Une Approche d'Indexation Conceptuelle de Documents Basée sur les Graphes CP-Nets

Fatiha Boubekeur[1, 2], Mohand Boughanem[1], Lynda Tamine-Lechani[1]

[1] IRIT-SIG/RFI, Université Paul Sabatier,
31062 Toulouse, France
[2] Université Mouloud Mammeri,
15000 Tizi Ouzou, Algérie
boubekeu@irit.fr, boughane@irit.fr, lynda.lechani@irit.fr

Résumé. Ce papier décrit une approche d'indexation conceptuelle basée sur les CP-Nets (*Conditional Preferences Networks*). Nous proposons d'utiliser le formalisme CP-Net comme langage d'indexation afin de représenter les concepts et les relations conditionnelles entre eux d'une manière relativement compacte. Les noeuds du CP-Net sont les concepts représentatifs du contenu du document et les relations entre ces nœuds expriment les associations conditionnelles qui les lient. Notre contribution porte sur un double aspect: d'une part, nous proposons une approche d'extraction des concepts en utilisant WordNet. Les concepts résultants forment les noeuds du CP-Net. D'autre part, nous proposons d'étendre et d'utiliser la technique de règles d'association afin de découvrir les relations conditionnelles entre les concepts nœuds du CP-Nets.

Mots-clés: Recherche d'information, Indexation conceptuelle, CP-Nets, WordNet, Règles d'association

1 Introduction

Le but principal d'un système de recherche d'information (SRI) est de retrouver l'information pertinente pour une requête utilisateur. Requêtes et documents sont généralement exprimés par un ensemble de mots-clés (sacs de mots) simples sensés représenter au mieux leurs contenus sémantiques. Les termes sont automatiquement extraits ou manuellement assignés aux documents et aux requêtes. L'évaluation consiste alors à apparier requête et documents pour retrouver les documents qui correspondent au mieux à la requête. Une caractéristique clé de tels systèmes est que le degré d'appariement de la requête et du document dépend du nombre de termes communs. Il est bien connu qu'une requête est habituellement une description vague et incomplète du besoin en information de l'utilisateur et les auteurs des documents utilisent un vocabulaire très diversifié pour exprimer les mêmes concepts. Ceci mène aux problèmes cruciaux de disparité des termes (*term mismatch*) et d'ambiguïté en recherche d'information (RI). Ce papier traite ces problèmes en proposant une solution au niveau de l'indexation des documents. Plus précisément, nous proposons d'utiliser (1) l'ontologie générale WordNet afin d'identifier les concepts représentatifs d'un document, (2) les règles d'association pour découvrir des relations

conditionnelles entre ces concepts, (3) les CP-Nets pour organiser concepts et relations en une représentation graphique. Le formalisme CP-Net est utilisé comme langage d'indexation, pour deux raisons. D'abord, les CP-Nets offrent un cadre unifié pour organiser de manière compacte et intuitive les concepts et les relations qui les lient. Ensuite, les CP-Nets permettent une représentation plus riche des documents puisqu'ils supportent les relations contextuelles et sémantiques entre concepts. Les concepts et les relations associées sont susceptibles de résoudre les problèmes de disparité et d'ambiguïté des termes et d'améliorer ainsi les résultats de la RI.

Le papier est structuré comme suit : en section 2, nous présentons les problèmes que nous souhaitons aborder, à savoir la disparité des termes et l'ambiguïté en RI, ainsi que les travaux qui permettent d'y apporter des solutions. Un résumé de notre contribution suivra. En section 3, nous détaillons notre approche d'indexation conceptuelle basée sur les CP-Nets. Une illustration de notre approche est présentée en section 4. La section 5 conclut le papier.

2 Problématique

La représentation des documents et requêtes est un problème fondamental en RI. La plupart des modèles de RI classiques utilisent toujours la technique usuelle en sac de mots. Cette technique présente cependant deux inconvénients majeurs qui mènent à de mauvaises performances du SRI:

- *La disparité des termes:* Les utilisateurs de SRI utilisent souvent pour décrire les concepts de leurs requêtes, des termes différents de ceux qu'utilisent les auteurs pour décrire les mêmes concepts dans leurs documents. Ainsi, un document peut être pertinent même s'il n'utilise pas les mêmes mots que ceux de la requête. Cependant, dans les SRI classiques, un document pertinent ne sera pas retrouvé en réponse à une requête si les représentations du document et de la requête ne partagent pas au moins un terme [7]. C'est ainsi par exemple qu'un document sur *la peur* pourtant pertinent pour une requête sur *la crainte*, ne sera pas retrouvé si le mot *crainte* est absent de ce document. La disparité des termes implique un silence documentaire.

