Semantic Information Retrieval Based on CP-Nets

Fatiha Boubekeur, Mohand Boughanem and Lynda Tamine-Lechani

Abstract—**This paper describes a flexible information retrieval approach based on Conditional Preferences Networks (CP-Nets). Our contribution focuses on document indexing and query evaluation using CP-NET theoretical foundations. More precisely, we propose first a conceptual document indexing approach using WordNet for identifying concepts and association rules for discovering contextual relations between concepts. Secondly, we propose an algorithm for query evaluation using graph similartity measures.**

I. INTRODUCTION

The main goal of an information retrieval system (IRS) is to find the information assumed to be relevant to a user query which is generally expressed by a set of keywords connected with Boolean operators. Weights are expressed in the query (as in [1]-[5]) in order to enable flexible query formulation and consequently flexible query evaluation. A key characteristic of such systems is that the degree of document query matching depends on the number of the corresponding shared terms. It is well known that a query is usually an incomplete and vague description of the user information need and authors of documents use a very wide vocabulary to express the same concepts. This leads to the problem of term mismatch and ambiguity in IR.

This paper addresses these problems proposing a semantic information retrieval approach based on CP-Nets. The CP-Net formalism is used both for document and query representation and for flexible query evaluation. We propose to use the CP-Net formalism for three main reasons: (1) to exploit an intuitive graphical formalism for representing flexible queries, (2) to perform a conceptual indexing of documents, (3) to allow a flexible query evaluation using a graph similarity measure between document and query CP-Nets.

The paper is structured as follows: In section II, we introduce the problems we aim to tackle, namely the term mismatch and ambiguity in IR, then we report some related works and summarize our contribution. In section III, we introduce backgrounds in CP-Nets and association rules. The proposed CP-Net based IR approach is detailed in section IV. Section V concludes the paper.

II. PROBLEM DEFINITION AND RELATED WORKS

Document and query representation is a hard and fundamental problem in IR. Most of classic IR models are still using the well known technique of "bag of words". However, this technique has limits. Indeed, in classic 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 the 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

There are a number of approaches that tackled these problems. In this context, two main issues can be distinguished: enhancing query formulation and document representation (indexing). Some new directions in document indexing are being recently undertaken: semantic indexing and conceptual indexing.

Semantic indexing is based on techniques for contextual word sense disambiguation (*WSD*) [6]-[12]. Indexing consists in associating the extracted words of document or query, to words of their own context [13]. Other more elaborate disambiguation approaches use hierarchical representations derived from ontologies to compute the *semantic distance* or *semantic similarity* between words to be compared [14]-[18].

Conceptual indexing is based on using concepts extracted from ontologies and taxonomies in order to index documents instead of using simple words [15], [19]-[21]. The indexing process runs mainly in two steps: first identifying multiword terms (that are generally noun phrases) in the document text [19], [22], [23], then matching the identified multiword terms with concepts in the ontology.

Improvemnts in query formulation are achieved by means of query structure refinement or query expansion.

Identifying the suitable query structure (words and related weights) is not an easy task for a user, particularly when the query contains conditional preferences [24]. Random or intuitive term weighting of a qualitative query leads to erroneous results from that expected for the preferences it attempts to express. Some works in IR tackled this problem using more intuitive qualitative preferences expressed by means of linguistic symbols such as: *important, very important*…., [2],[3], or using preference relation "$\succ$" as in [24].

Attempts in query expansion use term relations in order to improve effectiveness of IRS. More precisely, query expansion is used to improve *recall* in IR. The IRS will retrieve not only the documents that contain the query words but also documents containing words that are related to the query words. The synonymy relation is most often used to expand query words, but also other types of relations are exploited. Term relations are derived from ontology as in [14], [15], [20], [21], [25], [26] or discovered, in document context by means of association rules as in [27]-[30].

Fatiha Boubekeur is with the Research Institute in Computer Sciences of Toulouse (IRIT), Paul Sabatier University, 31062 Toulouse CEDEX 9 , France(email: boubekeu@irit.fr), and with the Department of Computer Sciences, Mouloud Mammeri University of Tizi-Ouzou, 15000, Algeria.

Mohand Boughanem and Lynda Tamine-Lechani are with the Research Institute in Computer Sciences of Toulouse (IRIT), Paul Sabatier University, 31062 Toulouse CEDEX 9, France(emails: boughane@irit.fr; Lynda.Lechani@irit.fr).

