Recherche d'Information Flexible Basée CP-Nets

Fatiha Boubekeur*, — Lynda Tamine-Lechani***

* *IRIT-SIG, Université Paul Sabatier,*
31062 Toulouse, France
boubekeu@irit. fr, tamine@irit. fr

** *Université Mouloud Mammeri*
15 000 Tizi-Ouzou, Algérie

RÉSUMÉ. *Ce papier décrit une approche de recherche d'information (RI) flexible fondée sur l'utilisation des CP-Nets (Conditional Preferences Networks). Le formalisme CP-Net est utilisé d'une part, pour la représentation graphique de requêtes flexibles exprimant des préférences qualitatives et d'autre part pour l'évaluation flexible de la pertinence des documents. Le raisonnement et l'inférence sur les préférences qualitatives n'étant pas aisés, nous devons quantifier les préférences. Nous proposons alors une approche de pondération automatique des requêtes CP-Nets. Cette pondération, correspondant à la quantification du CP-Net requête par des valeurs d'utilités, conduit à un UCP-Net. L'UCP-Net correspond à une requête booléenne pondérée. Une utilisation des CP-Nets est également proposée pour la représentation des documents dans la perspective d'une évaluation flexible des requêtes.*

ABSTRACT. *This paper describes a flexible approach for information retrieval (IR) based on the use of CP-Nets (Conditional Preferences Networks). The CP-Net formalism is used in one hand, for the graphical representation of flexible queries expressing qualitative preferences and in the other hand, for flexible relevance evaluation of documents. The reasoning and the inference on qualitative preferences not being easy, we must quantify the preferences. We propose an approach for automatically weighting the CP-Net queries. This weighting corresponds to the quantification of the CP-Net query by utility values, leading to a UCP-Net. The UCP-Net corresponds to a weighted Boolean query. CP-Nets are also used for representing documents in the prospect of flexible queries evaluation.*

MOTS-CLÉS : *RI flexible, modèle Booléen étendu, préférences utilisateur, CP-Nets, UCP-Nets.*

KEYWORDS : *Flexible IR, Extended Boolean model, user preferences, CP-Nets, UCP-Nets.*

1. Introduction

Le but principal d'un système de recherche d'informations (SRI) est de retrouver l'information considérée comme pertinente pour une requête utilisateur traduisant son besoin en information mais aussi ses préférences sur les informations recherchées. La pondération des termes de la requête (Buell et *al.*, 1981), (Bordogna et *al.*, 1991), (Pasi, 1999) a permis d'exprimer les préférences utilisateur sur les critères de recherche. Cependant, différentes sémantiques sont associées au poids (Crestani et *al.*, 1999) impliquant des définitions différentes de la fonction d'évaluation de la pertinence. Outre le problème de la sémantique du poids d'un terme, les poids numériques des requêtes forcent l'utilisateur à quantifier le concept qualitatif et vague d'importance. Cette tâche n'est pas évidente, en particulier si la requête exprime des préférences conditionnelles, d'une part, car il n'existe pas de bonne méthode pour pondérer correctement les termes de la requête, d'autre part, lorsque le nombre de valeurs sur lesquelles portent les préférences est élevé, il est quasiment impossible d'énumérer un poids valide pour tous les termes de la requête. De ce fait, des travaux se sont orientés vers l'utilisation de préférences qualitatives plus simples et plus intuitives, formulées à partir de termes linguistiques tels : *important, très important…* (Bordogna et *al.,* 1993), (Bordogna et *al.*, 1995). Cependant, le problème de la définition des poids numériques des termes est reporté sur la définition de la sémantique du concept flou *important* et des modulateurs linguistiques *très, peu, moyennement…*

Nous proposons, dans ce papier, une approche mixte d'expression des préférences utilisateur combinant l'expressivité et la simplicité du formalisme qualitatif à la puissance *calculatoire* du formalisme quantitatif. Nous nous intéressons particulièrement aux préférences conditionnelles. De telles formes de préférences, n'ont pas, à notre connaissance, été spécifiquement pris en charge dans les SRI existants. Une représentation qualitative, naturelle, simple et compacte de telles formes de préférences est supportée par les CP-Nets (Boutilier et *al.*, 1999). Nous utiliserons les CP-Nets comme outil d'expression des requêtes utilisateur flexibles (portant sur les préférences), puis nous proposons une méthode de pondération automatique de la requête. Cette pondération correspond à la quantification du CP-Net par des valeurs de préférence (ou valeurs d'utilité). L'extension des CP-Nets par association de valeurs d'utilités, conduit à un UCP-Net (Boutilier et al., 2001), correspondant à une requête pondérée correcte. La requête CP-Net ainsi pondérée doit être évaluée. Nous proposons une approche d'évaluation flexible des requêtes basée sur la sémantique des CP-Nets.