- *L'ambiguïté:* La plupart des SRI représente les documents et les requêtes par les termes qu'ils contiennent. Cependant, les termes sont ambigus. Cette ambiguïté est divisée en homonymie et polysémie. L'homonymie traduit la propriété qu'ont certains termes à être représentés par une même chaîne de caractères, et associés à différents sens. La polysémie est liée à la propriété qu'ont certains termes à exprimer différents sens. "The *bark* of a dog" versus "the *bark* of a tree" est un exemple d'homonymie, "*opening* a door" versus "*opening* a book" est un exemple de polysémie [7]. Dans les SRI classiques, l'ambiguïté implique que des documents non pertinents sont retrouvés. Ainsi, un document qui traite de l'aboiement des chiens (*ba*rk of dogs) sera retrouvé comme pertinent pour une requête portant sur l'écorce des arbres *(bark of trees)* si le mot *bark* figure dans le document et la requête. L'ambiguïté des termes implique un bruit documentaire.

Les SRI basés sur la technique classique en *" sacs de mots "* présentent ainsi de sérieux problèmes au niveau performance du fait de leur incapacité à traiter avec l'ambiguïté de la langue et l'imprécision sémantique des mots simples. Dans les

dernières décennies, de nombreux travaux de recherche en RI se sont orientés vers la prise en compte de la sémantique des mots dans le processus d'indexation. Les méthodes utilisées sont sensées améliorer les performances d'un SRI en termes de rappel et de précision en le rendant capable de traiter avec l'ambiguïté des mots. Deux grandes tendances existent : l'indexation sémantique et l'indexation conceptuelle. L'indexation sémantique se base sur des techniques de désambiguïsation contextuelle des mots dans les documents et requêtes. L'idée est que le sens d'un mot est complètement déterminé par les autres mots occurrant dans le même contexte. Yarowsky dans [11], associe aux mots d'index, des mots du contexte qui aident à déterminer leur sens. Tandis que Voorhees dans [10], en se basant sur WordNet comme outil de désambiguïsation des mots, calcule la distance sémantique entre chaque synset (sens associé dans WordNet) possible du mot à désambiguïser avec les synsets des mots occurant dans la même phrase. Le synset qui est le plus proche des autres mots de la phrase est alors choisi. L'indexation conceptuelle se base sur des concepts tirés d'ontologies et de taxonomies pour indexer les documents contrairement aux listes de mots simples. Une ontologie souvent utilisée dans ce sens est l'ontologie générale WordNet. Khan dans [6], utilisant la notion de concept, propose un algorithme permettant d'attacher les termes d'un texte aux concepts de l'ontologie en se basant sur la notion de région d'ontologie et de distance sémantique entre concepts. Une approche similaire est proposée dans [2], dans laquelle les termes d'un texte sont attachés aux concepts de Wordnet en se basant sur la notion de similarité sémantique entre concepts. Un document est finalement représenté par un réseau sémantique de concepts et de relations sémantiques entre eux. L'utilisation des relations entre termes est une issue prometteuse pour de meilleures performances de RI. Principalement, l'utilisation des relations entre termes permet d'améliorer le rappel du SRI, c.-à-d. le nombre de documents pertinents retrouvés. Le SRI retrouvera non seulement les documents qui contiennent les mots de la requête mais également des documents contenant les mots qui sont en relation avec les mots de la requête. Les relations entre termes étant extraites d'ontologies [6], [2], ou découvertes dans le contexte du document, au moyen de règles d'association [5].

Nous proposons, dans ce papier, une approche d'indexation conceptuelle de documents basée sur les CP-Nets. Les nœuds du CP-Net sont des concepts. Les relations du CP-Net traduisent des dépendances contextuelles entre concepts. Nous définissons alors : (1) une approche d'extraction des termes du document, (2) une formule de pondération des termes tenant compte de leur sémantique, (3) une méthode de désambiguïsation des termes basée sur l'utilisation de WordNet, (4) une approche pour découvrir les relations contextuelles entre concepts par les règles d'association et (5) une approche pour combiner les concepts et les relations correspondantes dans une représentation graphique compacte, à savoir le CP-Net.

3 Approche d'Indexation Conceptuelle Basée sur les CP-Nets

Nous utilisons WordNet et les règles d'association afin de construire le graphe CP-Net représentatif du document. Le processus d'indexation du document est effectué en quatre étapes principales. Nous les décrivons dans les paragraphes suivants.

