ON USING GRAPHICAL MODELS FOR SUPPORTING CONTEXT AWARE INFORMATION RETRIEVAL

Lynda Tamine-Lechani, Fatiha Boubekeur, Mohand Boughanem
Institut de recherche en informatique de Toulouse (France)
lechani@irit.fr, boubekeur@irit.fr, bougha@irit.fr

Keywords: Contextual information retrieval, Graphical models, Influence diagrams, CP-Nets

Abstract: It is well known that with the increasing of information volumes across the Web, it is increasingly difficult for search engines to deal with ambiguous queries. In order to overcome this limit, a key challenge in information retrieval nowadays consists in enhancing an information seeking process with the user's context in order to provide accurate results in response to a user query. The underlying idea is that different users have different backgrounds, preferences and interests when seeking information and so a same query may cover different specific information needs according to who submitted it. This paper investigates the use of graphical models to respond to the challenge of context aware information retrieval. The first contribution consists in using CP-Nets as formalism for expressing qualititative queries. The approach for automatically computing the preference weights is based on the predominance property embedded within such graphs. The second contribution focuses on another aspect of context, namely the user's interests. An influence-diagram based retrieval model is presented as a theoretical support for a personalized retrieval process. Preliminary experimental results using enhanced TREC collections show the effectiveness of our approach.

1 INTRODUCTION

It is well known that users often submit very short queries, usually between two and three words in length due to cognitive or linguistic limitations (Jansen et al., 2000). This leads to ambiguous information needs for what keyword based search engines provide inaccurate results (Nunberg, 2003). One effective technique for eliciting queries is query expansion (Efthimiadis, 1996). The underlying idea consists in increasing the length of user's queries via automatic and interactive techniques. However query expansion techniques are less efficient within a huge information context due to the task complexity, the amount of extra time required to achieve it and the user's lack of additional cognitive resources (Ruthven and Lalmas, 2003). Another approach for eliciting user's information need is based on Ingwersen's cognitive theory (Ingwersen, 1996) that suggests that during the information retrieval process "*additional cognitive structures concerned with domain, tasks/interests, and problems/goals or uncertainty are present*". Therefore, we consider that information retrieval (IR) takes place in a context determined by various elements such as users' goals, preferences and interests that have a huge impact on the user's relevance statement of the information returned in response to his information need. Based on this finding, numerous works in IR address nowadays two critical questions (Crestani and Ruthven, 2007): (1) what aspects of context can we recognise (2) how should context be utilised within a retrieval system to improve search performance?
This paper attempts to respond to the challenge of contextual IR by considering two important contextual factors: (1) user's preferences expressed within a query (2) user's domains of expertise and knowledge. More particularly, the main research questions addressed are the following:

- How to model the context-dependent information need specification? In order to achieve this goal, the user's preferences have to be formally represented to allow accurate document relevance estimation.

- How to model a query evaluation process embedded within a multi-domain user's expertise? Domains of expertise and knowledge consitute a context that connects concepts via semantic relations that can be exploited in order to overcome the basic out-context statistical information processing.

The effectiveness of these models are therefore weakly related to their flexibility, intended by their capability to deal with vagueness and uncertainty. The vagueness concerns mainly the user's information need representation and the user's interests description via the user-system interaction. Uncertainty concerns the user's relevance statement expressed through the usefulness of the information according to the specified query and the prevalent user's interests. For these reasons, we have been attracted by exploring the use of graphical models (Jensen, 2001) to suport context aware information retrieval processes. Graphical models are families of probability or utility value distributions defined in terms of directed or undirected graphs. The nodes in the graph are identified with random variables and joint probability/utility distributions are defined by appropriate functions applied on subsets of nodes.
More precisely, in order to answer the above questions, we investigate in one hand, the use of CP-Nets (Boutilier et al., 1999) to supporting information need elicitation and, in the other hand, influence diagrams (Jensen, 2001) to solving the problem of domain-dependent query evaluation. In section 2, we discuss related work and then we address two main issues. The first one, developped in section 3, concerns the specification of the user's query formulation via CP-Nets. In order to achieve this goal, we present the CP-Net based model for representing the user's preferences. The section 4 addresses the information personalization problem via influence diagrams. The qualitative and quantitative components of the model are presented and then preliminary experimental results are discussed. Section 5 concludes the paper.

