation of the document index. Our approach allows a richer and more accurate representation of documents when supporting both contextual and semantical relations between concepts. Our main propositions presented in this paper concern a novel conceptual document indexing approach based on concepts and contextual dependencies between them. This approach is based on the use of the WordNet general ontology as a source for extracting representative document concepts, and on association rules based techniques in order to discover the latent concept contextual relations. More precisely, our conceptual indexing approach is supported by three main steps: (1) Identifying the representative concepts in a document, (2) Discovering context-dependent relations between semantical entities namely the concepts (rather than between lexical entities that are terms) using a proposed variant of association rules namely semantic association

rules, and (3) combining both concepts and related semantic associations into a compact graphical representation.

3 A conceptual document indexing approach

This section details our conceptual document indexing approach based on (1) the exploitation of WordNet for identifying concepts, (2) association rules for discovering contextual relations between concepts. Both concepts and related relations are finally organized into a compact conditional graph.

3.1 Outline

We propose to use WordNet ontology and association rules in order to build the document conceptual index. The document indexing process is handled through three main steps:

(1) *Identifying the representative concepts of a document*: index concepts are identified in the document using WordNet.

(2) *Mining association rules*: contextual relations between selected concepts are discovered using association rules.

(3) *Building a graphical conceptual index*: both concepts and related relations are then organized into a graph. Nodes in the graph are concepts and edges traduce relations between concepts.

The above steps are detailed in the following.

3.2 Index concepts identification

3.2.1 Overview

The whole process of identifying the representative concepts of the document has been well described in [4], we summarize it in the following. The process relies on the following steps:

(1) *Term identification*: the goal of this step is to identify significative terms in a document. These terms correspond to entries in the ontology.

(2) *Term weighting*: in this step, an alternative of *tf*idf* weighting scheme is proposed. The underlying goal is to eliminate the less frequent (less important) terms in a document in order to select only the most representative ones (index terms).

(3) *Disambiguation*: index terms are associated with their corresponding concepts in the ontology. As each extracted term could have many senses (related to different concepts), we disambiguate it using similarity measures. A score is thus assigned to each concept based on its semantic distance to other concepts in the document. The selected concepts are those having the best scores.

3.2.2 Basic Notions

Terms are represented as lists of word strings (a word string is a character string that represents a word). The length of a term t, noted $|t|$, is then the number of words in t. A mono-word term consists in a one word list. A multi-word term consists in a word list of more than one word. Let t be a term represented as a list of words, $t = [w_1, w_2, ..., w_i, ., w_l]$. Elements in t could be identical, representing different *occurrences* of a same word in t. We note w_i the i^{th} word in t. We recursively define the position of the word w_i in the list t by the following:

$$\begin{cases} pos_t(w_1) = 1 \\ pos_t(w_i) = pos_t(w_{i-1}) + 1, \forall i = 1..l \end{cases}$$

Let $t_1 = [w_1, w_2, ..., w_m]$, $t_2 = [y_1, y_2, ..., y_l]$ be two given terms.

Definition 1 t_2 is a sub-term of t_1 if the whole sequence of words in t_2 occurs in t_1. t_1 is then called sur-term of t_2.

3.2.3 Identification of index terms

Before any document processing, in particular before pruning stop words, an important process for the next steps consists in extracting mono-word and multi-word terms from texts that correspond to entries in WordNet. The technique we propose performs a word by word analysis of the document. It is described in the following:

Let w_i be the next word (assumed not to be a stop word), to analyze in the document d. We extract from WordNet, the set S of terms containing w_i: let $S = \{C^i_1, C^i_2, \ldots C^i_m\}$, S is composed of multi-word and mono-word terms. S is ranked in decreasing order of term length as follows: $S = \{C^i_{(1)}, C^i_{(2)}, \ldots C^i_{(m)}\}$ where $(j)= (1) \ldots (m)$ is an index permutation such as $|C^i_{(1)}| \geq |C^i_{(2)}| \geq \ldots \geq |C^i_{(m)}|$. Terms with identical sizes are indifferently placed one beside another. For each element $C^i_{(j)}$ in S, we note $Pos_{C^i_{(j)}}(w_i)$ the position of w_i in the $C^i_{(j)}$ list of words. There are

$\left(pos_{C^i_{(j)}}(w_i) - 1\right)$ words on the left of w_i in $C^i_{(j)}$.