• **Notions préliminaires.** Le but du processus d'indexation est d'identifier et d'extraire les termes qui sont susceptibles de représenter le contenu sémantique du document. Les termes sont généralement représentés comme listes de mots. La longueur d'un terme t notée $|t|$ est alors définie comme le nombre de mots dans *t*. Un terme mono-mot consiste en une liste d'un seul mot. Un multi-terme est une liste de plusieurs mots. Soit t un terme représenté comme liste de mots w_i, $t=[w_1, w_2, \ldots, w_n]$. Les éléments dans t peuvent être identiques, représentant différentes occurrences d'un même mot. On note w_i le $i^{ème}$ mot dans *t*. Nous définissons la position du mot w_i dans la liste t comme suit: $pos_t(w_1)=1$; $pos_t(w_{i+1})=pos_t(w_i)+1, \forall i=1..l$.

Définition 1. Soient $L_1=[w_1, w_2, \ldots, w_m]$, $L_2=[y_1, y_2, \ldots, y_n]$, deux listes de mots données. L_2 est une sous liste de L_1 si la séquence de mots dans L_2 apparaît aussi dans L_1. Formellement: $L_2 = sub(L_1, p, l)$ *si* $\exists w_i \in L_1$ *tel que* $p = pos(w_i)$ et $\forall j, 0 \le j \le (l-1), w_{p+j} = y_{j+1}$.

*S*oient t_1, t_2 deux termes donnés par les listes de mots respectives L_1 et L_2:
t_2 est un sous-terme de t_1 si L_2 est une sous liste de L_1.
t_1 est un sur-terme de $t_2 \Leftrightarrow$ t_2 est un sous-terme de t_1.

• **Identification des Termes du Document.** Etant donné un document *d*, nous procédons dans cette étape à l'extraction de l'ensemble de ses termes représentatifs. Pour cela, et avant toute élimination de mots vides, nous procédons à une analyse mot par mot du document *d*. Soit w_i, le prochain mot, non vide, à analyser dans *d*. On extrait de WordNet l'ensemble *S* des termes C_j^i qui contiennent (au sens inclusion de chaînes de caractères) le mot w_i. Soit donc $S=\{C_1^i, C_2^i, \ldots, C_m^i\}$cet ensemble. *S* est composé de mono et de multi-termes. On ordonne alors *S* comme suit : donc S = $S=\{C_{(1)}^i, C_{(2)}^i, \ldots, C_{(m)}^i\}$ tel que (*j*)=(1)…(*m*), est une permutation d'indices telle que $|C_{(1)}^i| \ge |C_{(2)}^i| \ge \ldots \ge |C_{(m)}^i|$. Les termes ayant des longueurs égales sont indifféremment placés l'un à côté de l'autre. Pour chaque élément $C_{(j)}^i$de S, *j* = 1..*m*, on note $Pos_{C_{(j)}^i}(w_i)$ la position de w_i dans la liste de mots $C_{(j)}^i$. Soit $pos_d(w_i)$ la position de w_i dans *d*.
On appellera contexte relatif du mot w_i dans *d* par rapport au terme $C_{(j)}^i$, la sous chaîne de caractères $CH_j^i = sub(d, p, l)$, $l=|C^i_{(j)}|$ et $p = pos_d(w_i) - \left(pos_{C_{(j)}^i}(w_i) - 1\right)$.

On extrait alors le contexte relatif de w_i dans d, soit $CH_j^i = sub(d, p, l)$ (Figure 1), puis on compare les listes de mots CH_j^i et $C_{(j)}^i$. Si $CH_j^i \ne C_{(j)}^i$, le terme suivant $C_{(j+1)}^i \in S$ est analysé, sinon le terme $t_k = C_{(j)}^i$, est identifié. Si t_k recouvre totalement un ou plusieurs termes adjacents le précédant (t_{k-1} *à* t_j , $j \le k-1$), ces termes sont

éliminés de la description du document. Le prochain mot de *d* à analyser est w_j, tel que $pos_d(w_j) = p + l$.

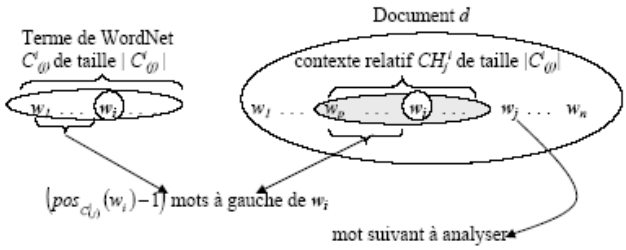

Fig. 1. Extraction du contexte relatif d'un terme

A l'issue de cette première étape, nous aurons identifié les termes t_i qui caractérisent le document *d*. En associant chacun d'eux à sa fréquence d'occurrence, on obtient l'ensemble $T(d) = \{(t_1, Occ_1), (t_2, Occ_2), \ldots, (t_n, Occ_n) / t_i \in d, Occ_i = Occ(t_i)$ la fréquence d'occurrence de t_i dans d, $1 \leq i \leq n\}$.