1-4244-1210-2/07/$25.00 ©2007 IEEE.

A. *Our Contribution*

Our main propositions presented in this paper are the following:

(1) A document CP-Net based indexing approach. The CP-Net nodes represent concept classes. The relations between nodes express the contextual dependencies between the related concepts. This approach is based on the use of WordNet general ontology as a source for extracting representative document concepts, and on association rules technique in order to discover the latent concept contextual relations.

(2) A flexible query evaluation approach based on a graph similarity measure between document and query CP-Nets.

III. PRELIMINARIES

A. *CP-Nets*

CP-Nets were introduced in 1999 [31] as graphical models for compact representation of qualitative preference relations. A CP-Net is a Directed Acyclic Graph, or *DAG*, $G= (V, E)$, where V is a set of nodes $X_1, X_2, .., X_n$, that represent the preference variables and E a set of directed arcs expressing preferential dependencies between them. Each variable X_i takes values in the set $Dom(X_i) = \{X_i^1, X_i^2, X_i^3, \ldots\}$. We denote by $Pa(X_i)$ the parent set of X_i in G. representing its predecessors in the graph.

To each variable X_i of the CP-Net, a conditional preference table ($CPT(X_i)$) is attached, specifying for each value of $Pa(X_i)$ a total preference order among $Dom(X_i)$ values. For a root node of the CP-Net, the *CPT* simply specifies an unconditional preference order on its values.

A UCP-Net [32] extends a CP-Net by quantifying the CP-Net nodes with conditional utility values (utility factors). A conditional utility factor $f_i(X_i ,Pa(X_i))$ (we simply write $f_i(X_i)$), is a real value attached to each value X_i given an instantiation of its parents $Pa(X_i)$, in order to express its conditional preference order.

B. *Association Rules*

The concept of extracting association rules was introduced in [33] as follows: given a large database of customer transactions, each transaction consists on items purchased by a customer in a visit. The goal of association rules is to generate all significant associations between items in the database. Formally [33], [34], let $I = \{I_1, \ldots, I_m\}$ be a set of m articles called *items*, $B= \{t_1, \ldots, t_n\}$ a set of transactions where each transaction t_i is a subset of I: An association rule is an implication of the form $X \rightarrow Y$ between two itemsets (i.e. sets of items) $X, Y \subset I$ such as $X \cap Y = \phi$. The rule r: $X \rightarrow Y$ holds in the transaction set B with *confidence c* if c% of transactions in B containing X also contains Y. It has *support s* in the transaction set B if $s\%$ of transactions in B contains $X \cup Y$.

Given a set of transactions B, the problem of extracting association rules is to generate all association rules that have support and confidence greater than a user-specified minimum support (called *minsup)* and minimum confidence (called *minconf*) respectively. This problem is supported by many existing algorithms, the most known is *Apriori* algorithm [34].

Association rules have been exploited [27]-[30] in order to extract term associations and then achieve query expansion.

IV. FLEXIBLE CP-NET-BASED INFORMATION RETRIEVAL

We describe in this section our flexible information retrieval approach based on CP-Nets. We show first of all how to use CP-Nets for expressing user qualitative queries, then for automatically weighting the related terms. We then present our approach for document CP-Net representation. Finally we present our CP-Net based flexible evaluation approach.

A. *Expressing Queries Using CP-Nets*

The user preferences are expressed using concepts represented by variables. Concepts are extracted from a controlled vocabulary formed by the whole of concepts belonging to document indexes. Each variable is defined on a domain of values (a value is therefore a query term). For each variable, the user must specify all of its preferential dependencies from which a CP-Net graph is built. Semantically, a CP-Net query defines a conjunction of disjunctions. Disjunctions link values of a same variable.

Figure 1 illustrates a CP-Net corresponding to a given user query. The variables of interest are V= *{City, Housing, Place}* where *Dom(City)={Paris, Lyon}*, *Dom(Housing) = {RH*[1]*, Studio}* and *Dom (Place) = {Center, Suburbs}.*

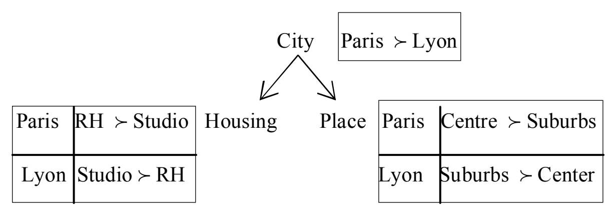

Fig.1. A CP-Net query

In addition, *CPT(City)* specifies that *Paris* is unconditionally preferable to *Lyon* (*Paris* $\succ$ *Lyon*), whereas *CPT(Housing)* for example, specifies a preference order on *Housing* values, under the condition of the *City* node values (thus for example, if *Paris* then *RH* $\succ$ *Studio*).