Le papier est organisé comme suit : en section 2, nous présentons les principes de base des CP-Nets et des UCP-Nets. La section 3, traite de la recherche d'information flexible basée sur les CP-Nets. Nous y présentons notre approche pour la pondération automatique des CP-Nets ainsi que notre méthode d'évaluation flexible des requêtes CP-Nets.

2. Les CP-Nets

2. 1. Concepts de base

Un CP-Net est un graphe orienté acyclique, ou *DAG*[1], $G = (V, E)$, où V est un ensemble de nœuds $\{X_1, X_2, X_3, \ldots X_n\}$ qui définissent les variables de préférence et E un ensemble d'arcs orientés entre les nœuds, traduisant des relations de dépendances préférentielles entre ces nœuds. Toute variable X_i du graphe est instanciable dans un domaine de valeurs $Dom(X_i) = \{x_{i1}, x_{i2}, x_{i3} \ldots\}$. Le prédécesseur d'un nœud X dans le graphe est dit son parent. On note $(Pa(X))$ l'ensemble des parents de X. ($X \cup Pa(X)$) constitue une famille du CP-Net. A chaque variable X du CP-Net, on associe une table de préférences conditionnelles ($CPT(X)$) spécifiant un ordre de préférence total sur les valeurs x_i de X étant donné chaque instance de ses parents. Pour un nœud racine, la table *CPT* spécifie un ordre de préférence inconditionnel sur les valeurs du nœud.

Un CP-Net induit un graphe complet de préférences ordonné, construit sur l'ensemble de ses alternatives. Une alternative du CP-Net est un élément du produit cartésien des domaines de valeurs de ses différents nœuds. Elle est interprétée comme une conjonction de ses éléments.

2. 2. Les UCP-Nets

Un CP-Net ne permet pas de comparer et d'ordonner toutes ses alternatives. Pour ce faire, on doit quantifier les préférences. Un UCP-Net étend un CP-Net en autorisant la quantification des noeuds par des valeurs d'utilité (ou facteurs d'utilité). Un facteur d'utilité, $f_i(X_i, Pa(X_i))$, associé à un nœud X_i étant donné l'ensemble de ses parents $Pa(X_i)$, (on notera plus simplement $f_i(X_i)$), est une valeur réelle qui définit l'ordre de préférence d'une instance de X_i étant donné une instance de $Pa(X_i)$.

Définir un UCP-Net revient à définir pour chaque famille de nœuds $\{X_i, Pa(X_i)\}$ du CP-Net, un facteur d'utilité $f_i(X_i)$. Ces facteurs servent à quantifier la table *CPT* dans le graphe. Sémantiquement, on traite les différents facteurs d'utilité comme étant généralisés additifs indépendants *(GAI)* (Formellement : si $V = \{X_1, \ldots, X_n\}$ est l'ensemble des nœuds du CP-Net, alors l'utilité globale d'une instance de V est donnée par : $u(V) = \sum_i f_i(X_i)$).

Pratiquement, étant donné un *DAG* quantifié G, X une variable de G, $Pa(X)$ l'ensemble de ses parents, Y_i les descendants de X et $x_1, x_2 \in Dom(X)$, en définissant : $Minspan(X) = \min_{x1, x2 \in Dom(X)} (\min_{p \in Dom(Pa(X))} (|f_X(x_1, p) - f_X(x_2, p)|))$ et $Maxspan(X) = \max_{x1, x2 \in Dom(X)} (\max_{p \in Dom(Pa(X))} (|f_X(x_1, p) - f_X(x_2, p)|))$, On dira que X domine ses descendants si : $Minspan(X) >= \sum_i Maxspan(Y_i)$.

1. Direct Acyclic Graph

Alors *G* est un UCP-Net valide si toute variable *X* de *G* domine ses descendants (Boutilier et *al,* 2001).