- **Pondération des Termes d'Index.** La pondération permet d'assigner à chaque terme de l'index son poids d'importance dans le document. Dans le cas des mono-termes (mots simples), des variantes de *tf*idf* sont utilisées. Le poids d'un terme *t* dans le document *d* est alors exprimé par : $W_{t,d} = tf(t) * idf(t)$, *tf* est la fréquence du terme, *idf* sa fréquence documentaire inverse telle que $idf(t) = \log\left(\frac{N}{df(t)}\right)$, *N* étant le nombre de documents dans le corpus et $df(t)$, fréquence documentaire du terme *t*, le nombre de documents du corpus qui contiennent *t*. Dans le cas des multi-termes, les approches de pondération des termes proposées dans la littérature s'appuient en général sur une analyse statistique et/ou syntaxique. Globalement, il s'agit d'additionner les fréquences de mots simples, ou multiplier le nombre d'occurrences du terme par le nombre de mots simples qu'il contient. Baziz et al. dans [2], proposent de calculer la fréquence d'un multi-terme *t* dans un document *d*, en additionnant le nombre d'occurrences du terme lui-même et le nombre d'occurrences de ses sous - termes. Formellement, si $sub(t_i)$ est l'ensemble de tous les sous-termes possibles t_i qui peuvent être dérivés de *t,* et $long(t_i)$ le nombre de mots dans t_i, alors : $tf(t) = Occ(t) + \sum_{t_i \in sub(t)} \frac{long(t_i)}{long(t)} Occ(t_i)$.

Pour notre part, nous proposons une nouvelle approche de pondération qui se base sur une analyse statistique et une analyse sémantique. Nous définissons alors une variante de *tf*idf* qui combine: (1) une mesure statistique des occurrences lexicales du terme lui-même, (2) une mesure statistique des occurrences lexicales du terme dans ses sur-termes, (3) une mesure probabiliste des occurrences du terme dans les sens de ses sous-termes. La formule proposée est définie comme suit : Soit $T(d)$ l'ensemble des termes descripteurs de d ; $t, t' \in T(d)$; $Sub_j(t) \in T(d)$ un sous-terme de t, et $Sur_i(t) \in T(d)$ un sur-terme de t. On pose $S(t)$ l'ensemble des concepts (synsets ou

sens) associés à *t* dans WordNet et *C* un synset donné. Nous définissons la probabilité que *t* soit un sens possible de $Sub_j(t)$ par :

$$P(t \in S(Sub_j(t))) = \frac{|\{C \in S(Sub_j(t)) / t \in C\}|}{|S(Sub_j(t))|} \quad \textbf{(1)}$$

Le poids $W_{t,d}$ du terme *t* dans le document *d* est alors défini par: $W_{t,d} = tf(t) * idf(t)$, tel que :

$$W_{t,d} = \left(Occ(t) + \sum_i Occ(Sur_i(t)) + \sum_j \left[P(t \in S(Sub_j(t))) * Occ(Sub_j(t)) \right] \right) * \ln\left(\frac{N}{df(t)} \right) \quad \textbf{(2)}$$

Où : *N* est le nombre de documents dans le corpus, *df(t)* (fréquence documentaire) est le nombre de documents du corpus qui contiennent le terme *t*.

L'index, *Index(d),* du document *d* sera construit sur la base des seuls termes dont les poids sont supérieurs à un seuil fixé.

- **Désambiguïsation des Termes**. Tout terme t_i dans *Index(d)* peut avoir plusieurs sens (synsets de WordNet) lui correspondant. Soit $S_i = \{C_1^i, C_2^i, \ldots, C_n^i\}$ l'ensemble des synsets associés au terme t_i. Ainsi, t_i possède $|S_i| = n$ sens. Nous admettons que chaque terme contribue à la représentation du contenu de *d* avec seulement un sens. Ainsi, nous devons choisir, pour chaque $t_i \in Index(d)$, son meilleur sens dans *d*. C'est la désambiguïsation.