The qualitative CP-Net query is then automatically weighted leading to a UCP-Net corresponding to the original CP-Net query. The approach we proposed to automatically weight CP-Net queries is detailed in [24].

Using the proposed method, the UCP-Net related to the query presented in Figure 1, is given in Figure 2.

B. *Building the CP-Net Conceptual Document Index*

We propose to carry out a conceptual indexing method to index documents in the corpus. Conceptual indexing is based on two main features: concepts and relations between them.

[1] Residence Hall

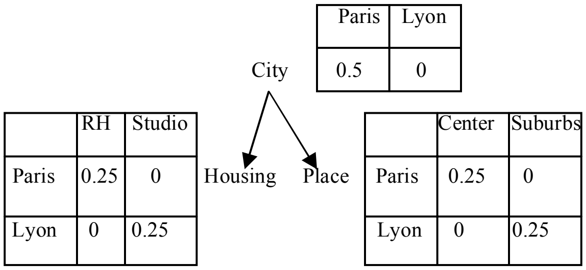

Fig.2. A UCP-Net query

Concepts are extracted using WordNet ontology. Concept relations are discovered by means of association rules. The goal is to build the document CP-Net graph. The document indexing process is handled through five main steps as presented in Figure 3.

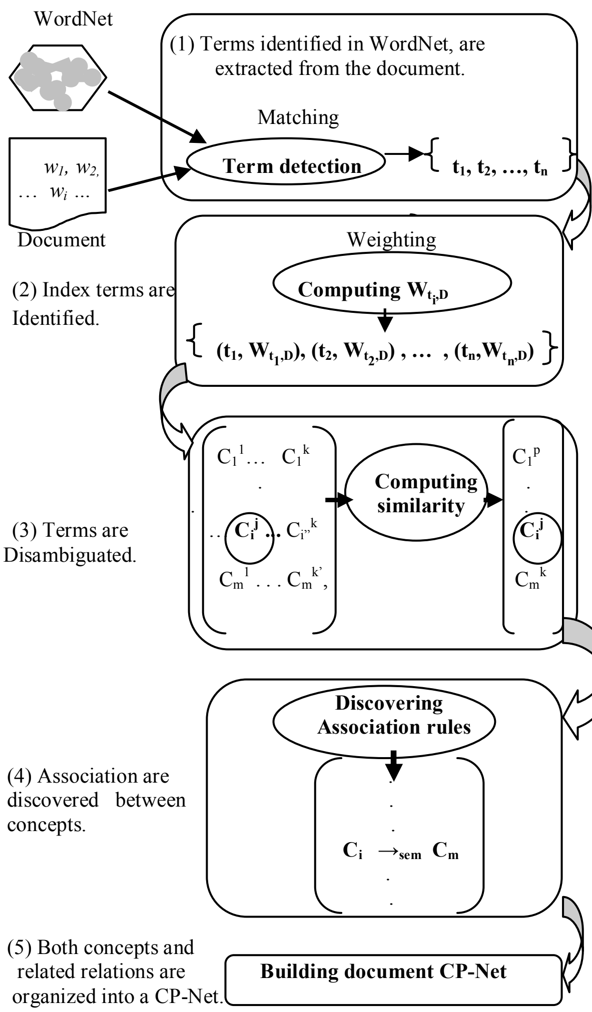

Fig.3. Steps for document CP-Net building

1) Identification of Index Terms: An important step for the indexing process consists in extracting monoword and multiword terms from texts. In order to identify multiword terms in document *D,* we propose a technique that performs a document word analysis. Let m_i be the next word (assumed not to be a stop word), to analyze in *D.* Using WordNet, we extract all the synsets C_i^j containing m_i: let $S=\{C_i^j, 1\leq j\leq |S| \,/\, m_i\in C_i^j\}$, pos_i be position of m_i in *D* (assuming that $pos_1 = 1$). We define the relative context of m_i occurrence in *D* given C_i^j namely CH_i^j , as the substring of *D* containing m_i, α words on the right and β words on the left of m_i. (α and β are determined by the position of m_i in C_i^j). For each C_i^j in *S,* we extract the m_i relative context given C_i^j , namely CH_i^j and then compare[2] strings CH_i^j and C_i^j . The longest concept C_i^j that matches CH_i^j is then retained as a representative term for *D*. The next word to analyze in *D* is m_j such that $pos_j = pos_i + (\alpha+1)$.