2 Related work

Numerous studies have proved that queries addressed to search engines are under-specified (Jansen et al., 2000). Furthermore, same queries cover generally several different intentions, depending on the users they submitted them. Contextual IR is an active research area that aims to outcome the limit of basic keyword IR models by dealing with user's context expressed via related user's preferences, interests and goals. This research has been carried out in order to achieve different goals: eliciting user's information needs, user modelling, personalizing search based on user's interests, enhancing queries with local user's context etc. In this paper, we focus on the advantage behind the utilization of graphical models to achieve two main goals among the above ones cited: eliciting user's information needs, personalizing search based on user's interests.
Preference elicitation is the process of extracting preference information from a user. This involves tools for modelling and representing such information. In this context, fuzzy set theory has been widely used by many authors leading to fuzzy retrieval models (Bordogna et al., 1991; Kantor, 1981). Fuzzy queries allow better representation of user preferences by means of query term weighting. The query term weights could be either numeric or linguistic values. Numerical valuation of term weights is a difficult task that forces the users to quantify the importance of each query term according to his real information need. In contrast linguistic valuation of term weights implies a more intuitive query expression. In (Bordogna and Pasi, 1993; Kraft et al., 1994) fuzzy linguistic approaches have been proposed for modelling flexible queries.
Numerous other works focused on customizing information according to the user's interests. In (Speretta and Gauch, 2005), the authors model the user's interests as weighted concept hierarchies extracted from the user's search history. Personalization is carried out by re-ranking the top documents returned to a query using an RSV[1] function that combines both similarity document-query and document-user. The profiling component of ARCH (Sieg et al., 2004) manages a user's profile containing several topics of interest of the user. Each of them is structured as a concept hierarchy derived from assumed relevant documents using a clustering algorithm in order to identify related semantic categories. Personalization is achieved via query reformulation based on information issued from selected and unselected semantic categories. WebPersonae (Gowan, 2003) is a browsing and searching assistant based on web usage mining. The different user interests are represented as clusters of weighted terms obtained by recording documents of interest to the user. The relevance of a document is leveraged by its degree of closeness to each of these clusters. In (Liu and Yu, 2004) user profiles are used to represent user's interests. A user profile consists of a set of categories, and for each category, a set of weighted terms. Retrieval effectiveness is improved using voting-based merging algorithms that aim to re-rank the documents according to the most related categories

[1] Relevance Status Value

to the query. Recently, extensions of the Page Rank algorithm (Qiu and Cho, 2006; Haveliwala, 2002) have been proposed. Their main particularity consist in computing multiple scores, instead of just one, for each page, one for each topic listed in the Open Directory.

3 Using CP-Nets for better eliciting information needs

The idea behind our approach is to offer to the user an intuitive and graphical formalism to express his conditional preferences in a qualitative and intuitive manner using the CP-Net formalism. We then propose an accurate algorithm based on UCP-Net features to automatically translate the user qualitative preferences into numerical query weights.

3.1 CP-Nets

CP-Nets were introduced in (Boutilier et al., 1999) as graphical models for compact representation of qualitative preference relations. They exploit conditional preferential dependencies in the structuring of the user preferences under the Ceteris-Paribus assumption (all else being equal). 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 variables of interestand 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. A set $\{X_i, Pa(X_i)\}$ defines a CP-Net family. To each variable X_i of the CP-Net is attached a Conditional Preference Table $CPT(X_i)$ 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. We call an alternative of the CP-Net each element of the set $Dom(X_1)xDom(X_2)...Dom(X_n)$.