Parmi les différentes méthodes de désambiguïsation proposées dans la littérature, nous nous sommes intéressés à l'approche proposée dans [2] pour sa simplicité. Cette approche se base sur le calcul d'un score (*C_Score*) pour chaque concept (ou sens) lié à un terme d'index. Ainsi, pour un terme t_i, le score de son $j^{ème}$ sens, noté C_j^i est donné par $C\_Score(C_j^i) = \sum_{\substack{l \in [1..m] \\ l \neq i}} \sum_{k \in [1..n_l]} Dist(C_j^i, C_k^l)$

Où *m* est le nombre de termes dans *Index(d)*, n_l représente le nombre de sens de WordNet propres à chaque terme t_l et $Dist(C_j^i, C_k^l)$ est une mesure de similarité sémantique entre les concepts C_j^i et C_k^l telle que définie dans [9] et [8]. Le concept C_j^i qui maximise le score est alors retenu comme meilleur sens du terme t_i.

Notre approche diffère principalement de celle en [2] dans la formule utilisée pour le calcul du score d'un concept. En effet, nous pensons que l'utilisation de la seule similarité sémantique entre concepts est insuffisante pour déterminer le meilleur sens d'un terme car cette mesure est indépendante du contexte (elle ne tient pas compte de la représentativité des termes dans le contexte de document). Nous croyons que le meilleur sens pour un terme t_i dans le document *d* est celui qui fortement corrélé avec les sens des termes *importants* dans *d*. Pour cela, nous définissons d'abord le poids

d'un concept (sens) $C_j^i \in S_i$ comme poids du terme correspondant t_i : $\forall C_j^i \in S_i, W_{C_j^i,d} = W_{t_i,d}$

Nous proposons alors de calculer le score comme suit:

$$Score\left(C_j^i\right) = \sum_{\substack{l \in [1,..,m] \\ l \neq i}} \sum_{1 \leq k \leq nl} WC_j^i, d * WC_k^l, d * Dist\left(C_j^i, C_k^l\right) \quad \textbf{(3)}$$

Le concept $C_m^i \in S_i$ tel que *Score* (C_m^i) = max_j *(Score* (C_j^i)*)* sera retenu comme meilleur sens du terme t_i dans *d*. L'ensemble des concepts retenus constituera le noyau sémantique *N(d)* du document *d*.

- **Représentation CP-Net d'un Document.** Le but de cette étape est de construire l'index conceptuel CP-Net. Les CP-Nets ont été introduits dans [4] comme outil de représentation compacte des relations de dépendances préférentielles conditionnelles entre des variables (caractéristiques) données. Nous proposons d'utiliser le formalisme CP-Net comme langage d'indexation, d'une part car les CP-Nets supportent naturellement les dépendances contextuelles, d'autre part les CP-Nets permettent une représentation compacte des relations sémantiques et contextuelles entre concepts, dans un formalisme graphique unifié. Dans ce qui suit, nous décrivons le processus de construction des noeuds et des relations du CP-Net.

Les Noeuds du CP-Net. Soit $N(d) = \{C_1, C_2, \ldots, C_n\}$ le noyau sémantique du document d. Notre approche pour construire les noeuds du CP-Net est basée sur les principes suivants :

- Les noeuds du CP-Net sont des variables attachées aux concepts C_i du noyau sémantique du document *d*. Dans ce qui suit, nous désignerons une variable noeud du CP-Net par le concept correspondant.

- Chaque variable C_i prend des valeurs dans l'ensemble $Dom(C_i) = \{C_1^i, C_2^i, C_3^i, ...\}$.

- Chaque valeur dans $Dom(C_i)$ est un concept $C_j^i \in N(d)$ tel que $C_j^i \; est-un \; C_i$ (*est-un* définit la relation de subsumption de WordNet).

A l'issue de cette étape, nous aurons construit l'ensemble $\eta(d) = \{(C_i, Dom(C_i)\}$, on notera plus simplement $\eta(d) = \{(X, Dom(X))\}$, des nœuds du CP-Net document.

Les Relations du CP-Net. Nous proposons d'utiliser les règles d'association pour découvrir les relations contextuelles latentes entre les concepts noeuds du CP-Net. Les règles d'association furent initialement introduites dans [1], dans le but de générer les associations significatives entre ensembles d'items dans une base de données transactionnelle. Elles ont été utilisées en RI pour découvrir des relations significatives entre termes d'indexation [5]. Dans notre contexte, les règles d'association sont utilisées pour découvrir les relations significatives entre concepts représentatifs du contenu d'un document. Le formalisme des règles d'association est alors étendu pour supporter les associations entre concepts. Le modèle formel est défini dans ce qui suit : soit $\eta(d) = \{(X, Dom(X))\}$ l'ensemble des concepts nœuds du CP-Net document. $X, Y \in \eta(d)$.