By this step, we will have identified the set of all (multiword or monoword) terms *t* that compose *D.* We then associate each term *t* with its count $occ_{t,D}$ (that is its occurrence frequency) in *D.*

2) Term Weighting: Term weighting gives an importance value to every detected term in order to select the most relevant terms in the document. We propose a variant of *tf*idf* weighting formula which combines a statistical measure and a probabilistic measure of occurrence frequency. The proposed weighting formula is formally defined in the following:

Let $W_{t,D}$ be the weight associated with term t in document *D* and *T(D)* the set of terms in *D*. $t, t' \in T(D)$. *t'* is a sub-term of *t* if *t* contains *t'*. *t* is called sur-term of *t'*.

Let $Sub_j(t) \in T(D)$ be a sub-term of *t*, $Sur_i(t) \in T(D)$ a sur-term of *t* and *Sn(t)* the set of synsets associated with *t*:

$$W_{t,D} = \left[occ_{t,D} + \sum_i occ_{Sur_i(t),D} + \sum_j \left(P\left(t \in Sn\left(Sub_j(t)\right)\right) * occ_{Sub_j(t),D} \right) \right] * \ln\left(\frac{N}{df(t)}\right).$$

Where *N* is the total number of documents in the corpus, *df(t)* (document frequency) is the number of documents in corpus that contain term *t* and $P\left(t\in Sn\left(Sub_j(t)\right)\right)$ is the probability that term *t* is a possible sense of its (more specific) sub-term $Sub_j(t)$, formally defined by:

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

The document index is then built by keeping only those terms whose weights are greater than a fixed threshold.

3) Term Disambiguation: Terms in document index may have a number of concept-senses corresponding to different WordNet synsets. We believe that each index term contributes to the content representation of D with only one sense. From where, we must choose for each term in document index, its best sense in the document. This is term disambiguation. The proposed term disambiguation

[2] We address here classical lexical comparison between two character strings.

approach relies on the computation of a score for every concept-sense

The proposed score of a concept sense measures its semantic relatedness to other concepts in the document. It is defined in the following:

Let $S_i = \{C_i^1, C_i^2, ..., C_i^n\}$ be the set of all synsets associated with index term t_i. We define the weight of a concept (sense) $C_i^j \in S_i$ as the weight of the related term t_i:

$$W_{C_i^j,D} = W_{t_i,D}$$

The score associated with C_i^j is then defined by:

$$Score(C_i^j) = \sum_{\substack{l=1..m \\ 1\le k\le n_l}} W_{C_i^j,D} * W_{C_k^l,D} * Dist(C_i^j, C_k^l)$$

Where $Dist(C_i^j, C_k^l)$ defines the semantic relatedness between concept-senses C_i^j and C_k^l. This can be for example the shortest way between nodes C_i^j and C_k^l in WordNet ontology or one of the other similarity measures proposed in literature (cited in section II).

The concept C_i^j which maximizes the score is then retained as the relevant sense of concept C_i in *D*. The set of all retained concepts defines the conceptual index of *D*.

4) Representing a Document as a CP-Net: The goal of this step is to build a CP-Net conceptual index. The use of CP-Net formalism is motivated by two reasons. First, CP-nets naturally support contextual associations. Second, CP-Nets allow a compact representation of both semantic and contextual relations between concepts, within a unified formalism namely the CP-Net graph. In the following, we describe the process of building the CP-Net nodes and relations.

a) The CP-Net Nodes: Our approach to build CP-Net nodes is based on the following principles:

- CP-Net nodes are random variables X_i attached with concepts C_i from the conceptual document index of *D*.

- A random variable corresponds then to a concept.