3. Recherche d'information flexible basée CP-Nets

Les préférences utilisateur sont exprimées sur des variables (représentant des concepts). Chaque variable est définie sur un domaine de valeurs (une valeur est un terme de la requête). Pour chaque variable, l'utilisateur spécifie toutes ses dépendances préférentielles à partir desquelles un graphe CP-Net est construit. (On supposera dans ce qui suit que la graphe résultant est un *DAG*). La requête CP-Net est ensuite pondérée par des facteurs d'utilité (poids de préférence). Notre processus de pondération automatique de la requête CP-Net est basé sur la propriété de dominance énoncée plus haut (section 2.2.). Nous le présentons ci-après.

3.1. Génération du UCP-Net ou la pondération automatique de la requête

Soit *X* un nœud de la requête CP-Net, tel que $|Dom(X)| = k$, et soit $u(i)$ le degré de préférence d'ordre *i* (en supposant un degré de préférence croissant lorsque *i* croît) sur les valeurs de X :

Pour tout nœud feuille X, nous générons les utilités de *X* comme suit : $u(1) = 0$ *et* $u(i) = u(i - 1) + (1 / (k - 1)), \quad \forall 1 < i <= k.$

Pour tout nœud interne *X* du CP-Net (*X* n'est pas un nœud feuille), on calcule la quantité : $S = \sum_i Maxspan(B_i)$ où les B_i sont les descendants de *X*. Comme *X* domine ses descendants on a *: Minspan(X) >= S*. Plusieurs valeurs répondent à la condition, nous choisirons la plus petite soit *S* et poserons *Minspan(X) = S*. Nous générons alors les utilités de *X* comme suit : $u(1) = 0$ et $u(i) = u(i - 1) + S, \quad \forall 1 < i <= k.$

On calcule alors *Minspan(X)* de manière triviale par $|u(i +1) - u(i)|$ et *Maxspan(X)* par $|u(k) – u(1)|$.

Les valeurs d'utilité obtenues pouvant être supérieures à 1 (cas des nœuds internes), nous proposons une normalisation des facteurs d'utilité individuels du CP-Net et des utilités globales de chacune de ses alternatives en divisant chaque valeur d'utilité du CP-Net par l'utilité globale de l'alternative la plus préférable.

3. 2. Evaluation de la requête CP-Net

Le processus de recherche est lancé sur l'ensemble des valeurs de chacun des nœuds du UCP-Net requête *Q* sans tenir compte de la pondération au préalable. Le résultat est une liste de documents pertinents probables. Chaque document *d* obtenu, est alors défini dans l'espace des termes de la requête *Q*, ce qui permet de le représenter par un CP-Net dont la structure est identique à celle du CP-Net requête.

Le document *d* (respectivement la requête *Q)* est alors interprété comme une disjonction de conjonctions, chacune d'elles étant construite sur l'ensemble des éléments du produit cartésien $Dom(X_1) * Dom(X_2) *... * Dom(X_n)$ où X_i $(1 \leq i \leq n)$ est un nœud du CP-Net requête, soit :

$$d = \vee_{j_i} (\wedge_i t_{i,j_i}^{p_{i,j_i}}), \qquad 1 \leq i \leq n \quad et \quad 1 \leq j_i \leq |Dom(X_i)| \qquad [1]$$

$$Q = \vee_{j_i} (\wedge_i t_{i,j_i}^{f_{i,j_i}}), \qquad 1 \leq i \leq n \quad et \quad 1 \leq j_i \leq |Dom(X_i)| \qquad [2]$$

Où $t_{i,j_i} \in Dom(X_i)$, p_{i,j_i} est le poids de t_{i,j_i} dans *d* (généralement fonction de sa fréquence d'occurrence, ce poids est nul si t_{i,j_i} n'appartient pas à *d*) et f_{i,j_i} est le poids du terme t_{i,j_i} (son utilité) dans *Q* étant donnée une valeur de son parent.

Soit $m = |Dom(X_1)| * |Dom(X_2)| *... *|Dom(X_n)|$, *e*n posant : $\wedge_i t_{i,j_i} = T_k$, les représentations [1] et [2] sont ramenées à :