Définition 2. Une règle d'association sémantique entre les concepts X et Y, notée $X \rightarrow_{sem} Y$, est définie comme suit :

$X \rightarrow_{sem} Y \Leftrightarrow \exists X_i \in Dom(X), \exists Y_j \in Dom(Y) / X_i \rightarrow Y_j$, tel que $X_i \rightarrow Y_j$ est une association entre les termes X_i et Y_j.

La signification intuitive de la règle $X \rightarrow_{sem} Y$ est que si un document *porte sur (is about)* le concept X, il tend également à *porter sur* le concept Y. L'*aboutness* du document exprime le *focus* de son contenu. Cette interprétation s'appliquant aussi à la règle $X_i \rightarrow Y_j$. Ainsi, la règle $R : X_i \rightarrow Y_j$ exprime la probabilité que le document *soit autour* de Y_j sachant qu'il *est autour* de X_i. La confiance associée à R se rapporte alors au degré d'importance de Y_j dans le document d, sachant le degré d'importance de X_i dans d. Elle est formellement définie dans ce qui suit.

Définition 3. La confiance de la règle $R : X_i \rightarrow Y_j$ est formellement donnée par:

$$Confiance(R) = \frac{Support(X_i \wedge Y_j)}{Support(X_i)} = \frac{\min(W_{X_i,d}, W_{Y_j,d})}{W_{X_i,d}}$$

Définition 4. La confiance de la règle d'association sémantique $R_{sem} : X \rightarrow_{sem} Y$ est définie comme suit:

$$Confiance(X \rightarrow_{sem} Y) = \max_{i,j} \left(Confiance \left(\begin{array}{c} R : X_i \rightarrow X_j \\ X_i \in Dom(X), Y_j \in Dom(Y) \end{array} \right) \right)$$

Remarque. $Confiance(X \rightarrow_{sem} Y)$ est toujours égale 1.

Dans notre contexte, le support de la règle d'association sémantique $X \rightarrow_{sem} Y$ se rapporte à la proportion de règles d'associations $X_i \rightarrow Y_j$ ($X_i \in Dom(X)$ et $Y_j \in Dom(Y)$, qui ont une confiance supérieure ou égale à un seuil minimal de confiance *minconf=1*. Le support est formellement défini dans ce qui suit.

Définition 5. Le support de la règle $R : X \rightarrow_{sem} Y$ est donné par:

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

Nous proposons de découvrir les relations entre concepts de $\eta(d)$ au moyen de règles d'association sémantiques. Les règles d'association sémantiques sont basées, dans notre contexte, sur les principes suivants: (1) une transaction est un document, (2) les items sont les valeurs des concepts noeuds du CP-Net (3) un itemset est un concept représentant un nœud du CP-Net, (4) une règle d'association $X \rightarrow_{sem} Y$ définit dans le CP-Net, un arc orienté du noeud X vers le

noeud *Y*. En utilisant les règles d'association, nous visons la construction d'une structure hiérarchique conditionnelle du *focus* du contenu de document. C'est-à-dire que nous visons à structurer des concepts décrivant le document, dans une hiérarchie conditionnelle correspondant à la sémantique des règles d'association extraites. Le problème de découverte des règles d'association entre les concepts est divisé en deux étapes, suivant le principe de l'algorithme A-priori (introduit en [1]). D'abord identifier les 1-itemsets fréquents, correspondant à des *concepts fréquents*. Un concept fréquent est dans notre contexte, un concept qui a un poids plus grand qu'un seuil minimum fixé. Ensuite, extraire les règles d'associations entre les itemsets fréquents. L'objectif est de ne garder que les seules règles qui ont un support et une confiance supérieurs à un seuil minimal de support *minsup* et un seuil minimal de confiance *minconf* respectivement.