- Each variable X_i takes values in the set $Dom(X_i) = \{X_i\} \cup \{X_i^1, X_i^2, X_i^3, ...\}$ such that X_i^j *is-a* X_i or $X_i^j\ is-a\ X_k$ and X_i , X_k are synonymous (*is-a* defines the WordNet hyponymy relation).

By this step, we will have built the set $\eta(D) = \{(X_i, Dom(X_i))\}$ of the CP-Net concept nodes.

b) The CP-Net Relations: We propose to use association rules to discover latent contextual relations between CP-Net concept nodes. Concept nodes are semantic entities. The existing formalism of association rules allows discovering relations between lexical entities namely terms, we thus propose to extend it so as to support semantic entity associations (namely concept node associations). We thus define a semantic association rule between X_i and X_j noted $X_i \rightarrow_{sem} X_j$, as follows:

$$X_i \rightarrow_{sem} X_j \Leftrightarrow \exists X_i^k \subset Dom(X_i), \exists X_j^l \subset Dom(X_j) / X_i^k \rightarrow X_j^l$$

Where $X_i^k \rightarrow X_j^l$ is a classical association between terms X_i^k and X_j^l.

Confidence of rule *R:* $X_i \rightarrow_{sem} X_j$ is formally given by:

$$Confidence(R) = \frac{Support(X_i X_j)}{Support(X_i)}$$

Where:

$$Support(X_i X_j) = \min(Support(X_i), Support(X_j))$$

$$Support(X_i) = \min_k (W_{X_i^k,D}) \quad , \forall X_i^k \in Dom(X_i)$$

Let $\eta(D) = \{(X_i, Dom(X_i))\}$ be the set of concept-nodes of document CP-Net. Contextual relations between concepts in *η(D)* are discovered by means of semantic association rules. Association rules are based, in our context, on the following principles:

- A transaction is a document.
- Items are CP-Net concept-node values.
- An itemset is a set of concepts belonging to a same concept- node domain.

Formally, a semantic association rule $X_i \rightarrow_{sem} X_j$ defines in the CP-Net, a directed edge from concept node X_i to concept node X_j. X_i is the parent node of X_j in the graph.

The CP-Net nodes are linked by conditional relations defined by the related association rules. To complete the CP-Net, we annotate each node X_i of the graph by unconditional existence table namely $CPT(X_i)$ such that:

$$\forall X_i^k \in Dom(X_i), CPT(X_i^k) = W_{X_i^K,D}$$

5) Illustration: Our approach is illustrated through the following example. Let *D ((Paris, 0.9), (Toulouse, 0.5), (Center, 0.1), (Studio, 0.4), (Suburbs, 0.7))* be a document described by the given weighted concepts.

Toulouse is-a City and *Paris is-a City*, thus *Paris* and *Toulouse* pertain to *City* concept node domain. Similarly, both of *Center* and *Suburbs* pertain to *Place* concept node domain, whereas *Studio* is associated with *housing* concept node. That is to say:

η(D) = {(City, Dom (City)), (Place, Dom(Place)), (Housing, Dom(Housing))} / Dom(City)={Toulouse, Paris}, Dom(Place)= {Suburbs, Center}, Dom(Housing)={Studio}.

We aim to discover associations between *City, Housing* and *Place* nodes. Applying Apriori algorithm leads to the following steps:

- Step1: *extracting frequent itemsets.* We assume the minimum support threshold is *minsup*>0.1.

The extracted frequent *k*-itemsets (*k*= 1, 2, 3) are given in Table I.

TABLE I.
GENERATING FREQUENT *K*-ITEMSETS

1-Itemsets	Itemset	Support
	Toulouse	0.9
	Paris	0.5
	Center	0.1
	Suburbs	0.7
	Studio	0.4
Frequent 2-itemsets	Toulouse, Studio	0.4
	Toulouse, Suburbs	0.7
	Paris, Studio	0.4
	Paris, Suburbs	0.5
	Studio, Suburbs	0.4
Frequent 3-itemsets	Toulouse, Studio, Suburbs	0.4
	Paris, Studio, Suburbs	0.4

Remark1. Support(*Center*) < *minsup* : the 1-itemset *Center* is not frequent. It is then pruned.

- Step2: *generating association rules between frequent itemsets.* The extracted association rules are given in table II. The corresponding confidence values are computed leading to the results given in Table III.