Let $Pos_d(w_i)$ be the position of w_i in d list of words.

Definition 2 The relative context of w_i *occurrence* in document d given the term $C^i_{(j)}$, is the word list CH^i_j defined by: $CH^i_j = sub(d, p, l)$ where:

$$p = pos_d(w_i) - \left(pos_{C^i_{(j)}}(w_i) - 1\right) \text{ and } l = |C^i_{(j)}|.$$

We so extract the relative context of w_i in d, namely $CH^i_j = sub(d, p, l)$ (c.f. Fig. 1), then we compare word string lists CH^i_j and $C^i_{(j)}$. .

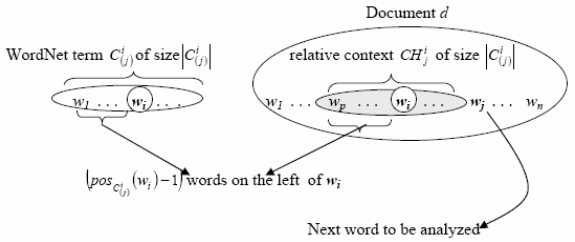

Figure 1: Identifying word context in d

If $CH^i_j \neq C^i_{(j)}$, the $C^i_{(j+1)}$ term of S is analyzed, otherwise, term $t_k = C^i_{(j)}$ is identified. The next word to analyze in d is w_j such that $pos_d(w_j) = p + l$.

During the identification process, three cases can happen as shown in Fig. 2:

Case a) Current identified term t_k is completely disjoint from t_{k-1}. It could be identical but we do not treat identities at this level. It is thus a new term which will be retained in document description.

Case b) Term t_k covers partially term t_{k-1}. The two terms are thus different and both identified even if they have common words.

Case c) Term t_k covers completely one or more preceding adjacent terms $t_{k-1} \ldots t_j$, $j \leq k-1$. In this case, to allow an effective disambiguation, we retain the longest term, namely t_k, and eliminate the adjacent terms it covers ($t_{k-1} \ldots t_j$, $j \leq k-1$).

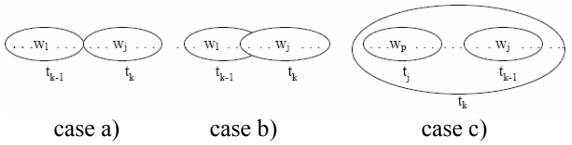

Figure 2: Term identification

By this step, we will have identified the set $T(d)$ of all the multi-word or mono-word terms that compose d: $T(d) = \{t_1, t_2, .., t_n\}$. We finally compute each term frequency in d and eliminate redundant terms, which leads to: $T'(d) = \{(t_1, Occ_1), (t_2, Occ_2), \ldots, (t_m, Occ_m) / t_i \in d, Occ_i = Occ(t_i)$ is the occurrence frequency of t_i in d, $1 \leq i \leq m$ and $m \leq n\}$.

3.2.4 Term weighting

Term weighting assignes to each term a weight that reflects its importance in a document. In the case of mono-word terms, variants of the *tf*idf* formula are used, expressed as follows: $W_{t,d} = tf(t)*idf(t)$, *tf* is term frequency, *idf* is the inverse document frequency such as $idf(t) = \log\left(\frac{N}{df(t)}\right)$, N is the number of documents in the corpus and $df(t)$ the number of documents in the corpus that contain the term t. In the case of multi-word terms, term weighting approaches use in general statistical and/or syntactical analysis. Roughly, they add single word frequencies or multiply the number of term occurrences by the number of single words belonging to this term.