Des problèmes peuvent surgir lors de la découverte des règles d'association, tels que les redondances et les cycles. Les règles redondantes découlent généralement des propriétés transitives : $X \rightarrow_{sem} Y$, $Y \rightarrow_{sem} Z$ *et* $X \rightarrow_{sem} Z$. Pour éliminer la redondance, nous proposons de construire la couverture minimale de l'ensemble de règles extraites (il s'agit de l'ensemble minimal de règles d'association non transitives). L'existence des cycles dans le graphe est due à la découverte simultanée des règles d'association $X \rightarrow_{sem} Y$ et $Y \rightarrow_{sem} X$, ou des règles d'association telles que $X \rightarrow_{sem} Y$, $Y \rightarrow_{sem} Z$ *et* $Z \rightarrow_{sem} X$. Pour résoudre ce problème, nous éliminons la règle de plus faible support parmi celles qui ont conduit au cycle. Dans le cas de support égal, nous éliminons aléatoirement une règle dans le cycle. Une fois le CP-Net construit, les noeuds du CP-Net sont annotés par une table inconditionnelle notée *CPT(Y)* dont les valeurs définissent les poids (inconditionnels) d'importance des termes dans le document correspondant :

$$\forall X_i \in Dom(X), CPT(X_i) = W_{X_i,d} \quad \textbf{(4)}$$

4 Illustration

Soit *d ((Toulouse, 0.9), (Paris, 0.5), (Center, 0.1), (Studio, 0.4), (Suburbs, 0.7)...)* un document décrit par les concepts pondérés donnés. *Toulouse* et *Paris* appartiennent au domaine de valeurs du concept noeud *City*. De même, *Center* et *Suburbs* appartiennent au domaine du concept nœud *Place*, tandis que *Studio* est associé au concept noeud *Housing*. Ainsi *η(d) = {(City, Dom (City)), (Place, Dom(Place)), (Housing, Dom(Housing))}/ Dom(City) ={Toulouse, Paris}, Dom(Place) = {Suburbs, Center}, Dom(Housing) = {Studio}*.

Nous visons à découvrir des associations entre les noeuds *City, Housing,* et *Place*. L'application de l'algorithme Apriori mène à: (1) l'extraction des itemsets fréquents, et (2) la génération des règles d'association entre les 1-itemsets fréquents. Nous nous intéressons aux relations entre les concepts individuels dans le document (plutôt qu'entre les ensembles de concepts), ainsi nous calculons seulement les *k*-itemsets pour $k = 1, 2$. En supposant un seuil minimum de support *minsup* = 0.1, les k-itemsets ($k = 1, 2$) fréquent extraits sont donnés en Table 1. *Support*(Center) <

minsup, ainsi le *1-itemset Center* n'est pas fréquent, il est alors éliminé. Les règles d'association extraites sont données en Table 2.

Table 1. Gération des k-itemsets fréquents

	Itemset	Support
1-itemset fréquent	Toulouse	0.9
	Paris	0.5
	Center	0.1
	Suburbs	0.7
	Studio	0.4
2-itemsets fréquents	Toulouse, Studio	0.4
	Toulouse, Suburbs	0.7
	Paris, Studio	0.4
	Paris, Suburbs	0.5
	Studio, Suburbs	0.4

Table 2. Les règles d'association découvertes

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

En appliquant la formule donnée en définition 3, les confiances des règles produites sont calculées conduisant aux résultats de la Table 3.

Table 3. Confiances des règles d'association générées

R_i	R_1	R_3	R_5	R_7	R_9
$Confiance(R_i)$	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

Si nous supposons un seuil de confiance minimal *minconf* =1, nous retenons seulement les règles dont la confiance est supérieure ou égale à *minconf*. Les règles ainsi sélectionnées sont données en Table 4. Ces règles sont d'abord utilisées pour construire les règles d'association sémantiques qui correspondent aux relations entre les noeuds du CP-Net. Ainsi, on déduit :

- de R_2: Studio → Toulouse et R_6: Studio → Paris: $Housing \rightarrow_{sem} City$
- de R_4: Suburbs → Toulouse : $Place \rightarrow_{sem} City$
- de R_7: Paris → Suburbs : $City \rightarrow_{sem} Place$
- de R_9: Studio →Suburbs : $Housing \rightarrow_{sem} Place$

Nous calculons alors le support de chaque règle sémantique. Les résultats sont donnés en Table 5.

Table 4. Règles d'association sélectionnées

R_2 : Studio → Toulouse
R_4 : Suburbs → Toulouse
R_6 : Studio → Paris
R_7 : Paris → Suburbs
R_9 : Studio →Suburbs

Table 5. Supports des règles d'association sémantiques

Housing $\rightarrow_{sem}$ City	1
Place $\rightarrow_{sem}$ City	0.5
Cit y $\rightarrow_{sem}$ Place	0.5
Housing $\rightarrow_{sem}$ Place	1