TABLE II.
GENERATED ASSOCIATION RULES

R_1: Toulouse, Studio → Suburbs	R_7: Toulouse → Studio, Suburbs
R_2: Toulouse, Suburbs → Studio	R_8: Suburbs → Toulouse, Studio
R_3: Studio, Suburb → Toulouse	R_9: Studio → Toulouse, Suburbs
R_4: Paris, Studio → Suburbs	R_{10}: Paris → Studio, Suburbs
R_5: Paris, Suburbs → Studio	R_{11}: Suburbs → Paris, Studio
R_6: Studio, Suburbs → Paris	R_{12}: Studio → Paris, Suburbs

TABLE III.
COMPUTED RULE CONFIDENCE S

R_i	R_1	R_2	R_3	R_4	R_5	R_6
Confidence(R_i)	1	0.57	1	1	0.8	1
	R_7	R_8	R_9	R_{10}	R_{11}	R_{12}
	0.44	0.57	1	0.8	0.57	1

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

These rules are first used to build semantic association rules that in fact correspond to CP-Net concept-node relations. Thus, we deduce:

(1) From R_1 and/or R_4: *City, Housing* $\rightarrow_{sem}$ *Place,*
(2) From R_3 and/or R_6: *Housing, Place* $\rightarrow_{sem}$ *City,*
(3) From R_9 and/or R_{10}: *Housing* $\rightarrow_{sem}$ *City, Place.*

TABLE IV.
SELECTED ASSOCIATION RULES

R_1: Toulouse, Studio → Suburbs
R_3: Studio, Suburbs → Toulouse
R_4: Paris, Studio → Suburbs
R_6: Studio, Suburbs → Paris
R_9: Studio → Toulouse, Suburbs
R_{12}: Studio → Paris, Suburbs

Clearly, retaining the three rules will lead to a cycle in the CP-Net graph. To avoid this, we randomly retain one of them. Suppose the rule *City, Housing* $\rightarrow_{sem}$ *Place* is selected, then *CPTs* are associated with concept nodes which leads to The CP-Net document given in Figure 4.

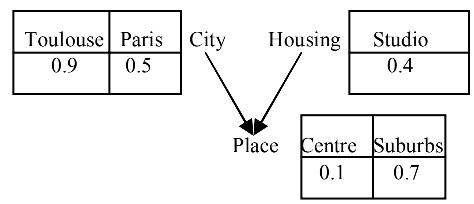

Fig.4. Document CP-Net .

C. CP-Net Based Query Evaluation

Query evaluation consists in selecting documents *(D)* assumed to be relevant to a query *(Q)*. For this aim, documents are ranked according to their relevance status value *(RSV(D,Q))* computed, in our approach, using a graph similarity measure that quantifies the degree of similarity between the query and document CP-Net graphs. Formally, this is expressed by:

$$RSV(D,Q) = SIM(G_D, G_Q)$$

Where G_D and G_Q are the CP-Net graphs that correspond respectively to the document D and the query Q. This similarity is computed as an aggregation of similarity graph across the shared concepts as follows:

$$SIM(G_D, G_Q) = \frac{|\eta(G_D) \cap \eta(G_Q)|}{|\eta(G_D) \cup \eta(G_Q)|} * \max_{X \in \eta(G_D) \cap \eta(G_Q)} (Sim^X(D,Q))$$

Where $\eta(G_D)$ and $\eta(G_Q)$ are the G_D and G_Q document and query CP-Net concept nodes respectively. $Sim^X(D,Q)$ is the partial similarity between D and Q at the X concept level. According to CP-Net graphs topology, this measure is computed as a combination of structural similarity and relational similarity measures:

$$Sim^X(D,Q) = \alpha * Sim^X_{struct}(D,Q) + (1-\alpha) * Sim^X_{relat}(D,Q).$$

Where $0 \le \alpha \le 1$, α being a given value that specifies the importance of structural similarity compared to relational similarity.

The structural similarity Sim^X_{struct} defines the proportion of common *X's* instances (values) of *D* and *Q*. A common instance of *D* and *Q* is a term of the query *Q* which belongs to the document *D*.

The relational similarity Sim^X_{relat} indicates the *X* degree of representativeness. This degree defines its importance as well in the document as in the query. It is measured using the depth associated with the concept in the hierarchy of the corresponding graph. Below are the formal definitions of the structural and relational similarity measures.