We propose a new weighting formula as variant of *tf*idf* defined as follows: let $W_{t,d}$ be the weight associated with the term t in the document d, $T'(d)$ the term set of d, $Sub_j(t) \in T'(d)$ a sub-term of t, $Sur_i(t) \in T'(d)$ a sur-term of t, $S(t)$ the set of synsets C associated with term t. We define the probability that t is a possible sense of $Sub_j(t)$ as follows:

$$P\left(t \in S\left(Sub_j(t)\right)\right) = \frac{|\{C \in S\left(Sub_j(t)\right) / t \in C\}|}{|S\left(Sub_j(t)\right)|}$$

We then define $W_{t,d} = tf(t) * idf(t)$ as follows:

$$W_{t,d} = \begin{pmatrix} Occ(t) \\ +\sum_i Occ(Sur_i(t)) \\ \sum_j \left[P\left(t \in S\left(Sub_j(t)\right)\right) * Occ\left(Sub_j(t)\right)\right] \end{pmatrix} * \ln\left(\frac{N}{df(t)}\right)$$

Where: N is the total number of documents in the corpus, $df(t)$ (document frequency) is the number of documents in the corpus that contain the term t .

The underlying idea is that the global frequency of a term in a document is quantified on the basis of both its own frequency and its frequency within each of its sur-terms as well as its probable frequency within its sub-terms. Document index, *Index(d)*, is then built by keeping only those terms whose weights are greater than a fixed threshold.

3.2.5 Term disambiguation

Each term t_i in *Index(d)* may have a number of related senses (i.e. WordNet synsets). *Let* $S_i = \{C_1^i, ..., C_j^i, . . ., C_n^i\}$ be the set of all synsets associated with term t_i . Thus, t_i has $|S_i| = n$ senses. We believe that each index term contributes to the semantic content representation of d with only one sense (even if that is somewhat erroneous, since a term can have different senses in the same document, but we consider here the only dominating sense). From where, we must choose, for each term belonging to the document index ($t_i \in Index(d)$), its best sense in d . This is term disambiguation.

Our disambiguation approach relies on the computation of a score (*C_Score*) for every concept (sense) associated with an index term. That is, for a term t_i, the score of its j^{th} sense, noted C_j^i, is computed as:

$$Score\left(C_j^i\right) = \sum_{\substack{l \in [1,..,m] \\ l \neq i}} \sum_{1 \leq k \leq nl} Wc_{j,d}^i * Wc_{k,d}^l * Dist\left(C_j^i, C_k^l\right)$$

Where m is the number of concepts from *Index(d)*, n_l represents the number of WordNet senses which is proper to each term t_l, $Dist\left(C_j^i, C_k^l\right)$ is a semantic relatedness measure between concepts C_j^i and C_k^l as defined in [11], [12], [16] and $W_{C_j^i,d}$ is the weight associated with concept C_j^i, formally defined as follows:

$$\forall C_j^i \in S_i, W_{C_j^i,d} = W_{t_i,d}$$

The concept-sense which maximizes the score is then retained as the best sense of term t_i. The underlying idea is that the best sense for a term t_i in the document d may be the one strongly correlated with senses associated with other important terms in d.

The set $N(d)$ of the selected senses represents the semantic core of the document d.

3.3 Discovering associations between concepts

Association rules introduced in [1] aim to generate all the significant associations between items in a transactional database. In our context, association rules are used in order to discover significant associations between concepts. The formal model is defined in the following:

Let $N(d) = \{C_1, C_2, ..., C_j, ...\}$ be the semantic core of the document d. Each concept C_i is represented by the term it refers to in the document. The set of its synonymous defines its value domain $Dom(C_i) = \{C_1^i, C_2^i, C_3^i, ...\}$. Each value in $Dom(C_i)$, is a concept $C_j^i \in N(d)$ such that C_j^i and C_i are synonymous.