Nous retenons évidemment les règles *Housing* $\rightarrow_{sem}$ *City* et *Housing* $\rightarrow_{sem}$ *Place* dont le support égale 1. Deux associations existent entre les concepts *City* et *Place* avec un même support. Nous gardons aléatoirement l'une des deux règles correspondantes. Supposons *Place* $\rightarrow_{sem}$ *City* est retenue. Il est clair que retenir les trois règles ainsi obtenues mènerait à un cycle dans le graphe CP-Net. Pour éviter ceci, nous éliminons la règle la plus faible (C'est à dire possédant le plus petit support), soit *Place* $\rightarrow_{sem}$ *City*. En conclusion, les seules règles sémantiques sélectionnées sont les suivantes: *Housing* $\rightarrow_{sem}$ *City* et *Housing* $\rightarrow_{sem}$ *Place*. Les tables *CPT* sont alors associées aux nœuds correspondants, en utilisant la formule (4), ce qui mène au CP-Net document donné en Figure 2.

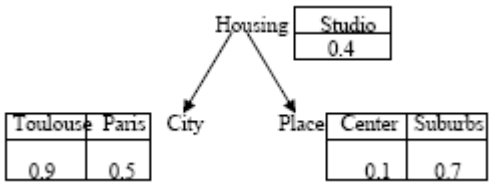

Fig. 2. Le CP-Net document

5 Conclusion

Nous avons présenté dans ce papier, une approche d'indexation conceptuelle basée sur les CP-Nets. Le formalisme CP-Net supporte naturellement concepts et associations contextuelles entre concepts, permettant ainsi une représentation plus riche et plus précise des documents. Ceci est susceptible de résoudre les problèmes de disparité et d'ambiguïté des termes en RI. En vue de pallier ces problèmes, nous

avons tenté de présenter des définitions formelles de contexte de terme, de concept-noeud, de relation entre concepts-noeuds, et de règles d'association sémantiques. Nous avons également présenté des techniques en vue d'une indexation conceptuelle graphique des documents. Notre contribution porte sur deux aspects principaux. Le premier consiste en l'indexation conceptuelle basée sur l'ontologie WordNet. L'approche n'est certes pas nouvelle mais nous avons proposé de nouvelles techniques pour identifier, pondérer et désambiguïser les termes. Le deuxième aspect de notre contribution consiste en une nouvelle approche d'organisation et de représentation de l'index conceptuel d'un document, en un graphe compact basé sur le formalisme CP-Net. Pour découvrir les associations entre concepts nous avons proposé une variante des règles d'association à savoir les règles d'association sémantiques. Les règles d'association sémantiques entre concepts permettent la découverte de relations dépendantes du contexte impliquant une représentation plus expressive du document.

En perspective, nous projetons de valider empiriquement notre approche sur une collection de test de RI.

Références

1. Agrawal R., Imielinski T., Swami A. : «Mining association rules between sets of items in large databases», In *Proceedings of the ACMSIGMOd Conference on Management of data*, Washington, USA, ACM Press, 1993, (p. 207–216).
2. Baziz M., Boughanem M., Aussenac-Gilles N. : «The Use of Ontology for Semantic Representation of documents», In *The 2nd Semantic Web and Information Retrieval workshop(SWIR), SIGIR 2004*, Sheffield UK., Ying ding, Keith van Rijsbergen, Iad Ounis, Joemon Jose (Eds.), 2004, (p. 38-45).
3. Boubekeur F., Boughanem M. , Tamine L. : «Towards Flexible Information Retrieval Based on CP-Nets». Dans : *Flexible Query Answering (FQAS)*, *Milan, Italie*, *07/01/06-10/06/06*, Henrik Legind Larsen, Gabriella Pasi, Daniel Ortiz-Arroyo (Eds.), World Scientific Publishing, Advances in Artificial Intelligence, p. 222-231, juin 2006.
4. Boutilier C., Brafman R., Hoos H., Poole D. : «Reasoning with Conditional Ceteris Paribus Preference Statements», In *Proceedings of UAI-1999*, (p.71–80).
5. Haddad M.H. : 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.
6. Khan L.R: Ontologie-based Information Selection*,* Phd Thesis, Faculty of the Graduate School, University of Southern California , 2000.
7. Krovetz R.: «Homonomy and Polysemy in Information Retrieval», In *the Proceedings of the COLING/ACL '97 conference.*
8. Lin D.: «An information-theoretic definition of similarity», In *Proceedings of 15th International Conference On Machine Learning* , 1998.
9. Resnik P. : «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, 1999, (p. 95-130).
10. Voorhees E.M.: «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 , 1993.
11. Yarowsky D. : «Unsupervised word sense disambiguation rivaling supervised methods», In *33rd Annual Meeting*, Association for Computational Linguistics, Cambridge, Massachusetts, USA, 1995, (p189-196).