- Structural similarity measure: Let $\eta(G_D)\cap\eta(G_Q)$ be the set of common concept-nodes between document CP-Net G_D and query CP-Net G_Q. $\forall X \in \eta(G_D)\cap\eta(G_Q)$, let's consider $Dom_{X,D}$ and $Dom_{X,Q}$ the domains of instances associated with concept-node *X* respectively in G_D and G_Q. The structural similarity of *D* to *Q* at the level of the concept *X* is defined by:

$$Sim^X_{struct}(D,Q) = \frac{|\, Dom_{X,D} \cap Dom_{X,Q} \,|}{|\, Dom_{X,D} \cup Dom_{X,Q} \,|}$$

- Relational similarity measure: For a shared concept *X*, we define $Deg_D(X)$, $Deg_Q(X)$ respectively its importance level in the document *D* and the query *Q*. The importance level of a concept-node *X* is inversely proportional to the depth of the corresponding node in the graph. Thus, for a graph of depth *n*, the root of the graph is of level 1 and of importance 1. Its direct descendants are of level 2 and of importance1/2…etc. The elements of level *n* have an importance of *1/n*. Let $W_{X_i,Q}, W_{X_i,D}$ be the weights associated with *X*'s value X_i respectively in *Q* and *D*. $W_{X_i,D}$ is an unconditional weight. Whereas X_i weight in *Q* is a conditional weight defined by $CPT(X_i / U_k)$ given a value U_k of its parents in the query CP-net graph. In a first time, we simply define:

$$W_{X_i,Q} = Average(CPT(X_i / U_k))$$

The relational similarity of *D* to *Q* at the level of the concept *X* is defined as follows:

$$Sim^X_{relat}(D,Q) = \sum_i \frac{\min\left(W_{X_i,D} * Deg_D(X), W_{X_i,Q} * Deg_Q(X)\right)}{\left(W_{X_i,D} * Deg_D(X) + W_{X_i,Q} * Deg_Q(X)\right)}$$

V. Conclusion

We described in this paper a novel approach for flexible IR based on CP-Nets. The approach focuses on two main aspects. The first one consists in a conceptual indexing based on CP-Nets. 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 between terms leading to a more expressive document representation. We also proposed a CP-Net based query evaluation. Our approach aims to evaluate document relevance to a given query on the basis of simple graph similarity measure. In future, we plan to validate empirically our approach using structured test collections such as Reuters collection.

References

[1] G. Bordogna, P. Carrara, G. Pasi, "Query term weights as constraints in fuzzy information retrieval," *Information Processing and Management,* 27(1) , 15-26, 1991.

[2] G. Bordogna and G. Pasi," A fuzzy linguistic approach generalizing Boolean information retrieval: a model and its evaluation," in Journal of the American Society for Information Science, 44(2), 70-82, 1993.

[3] G. Bordogna and G. Pasi, "Linguistic aggregation operators of selection criteria in fuzzy information retrieval*," in International Journal of Intelligent Systems*, (10), 233-248, 1995.

[4] D. A. Buell and D. H. Kraft, " A model for a weighted retrieval system," *in Journal of the American Society for Information Science*, 32(3), 211-216,1981.

[5] D. H. Kraft and D. Buell,"Fuzzy sets and generalized Boolean retrieval systems," *in International Journal of Man-Machine Studies*, 19(1, 45-56, 1983.

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

[7] R. Krovetz," Homonomy and Polysemy in Information Retrieval," in the Proceedings of the COLING/ACL '97 conference.

[8] R. Mihalcea and D. Moldovan, " Semantic indexing using WordNet senses," *in Proceedings of ACL Workshop on IR & NLP*, HongKong, 2000.

[9] R. Mihalcea, "*Co*-training and Selftraining for Word Sense disambiguation," in *Proceedings of the Conference on Natural Language Learning (CoNLL 2004)*, Boston, USA.

[10] M. Sanderson," Word sense disambiguation and information retrieval," *in Proc. of the 17th Annual International ACM-SIGIR Conference on Research and development in Information Retrieval.* (pp. 142-151), Springer- Verlag, 1994.

[11] M. Sanderson,"Word Sense disambiguation and Information Retrieval,". Phd Thesis, Technical Report (TR-1997-7) of the department of Computing Science at the University of Glasgow, Glasgow G12 8QQ, UK, 1997.

[12] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in 33rd Annual Meeting, (pp189-196), Association for Computational Linguistics,Cambridge, Massachusetts, USA, 1995.