3.2 UCP-Nets

A UCP-Net (Boutilier et al., 2001) extends a CP-Net by quantifying the CP-Net nodes with conditional utility values (utility factors). A conditional utility factor $f_i(x_i^j, p)$, where p is an assignment of $Pa(X_i)$, (we simply write $f_i(x_i^j)$), is a real value attached to each $x_i^j \in Dom(X_i)$ in order to express a conditional preference order over x_i^j value being given an instance of $X_i's$ parents $Pa(X_i)$. Utility factors traduce quantitative ordering of the qualitative preferences expressed in the corresponding CP-Net. Therefore defining a UCP-Net amounts to define for each set $\{X_i, Pa(X_i)\}$of the CP-Net, the utility factors $f_i(x_i^j)$ for each $x_i^j \in Dom(X_i)$. These factors are used to quantify the CPTs in the graph. The utility factors are generalized additive independent (GAI) (Boutilier et al., 2001). Formally, for a UCP-Net $G = \{V, E\}$ where $V = \{X_1, X_2, \ldots, X_n\}$, we compute the global utility of a given alternative A, denoted $u(A)$ as follows:

$$u(A) = \sum_i f_i(x_i^j) \quad (1)$$

The validity of a UCP-Net is based on the principle of predominance defined as follows (Boutilier et al., 2001): Let $G = \{V, E\}$ a quantified CP-Net. G is a valid UCP-Net if:

$$\forall\, X \in V,\ Y_i \in V / X = Pa(Y_i), \quad Minspan(X) \geq \sum_i Maxspan(Y_i) \quad (2)$$

Where,

$$Minspan(X) = min_{x_1, x_2 \in Dom(X)} \left(min_{p \in Dom(Pa(X))}(|f_x(x_1, p) - f_x(x_2, p)|)\right) \quad (3)$$

$$Maxspan(X) = max_{x_1, x_2 \in Dom(X)} \left(max_{p \in Dom(Pa(X))}(|f_x(x_1, p) - f_x(x_2, p)|)\right) \quad (4)$$

3.3 The problem

A user query **Q** is generally expressed by a set of keywords (terms) connected with boolean operators. Query terms are weighted (Bordogna et al., 1986) so as to allow expressing user preferences on the search criteria. Considering the query **Q**, the IR system computes, for each document **D**, its relevance status value, called $RSV(Q, D)$, that measures its degree of matching to the query and then ranks the retrieved documents in decreasing order of their RSV. The RSV could be defined as:

$$RSV(Q, D) = \Psi(f(u_i, a_i)) \quad (5)$$

where u_i is the weight associated with the query term t_i in **Q**, a_i is the weight associated to the term t_i in the document D, $f(u_i, a_i)$ is a function that matches term-term in both **D** and **Q** in order to compute the partial relevance of the document **D** according to this query term, Ψ is an aggregation function that combines all the partial scores of **D** in a global relevance score namely $RSV(Q, D)$. This formula highlights that a good scoring depends on:

1. a *good* aggregation functionΨ,
2. an *efficient* partial relation f combining query term weights in both **D** and **Q**,
3. a *good* weigthing of the terms in both **D** and **Q**.

The following section details our approcah to address the third above point.

3.4 The model

3.4.1 Representing user's preferences using UCP-Nets

Our goal at this level, is to build CP-nets that represent and allow automatic quantification of user's preferences. For this aim, the user preferences are first expressed using concepts represented by variables. Each variable is defined on a domain of values (a value is therefore a query term). For each variable, the user should specify all of its preferential dependencies from which a CP-Net graph is built. The CP-Net query is then weighted by preference weights corresponding to utility factors. Our automatic weighting process is based on the predominance property (Boutilier et al., 2001). We present it in the following: Let $G = \{V, E\}$ be a CP-Net query Q which expresses the qualitative conditional preferences of a user on n concepts (variables), X be a variable of Q, such as $|Dom(X)| = k$, and let $u(i)$ be the i^{th} preference order on $X's$ values (one assume $u(i)$ growing when i grows): For any leaf node X, we simply generate the utilities as uniform preference orders over the set $[0..1]$ as follows:

$$u(i) = \begin{cases} 0 \quad if\ i = 1 \\ u(i-1) + \frac{1}{k-1} \\ \forall\ 1 < i \leq k \end{cases} \tag{6}$$