Thus, let $\eta(d) = \{(C_i, Dom(C_i))\}$, we simply note $\eta(d) = \{(X, Dom(X))\}$ the semantic core of document d. We propose to use association rules mining to discover *latent* (*hidden*) contextual relations between the index concepts. Concepts are semantic entities. As association rules allow discovering relations between lexical entities, namely terms, we propose to extend them in order to support semantic entity associations (namely concept associations). The proposed extension is a revised version of the one defined in [4].

Definition 3 A *semantic association rule* between $X \text{ and } Y \in \eta(d)$, we note $X \rightarrow_{sem} Y$, is defined as follows:

$$X \rightarrow_{sem} Y \Leftrightarrow \begin{cases} \exists X_i \in Dom(X), \exists Y_j \in Dom(Y) / \\ X_i \rightarrow Y_j \end{cases}$$

Such that $X_i \rightarrow Y_j$ is an association between terms (frequent 1-itemsets) X_i and Y_j.

The intuitive meaning of $X \rightarrow_{sem} Y$ rule is that if a document *is about* a concept *X*, it tends to also be *about* the concept *Y*. The *aboutness* of a document expresses the *topic focus* of its content. This interpretation also applies to the rule $X_i \rightarrow Y_j$. Thus, the rule $R: X_i \rightarrow Y_j$ expresses the probability that the document is about Y_j knowing that it is about X_i. The confidence associated with *R* thus rely on the importance degree of Y_j in a document *d*, knowing the importance degree of X_i in *d*. It is formally defined in the following.

Definition 4 The confidence of the rule $R: X_i \rightarrow Y_j$ is formally given by:

$$Confidence(R) = \frac{Support(X_i \text{ and } Y_j)}{Support(X_i)} = \frac{\min(W_{X_i,d}, W_{Y_j,d})}{W_{X_i,d}}$$

Definition 5 The confidence of a semantic association rule $R_{sem}: X \rightarrow_{sem} Y$ is defined as:

$$Confidence(X \rightarrow_{sem} Y) = \max_{i,j}\left(Confidence\left(\underset{X_i \in Dom(X), Y_j \in Dom(Y)}{R: X_i \rightarrow X_j} \right) \right)$$

Remark 1. $Confidence(X \rightarrow_{sem} Y)$ always equal 1.

In our context, the support of a semantic association rule $X \rightarrow_{sem} Y$ relies on the amount of individual association rules $X_i \rightarrow Y_j$ ($X_i \in Dom(X)$ and $Y_j \in Dom(Y)$, having confidence greater than the fixed minimum confidence threshold *minconf=1*. The support is formally defined in the following.

Definition 6 The support of the rule $R: X \rightarrow_{sem} Y$ is formally given by:

$$Support(R) = \frac{\left| \{ X_i \rightarrow Y_j / Confidence(X_i \rightarrow Y_j) \geq \min conf \} \right|}{\left| \{ X_i \rightarrow Y_j, (X_i, Y_j) \in Dom(X) \times Dom(Y) \} \right|}$$

We propose to discover relations between concepts in *η(d)* by means of semantic association rules. Semantic association rules are based, in our context, on the following principles: (1) a transaction is a document, (2) items are values from concept domains, (3) an itemset is a concept, (4) a semantic association rule $X \rightarrow_{sem} Y$ defines an implication (i.e. a conditional relation) between concepts *X* and *Y*. By using association rules, we aim to build a conditional hierarchical structure of the *topic focus* of the document content. That is to say, we aim at structuring concepts describing the document, in a conditional hierarchy which is supported by the semantics of extracted association rules.