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

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

[15] L. R. Khan, "Ontology-based Information Selection," Phd Thesis, Faculty of the Graduate School, University of Southern California, 2000.

[16] C. Leacock, G. A. Miller and M. Chodorow, " Using corpus statistics and WordNet relations for sense identification," *Comput. Linguist.* 24, 1, (pp. 147-165), 1998.

[17] D. Lin,"An information-theoretic definition of similarity," *in Proceedings of 15th International Conference On Machine Learning*, 1998.

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

[19] M. Baziz,, M. Boughanem and N. Aussenac-Gilles, "The Use of Ontology for Semantic Representation of documents," in *The 2nd Semantic Web and Information Retrieval Workshop(SWIR), SIGIR* Sheffield UK, (pp. 38-45), Ying ding, Keith van Rijsbergen, Iad Ounis, Joemon Jose (Eds.), 2004

[20] N. Guarino, C. Masolo, C. and G. Vetere, " OntoSeek: Using Large Linguistic Ontologies for Accessing On-Line Yellow Pages and

Product Catalogs," National Research Council, LADSEBCNR: Padavo, Italy, 1999.

[21] M. A. Stairmand and J. B. William, " Conceptual and Contextual Indexing of documents using WordNet-derived Lexical Chains," *in Proc. of 18th BCS-IRSG Annual Colloquium on Information Retrieval Research*, 1996.

[22] B. Daille, E. Gaussier and J. M.Lange, "Towards Automatic Extraction of Monolingual and Bilingual Terminology," *in Proc. of COLING 94,* 515-521.

[23] J. S. Justeson, and S. M. Katz, **"** Technical terminology: some linguistic properties and an algorithm for identification in text," *in Natural Language Engineering*, 1:9-27, 1995.

[24] F. Boubekeur, M. Boughanem and L.Tamine-Lechani,"Towards Flexible Information Retrieval Based on CP-Nets,". *in Proc. Of Flexible Query Answering (FQAS),* Milano, Italy, Henrik Legind Larsen, Gabriella Pasi, Daniel Ortiz-Arroyo (Eds.), World Scientific Publishing, Advances in Artificial Intelligence, p. 222-231, june 2006.

[25] A. Stein, J. A. Gulla, A. Müller and U. Thiel, "Conversational interaction for semantic access to multimedia information," in Maybury, M. T., ed., Intelligent Multimedia Information Retrieval, (pp. 399–421), Menlo Park, CA: AAAI/The MIT Press, 1997.

[26] E. M. Voorhees, " Query expansion using lexical-semantic relations," *in Annuel International ACM SIGIR conference on research and development in information retrieval (SIGIR'94),* (pp pages 63–69). ACM press, Dublin, Ireland.

[27] M. H. Haddad, " Extraction et Impact des connaissances sur les performances des Systèmes de Recherche d'Information," Thèse de Doctorat en Informatique de L'Université Joseph Fourier, Grenoble, 2002.

[28] H. Haddad, J.P. Chevallet and M. F. Bruandet, "Relations between Terms Discovered by Association Rules," 4th European conference on Principles and Practices of Knowledge Discovery in Databases PKDD'2000, Workshop on Machine Learning and Textual Information Access, France (2000).

[29] J. Wei, S. Bressan and B. Chin Ooi, "Mining Term Association Rules for Automatic Global Query Expansion: Methodology and Preliminary Results," *wise*, p. 0366, First International Conference on Web Information Systems Engineering (WISE'00)-Volume 1, 2000.

[30] M. Song , Y. Song b, X. Hu and R. B. Allen, "Integration of association rules and ontologies for semantic query expansion,", in Data Knowledge Engeenering (2007), doi: 10.1016/ j.datak.2006.10.010

[31] C. Boutilier, R. Brafman, H. Hoos, and D. Poole," Reasoning with Conditional Ceteris Paribus Preference Statements," *in Proc. of UAI*, 71 –80, 1999.

[32] C. Boutilier, F. Bacchus and R. Brafman,"UCP-networks: A directed graphical representation of conditional utilities," *in Proc. of UAI*, 56 – 64, 2001.

[33] R. Agrawal, T. Imielinski, and A. Swami. "Mining association rules between sets of items in large databases". In *Proceedings of the ACMSIGMOd Conference on Management of data*, Washington, USA (pp. 207–216). ACM Press, 1993.

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