For any internal node X (X is not a leaf node), we compute $S = \sum_i Maxspan(B_i)$ where B_i represents the descendants of X. The predominance property imposes that $Minspan(X) \geq S$. Several values answer the condition correctly, the smallest one S is choosen, so $Minspan(X) = S$. The utilities are computed as follows:

$$u(i) = \begin{cases} 0 \ \ if\ i = 1 \\ u(i-1) + S \\ \forall\ 1 < i \leq k \end{cases} \tag{7}$$

therefore we easily compute:

$$Minspan(X) = |u(i+1) - u(i)|\ and \\ Maxspan(X) = |u(k) - u(1)| \tag{8}$$

The utility values obtained can be higher than 1 (particularly in the case of internal nodes), we propose a normalisation of the individual utility factors of the CP-Net and of the total utilities of each alternative as follows: For each CP-Net node X_j, let $Max(X_j) = max_i(u(i))$ be the highest preference order on X_j values, then:

$$\forall\ X_j,\ \forall u(i),\ 1 \leq i \leq Dom(X_j),\ \ u(i) = \frac{u(i)}{\sum_j Max(X_j)} \tag{9}$$

3.4.2 Illustration

Let us consider the following user need: *"I am looking for housing in Paris or Lyon of studios or residence Hall (RH) type. Knowing that I prefer to be in Paris rather than to be in Lyon, if I should go to Paris, I will prefer being in a residence hall(we will treat residence hall as a single term), whereas if I should go to Lyon, a studio is more preferable to me than a room in residence hall. Moreover the Centre town of Paris is more preferable to me than its suburbs; whereas if I must go to Lyon, I will rather prefer to reside in suburbs that in the centre"*.
Figure 1 illustrates the CP-Net corresponding to the above query. The variables of interest are $V = \{City, Housing, Place\}$ where $Dom(City) = \{Paris, Lyon\}$, $Dom(Housing) = \{RH, Studio\}$ and $Dom(Place) = \{Centre, Suburbs\}$. In addition, $CPT(City)$ specifies that *Paris* is unconditionally preferable to *Lyon* (we denote 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 $RH \succ Studio$). Following the approach de-

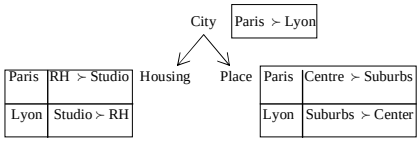

Figure 1: A CP-Net quey

scribed above, we obtain the UCP-Net query given in Figure 2 below.

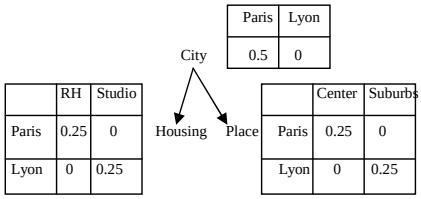

Figure 2: A UCP-Net quey

Appropriate graph based similarity measures could then be performed in order to identify the accurate documents being relevant for such CP-Net queries (Boubekeur et al., 2007).

4 Using influence diagrams for personalizing information

At this level, the basic underlying idea behind our contribution is to substitute the basic RSV function which measures the degree of matching document-query, by a function leveraged by the user's domains of interest. To the best of our knowledge, personalized information retrieval has not been addressed in earlier works as a decision-making problem by means of a utility theory. It is a novel direction that we explore in this paper at both theoretical and empirical levels.

4.1 Influence diagrams: an extension of Bayesian Networks

A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest (Jensen, 2001). A Bayesian network uses qualitative and quantive components to model and manipulate n-dimensional probability distributions. The qualitative component is carried out through a Directed Acyclic Graph (DAG), $G =< V, E >$ where each node in $X_i \in V$ encodes the random variable of interest and E encodes the relationships among these variables. We note $Pa(X_i)$ the parent set of X_i in G. The quantitative component outlines the estimation of the conditional dependencies among the variables. More precisely, for each variable $X_i \in V$, is attached conditional probability distributions $p(X_i/pa(X_i))$ where $pa(X_i)$ represents any combination of the values of the variables in $Pa(X_i)$. The inference of new sources of evidence is possible using the joint distribution low:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} (p(X_i/pa(X_i)) \quad (10)$$

An influence diagram (Shachter, 1988) is an extension of the Bayesian probabilistic model for solving decision making problems. In practice, an influence diagram is represented by an acyclic DAG containing three types of nodes (chance, decision and utility nodes) and two types of arcs (influence and informative arcs). The dependencies between chance nodes, representing random variables, is carried out using classical Bayesian probability distributions. For each utility node U related to a decision node D is attached a real-valued fuction over $pa(U)$. Given a particular situation, the diagram evaluation is carried out using an evidence propagation algorithm which aims to determine the optimal decision utility.
Prior works have specified simple Bayesian retrieval models dealing with unstructured (Acid et al., 1988) and structured document collections (Campos et al., 2004).