$$Q = \vee_k T_k^{U_k} = \vee T_k^{U_k} \qquad 1 \leq k \leq m \qquad [3]$$

$$d = \vee_k T_k^{S_k} = \vee T_k^{S_k} \qquad 1 \leq k \leq m \qquad [4]$$

Où $U_k = \sum_i f_{i,j_i}$ (les facteurs X_i étant *GAI*) *et* S_k est la valeur agrégée des poids p_{i,j_i} introduits en [1]. Pour calculer ce poids, nous utilisons une propriété clé des CP-Nets : un nœud parent est plus important que ses descendants. Nous associons une importance de position G_X aux nœuds *X* du CP-Net selon leurs niveaux dans le graphe. Pour tout nœud feuille *X*, on donne : $G_X = 1$; Pour tout autre noeud *X* tel que B_l sont les descendants de *X* et G_{B_l} leurs importances de position respectives, on a : $G_X = Max_l\, G_{B_l} + 1$. Le poids S_k introduit en [4] est la moyenne pondérée des poids p_{i,j_i} des termes t_{i,j_i} composant T_k , soit : $S_k = \sum_i p_{i,j_i} * G_{X_i} / \sum_i G_{X_i}$, où X_i est le nœud contenant le terme (t_{i,j_i}) de *d.*

Pour évaluer la pertinence du document *d* pour la requête pondérée *Q*, nous proposons d'adapter et d'utiliser le minimum pondéré (Dubois et al, 1986), (Yager, 1987) comme suit :

Soit U_k le poids d'importance de T_k dans *Q*, $F(d, T_k) = S_k$, le poids de T_k dans le document *d*, on note $RSV_{T_k}(F(d, T_k), U_k)$ la fonction d'évaluation de T_k par rapport au document *d*. Les différentes conjonctions pondérées $T_k^{U_k}$ étant liées par une disjonction, ce qui donne :

$$RSV_{T_k}(F(d, T_k), U_k) = Min (S_k, U_k) \qquad [5]$$

La pertinence globale du document *d* pour la requête *Q* est donc:

$$RSV(d, Q) = Max_k (Min (S_k, U_k)) \qquad [6]$$

4. Conclusion

Nous avons présenté dans cet article une approche de RI flexible basée sur les CP-Nets. Cette approche est fondée d'une part sur l'expression de requêtes flexibles

traduisant les préférences d'un utilisateur, utilisant les CP-Nets. Le formalisme utilisé est graphique et qualitatif ce qui permet une formulation naturelle et intuitive et une représentation simple et compacte des préférences. Le formalisme qualitatif possède une puissance d'expression élevée mais décline en puissance de calcul. Nous avons proposé de l'allier aux utilités. Les utilités, représentant des poids d'importance conditionnelle de termes de la requête, sont calculés automatiquement. L'utilisateur est ainsi déchargé de cette tâche fastidieuse et non moins improbable de pondération, et les poids générés sont corrects puisque basés sur des fondements théoriques établis (validité d'un UCP-Net). D'autre part, cette approche est fondée sur l'évaluation flexible, des requêtes dans la sémantique CP-Net basée sur l'utilisation d'un opérateur d'agrégation flexible, en l'occurrence le minimum pondéré, que nous avons adapté à la sémantique CP-Net. Notons cependant que le graphe CP-Net introduit par l'utilisateur peut être incorrect (inconsistant). Des outils d'aide à la formulation et des outils de correction automatique seraient nécessaires pour garantir la validité des descriptions fournies par l'utilisateur.

6. Bibliographie

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

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

Bordogna G., Pasi G., Linguistic aggregation operators of selection criteria in fuzzy information retrieval, *International Journal of Intelligent Systems*, 10, 1995, p. 233-248.

Boutilier C., Brafman R., Hoos H., and Poole D., Reasoning with Conditional Ceteris Paribus Preference Statements, *Proc. of UAI-1999*, p.71–80.

Boutilier C.,, Bacchus F., and Brafman R., UCP-Networks : A directed graphical representation of conditional utilities, *UAI'2001*, p. 56–64.

Buell D. A., Kraft D. H., A model for a weighted retrieval system, *Journal of the American Society for Information Science*, 32[3], May 1981, p. 211-216.

Crestani F., Pasi G., Soft information retrieval : Applications of fuzzy Set Theory and Neural Networks, "Neuro Fuzzy Techniques For Intelligent Information Systems". N. Kasabov and Robert Kozmz Editors, Physica –Verlag, Springer-Verlag Group, 1999, p. 287-313.

Dubois D., Prade H., Weighted minimum and maximum operations in fuzzy set theory. *Information Sciences*, 39, 1986, p. 205-210.

Pasi. G., A logical formulation of the boolean model and of weighted boolean model, Proceedings of the *Workshop on Logical and Uncertainty Models for Information Systems*, London, UK, 1999, p. 1–11.

Yager R., A note on weighted queries in information retrieval systems, *Journal of American Society for Information Science*, 38[1], 1987, p. 23-24.