The problem of discovering association rules between concepts is devided into two steps, following the principle by A-priori algorithm.. First, identifying all the frequent 1-itemsets, corresponding to individual concepts. A frequent concept is in our context, a concept that have a weight greater than a fixed minimum threshold. Second, mining association rules between the frequent itemsets. The objective of mining semantic association rules between concepts is to retain the only ones that have a support and a confidence greater than a fixed minimum threshold *minsup* and *minconf* respectively.

Some problems can occur when discovering association rules, such as redundancies and cycles. Redundant rules come generally from transitive properties: $X \rightarrow_{sem} Y$, $Y \rightarrow_{sem} Z$ and $X \rightarrow_{sem} Z$. In order to eliminate redundancy we propose to build the minimal covers of the set of extracted rules (that is the minimal set of non transitive rules). The existence of cycles in the graph would be due to the simultaneous discovery of association rules $X \rightarrow_{sem} Y$ and $Y \rightarrow_{sem} X$, or of association rules such as $X \rightarrow_{sem} Y$, $Y \rightarrow_{sem} Z$ and $Z \rightarrow_{sem} X$. To solve this problem, we eliminate the weakest rule

having the lowest support, among that leading to the cycle. If all the rule supports are equal, we randomly eliminate one rule in the cycle.

Finally, concepts and related association rules are structured into a network leading to a graphical index. The process of building the index graph is based on the following principles:

Nodes in the graph are are concepts from the semantic core of the document *d*, $\eta(d) = \{(C_i, Dom(C_i))\}$. Thus, nodes in the graph are represented by variables attached to the concepts C_i , instantiated in the value domain $Dom(C_i) = \{C_1^i, C_2^i, C_3^i, ...\}$.

Node relations are conditional dependencies discovered between concepts, by means of semantic association rules.

Each node *X* in the graph is annotated by a unconditional table named *T(X)* such that:

$$\forall X_i \in Dom(X),\ T(X_i) = W_{X_i,d} \qquad (1)$$

In a previous work [4], we used particularly CP-Nets [5] as the graphical model supporting this conceptual index.

3.4 Illustration

The document indexing process presented above is illustrated through the following example. Let *d((U.C[1].,0.5),(Metropolis,0.9), (Impoverishment, 0.1),(People,0.4),(Poorness, 0.7),…)* a document described by the given weighted concepts. *Metropolis,* and *U.C are synonymes of City*, thus *U.C* and *Metropolis* pertain to *City* concept node domain. Similarly, both of *Impoverishment* and *Poorness* pertain to *Poverty* concept node domain, whereas *People* is associated with *population* concept node. That is to say *η(d) = {(City, Dom (City))*, *(Poverty, Dom(Poverty))*, *(Population, Dom(Population))* |*Dom(City)= {Metropolis, U.C}*, *Dom(Poverty)= {Poorness,Impoverishment}*, *Dom(Population)= {People}*.

[1] *U.C. = Urban Center*

We aim to discover associations between *City, Population* and *Poverty* nodes. Applying Apriori algorithm [2] leads: (1) to extract frequent itemsets, then (2) to generate association rules between frequent 1- itemsets. Relations of interest are between individual concepts in the document (rather than between sets of concepts), thus we only have to compute the *k*-itemsets , *k*=1, 2. Assuming a minimum support threshold *minsup*= 0.1, the extracted frequent *k*-itemsets (*k*= 1, 2) are given in Table 1.

Support(*Impoverishment*) < *misup* : the 1-itemset *Impoverishment* is not frequent, it is then pruned. The extracted association rules are given in Table 2.

Table 1: Generating frequent *k*-itemsets

1-Itemsets	Itemset	Support
	Metropolis	0.9
	U.C	0.5
	Impoverishment	0.1
	Poorness	0.7
	People	0.4
Frequent 2-itemsets	Metropolis, People	0.4
	Metropolis, Poorness	0.7
	U.C, People	0.4
	U.C, Poorness	0.5
	People, Poorness	0.4

Table 2: Generated association rules

R_1: Metropolis → People	R_2: People →Metropolis
R_3: Metropolis →Poorness	R_4: Poorness→Metropolis
R_5: U.C →People	R_6: People → U.C
R_7: U.C → Poorness	R_8: Poorness → U.C
R_9: People →Poorness	R_{10}: Poorness → People

By applying the formula given in definition 4, generated rule confidences are computed leading to the results given in Table 3.

If we suppose a confidence minimum threshold *minconf* =1, we retain only rules whose confidences are equal or greater than *minconf*. The selected rules are shown in Table 4.

Table 3: Confidence rules

R_i	R_1	R_3	R_5	R_7	R_9
Confidence(R_i)	0.57	0.77	0.8	1	1
	R_2	R_4	R_6	R_8	R_{10}
	1	1	1	0.71	0.57

These rules are first used to build semantical association rules that in fact correspond to the index graph concept- node relations. Thus, we deduce:

(1) From R_2: People → Metropolis and R_6: People → U.C: *Population* $\rightarrow_{sem}$ *City*

(2) From R_4: Poorness → Metropolis : *Poverty* $\rightarrow_{sem}$ *City*

(3) From R_7: U.C → Poorness : *Cit y* $\rightarrow_{sem}$ *Poverty*

(4) From R_9: People →Poorness : *Population* $\rightarrow_{sem}$ *Poverty*

Table 4: Selected association rules

R_2: People → Metropolis
R_4: Poorness → Metropolis
R_6: People → U.C
R_7: U.C → Poorness
R_9: People →Poorness

We then compute the support of each semantic association rule. The results are given in Table 5.

Table 5: Semantic association rules supports

Population $\rightarrow_{sem}$ City	1
Poverty $\rightarrow_{sem}$ City	0.5
City $\rightarrow_{sem}$ Poverty	0.5
Population $\rightarrow_{sem}$ Poverty	1

We obviously retain these rules that have a support equal to 1. Two associations exist between concepts *City* and *Poverty* , with the same support. We thus randomly keep one of them. Suppose *Poverty* $\rightarrow_{sem}$ *City* is retained. The set of selected semantic association rules is highlighted in Table 5. Clearly, retaining the three rules will lead to a cycle in the index graph. To avoid this, we prune the weakest rule (that is the one with lowest support) *Poverty* $\rightarrow_{sem}$ *City*. Finally, the only selected semantic rules are the following: *Population* $\rightarrow_{sem}$ *City* and *Population* $\rightarrow_{sem}$ *Poverty*. *CPT's* are then associated with concept nodes *Population*, *City* and *Poverty* respectively, using formula 1, which leads to the graphical document index given in Fig. 3.

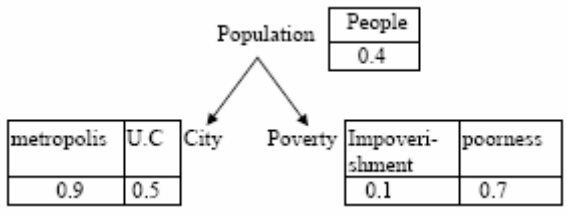

Figure 3: A graphical document index

4 Conclusion

We described in this paper a novel approach for conceptual document indexing. The approach is founded on the joint use of both ontology for identifying, weighting and disambiguating terms, and association rules to derive context dependent relations (the context here refers to the document content) between terms leading to a more expressive document representation. The approach foundation is not new but we proposed new techniques for identifying, weighting and disambiguating terms and for discovering relations between related concepts by means of the proposed semantic association rules. Semantic association rules allow to derive context dependent relations between concepts leading to a more expressive document representation.

The work is still in progress and results from a comparative study with other existing conceptual indexing approaches will be soon available.

References

[1] R.Agrawal, T. Imielinski, and A. Swami (1993). Mining association rules between sets of items in large databases. In *Proceedings of the ACMSIGMOD Conference on Management of data*, pages 207-216, Washington, USA, 1993.

[2] R. Agrawal and R. Srikant (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th VLdB Conference*, pages 487-499.

[3] M. Baziz, M.Boughanem, N.Aussenac-Gilles N. and C. Chrisment (2005). Semantic Cores for Representing document in IR. In *SAC'2005- 20th ACM Symposium on Applied Computing*, pages 1011-1017. Santa Fe, New Mexico, USA, 2005 .

[4] F. Boubekeur, M. Boughanem and L. Tamine-Lechani (2007). Semantic Information Retrieval Based on CP-Nets, in *IEEE International Conference on Fuzzy Systems,* London, U.K, july 2007.

[5] C. Boutilier, R. Brafman, H. Hoos and D. Poole (1999). Reasoning with Conditional Ceteris Paribus Preference Statements. In *Proceedings of UAI-1999*, pages 71-80.

[6] W.B. Croft (1993). Knowledge-based and statistical approaches to text retrieval. In *IEEE Expert* , 8(2), pages 8-12.

[7] N. Guarino, C. Masolo and G. Vetere (1999). OntoSeek: Using Large Linguistic Ontologies for Accessing On-Line Yellow Pages and Product Catalogs. *National Research Council, LADSEBCNR*, Padavo, Italy, 1999.

[8] J. Gonzalo, F. Verdejo, I. Chugur, J. Cigarran. (1998). Indexing with WordNet synsets can improve retrieval. *In proceedings of COLING/ACL Work*, 1998.

[9] L.R Khan (2000). Ontology-based Information Selection. *Phd Thesis, Faculty of the Graduate School, University of Southern California*.

[10] R. Krovetz and W.B. Croft (1992). Lexical Ambiguity and Information Retrieval. In *ACM Transactions on Information Systems*, 10(1).

[11] C. Leacock, G.A. Miller and M. Chodorow (1998*). Using corpus statistics and WordNet relations for sense identification*. *Computational Linguist*ics 24(1,) pages 147-165.

[12] D. Lin (1998). An information-theoretic definition of similarity. In *Proceedings of 15th International Conference On Machine Learning*.

[13] R. Mihalcea, and D. Moldovan (2000). Semantic indexing using WordNet senses. In *Proceedings of ACL Workshop on IR and NLP*, HongKong.

[14] M. Mauldin, J. Carbonell and R. Thomason (1987). Beyond the keyword bariier: knowledge-based information retrieval. *Information services and use* 7(4-5), pages 103-117.

[15] J.A. Olivas, P.J. Garcés and F.P. Romero (2003). An application of the FIS-CRM model to the FISS metasearcher: Using fuzzy synonymy and fuzzy generality for representing concepts in documents. *International Journal of Approximate Reasoning,* Volume 34, Issues 2-3, pages 201-219, November 2003.

[16] P. Resnik (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research (JAIR)*, 11. pages 95-130.

[17] J.J. Rocchio (1971). Relevance feedback in information retrieval. *In The SMART Retrieval System-experiments in Automatic Document Processing*, G.Salton editor, Prentice-Hall, Englewood Cliffs, NJ, pages 313-323.

[18] M. Sanderson (1994). Word sense disambiguation and information retrieval. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and development in Information Retrieval*, pages 142-151.

[19] M.A. Stairmand and J.B. William (1996). Conceptual and Contextual Indexing of documents using WordNet-derived Lexical Chains. In *Proceedings of 18th BCS-IRSG Annual Colloquium on Information Retrieval Research.*

[20] E.M. Voorhees (1993). Using WordNet to disambiguate Word Senses for Text Retrieval. In *Proceedings of the 16thAnnual Conference on Research and development in Information Retrieval, SIGIR'93*, Pittsburgh, PA.

[21] D. Yarowsky (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting*, *Association for Computational Linguistics,* pages 189-196, Cambridge, Massachusetts, USA.