4.2 The problem

Intuitively, the problem of personalizing IR may be expressed basically as follows:
Given a query **Q**, the IR problem is to rank documents **D** according to their relavance to the information need of the user **U**. From the probabilistic point of view, the goal is to find the *a posteriori* most likely documents for which the probability of relevance of the document **D** considering the query **Q** and the user **U**, noted $p(d/q,u)$, is highest. By Bayes' law,

$$p(d/q,u) = \frac{p(q/d,u)p(d/u)}{p(q/u)} \quad (11)$$

where d, q and u are the random variables associated to respectively D, Q and U. As the denominator $p(q/u)$ is a constant for a given query and user, we can use only the numerator in order to rank the documents. Thus we define the RSV of a document as:

$$RSV_U(\mathbf{Q},\mathbf{D}) = p(q/d,u)p(d/u) \quad (12)$$

The first term of equation (2) is query dependent reflecting the closeness of the document **D** and the query **Q** according to the user **U**. The second term is query independent, highlighting the usefulness of the document to the user. This may express the suitability of the document to the whole domains of interest of the user when seeking information. In the case that we state that the user is modelled using a set of topics $C_1, C_2, ..., C_n$, the formula (2) gives:

$$RSV_U(Q,D) = p(q/d,c_1,c_2,...,c_n)p(d/c_1,c_2,...,c_n) \quad (13)$$

where c_i refers to a random variable associated to the user's interest C_i. The formula (3) highlights that:

1. two key conditions are prevalent when computing the relevance of documents : (1) relevance condition that ensure that the selected documents are close to the query, (2) the usefulness condition that ensure that the selected documents are consistent with the user's topics of interest,
2. maximum likelihood of a document is achieved when maximizing the coverage of the information according to the different topics. The user may choose the degree of relevance to integrate either all or a sublist of topics of interest during the personalization process.

By considering this manner of addressing the information personalization problem in the context of multi-user interests, we are hence attracted by formulating it in a mathematical model based on a utility theory supported by ID wich are extension of

Bayesian models. The problem is globally expressed through $ID(D,C,\mu)$:

- document variable set $D = \{d_1, d_2, ..., d_n\}$ where n is the number of documents in the collection,
- user's interests variable set $C = \{c_1, c_2, ..., c_u\}$ where u is the u^{th} topic of interest,
- utility set $\mu = \mu_1, \mu_2, ..., \mu_u$ where μ_j expresses the utility of the positive decision r about the relevance of a document **D** according to the user interest C_i, noted below $\mu(r/c_j)$. r is a decison variable within the set $R = \{r_1, r_2, ..., r_n\}$

The problem of information personalization takes then the form of ordering the documents $D_i \in D$ according to $\mu_\Omega(D_i) = \Psi(\mu_1, \mu_2, ..., \mu_u)$ where Ψ is an appropriate aggregation operator that combines evidence values from $C_1, C_2, ..., C_u$. With respect to the probabilistic view illustrated above, the problem takes form of:

$$RSV_U(Q,D) = \Psi_{j=1..u}(\mu(r/c_j)p(q/d,c_j)) \quad (14)$$

The following section gives formal details of our personalized information retrieval based on the above specification.

4.3 The model

The model topology is presented in figure (3).

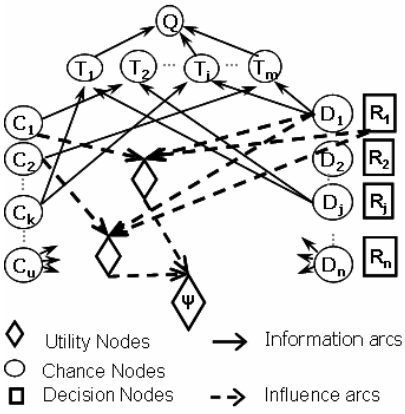

Figure 3: The diagram topology

Following the decision theoritical support of our approach, the personalized $RSV_U(Q,D)$ measures the accuracy of the decisions related to the relevance of the documents to be presented according to the query and the whole user interests. More precisely, given a query Q, the retrieval process starts placing the evidence alternatively in each observed document node then, the inference process is run as in a making-decision problem, by maximizing a re-ranking utility measure. We propose the following mapping fuction which ranks the documents according to the quotient between the expected utility of retrieving them and the expected utility of not retrieving them, computed as:

$$RSV_U : \begin{cases} R \longrightarrow R \\ RSV_U(Q,D_j) \mapsto \frac{EU(r/d_j)}{EU(\overline{r}/d_j)} \end{cases} \quad (15)$$

where $EU(r/d_j)$ (resp. $EU(\overline{r}/d_j)$) is the expected utility of the decision *"D_j is relevant, to be presented"* (resp.*"D_j is irrelevant, not to be presented"*)

$EU(r/d_j)$ is computed as follows (when assuming that the prior probabilities $p(d_j)$ and $p(c_k)$) are equal):

$$EU(r/d_j) = \Psi_{k=1..u}[\mu(r/c_k) * p(q/d_j, c_k)] \quad (16)$$

By applying the joint law and assuming that documents and user's interests are independent, and terms are also independent $EU(r/D)$ is computed as:

$$EU(r/D_j) = \Psi_{k=1..u}\, \mu(r/c_k) * \left(\sum_{\theta^s \in \theta} p(q/\theta^s) * \prod_{T_i \in Q \cap (D_j \cup C_k)} p(\theta_i^s/d_j) * p(\theta_i^s/c_k)\right) \quad (17)$$

$EU(\overline{r}/D_j)$ is consequently computed as:

$$EU(r/D_j) = \Psi_{k=1..u}\, \mu(\overline{r}/c_k) * \left(\sum_{\theta^s \in \theta} p(q/\theta^s) * \prod_{T_i \in Q \cap (D_j \cup C_k)} p(\theta_i^s/d_j) * p(\theta_i^s/c_k)\right) \quad (18)$$

where d_j traduces, as in the Turtle model (Turtle and Croft, 1990), that the document D_j has been observed and so introduces evidence in the diagram, all the remaining document nodes are set to $\overline{d_j}$ alternatively to compute the posterior relevance. Similarly, c_k and $\overline{c_k}$ express respectively that the user's interest C_k is observed or not observed, θ represents the whole possible configurations of the terms in $pa(Q)$, θ^s the s order configuration, and θ_i^s the s order configuration of term T_i in $pa(Q)$, Ψ an aggregation operator, $\mu(r/c_k)$ is the utility of the decision related to state that the document is relevant considering the user's interest C_k, $p(q/\theta^s)$ is the probability that the query Q is satisfied considery the configuration of it's parents, $p(\theta_i^s/d_j)$ and $p(\theta_i^s/c_k)$ are respectively the probability of relevance of term T_i in the configuration θ_i^s considering the document d_j and the user's interest c_k. The quantitative components $\mu(r/c_k)$, $p(q/\theta^s)$, $p(\theta_i^s/d_j)$ and $p(\theta_i^s/c_k)$ are specified below.

The utility value. The utility node joins an observed user's interest C_k to the decision related to the presentation of an observed document D_j. According to this, an utility value expresses the degree of the closeness between the document D_j and the user's interest C_k. We propose the following formula to compute $\mu(r_j/c_k)$:

$$\mu(r_j/c_k) = \frac{1+\sum_{T_i \in D_j} nidf(T_i)}{1+\sum_{T_i \in D_j - C_k} nidf(T_i)} \quad (19)$$

where $nidf(T_i)$ is the normalised idf of the term T_i, $\mu(\overline{r_j}/c_k)$ is computed as: $\mu(\overline{r_j}/c_k) = \frac{1}{\mu(r_j/c_k)}$

Computing $p(Q/pa(Q))$. the query is a leaf node that has as many parents as terms are belonging to it's representation, noted by $Pa(Q)$. Therefore, it should store 2^k configuration, k being the number of parents. Taking into account only the positive configuration terms parents $R(pa(Q))$ (noted further θ), we can compute the probability function attached to a query node using the *noisy-Or* aggregation operator (Pearl, 1988) such as:

$$p(Q/pa(Q)) = \begin{cases} 0 \; if \; (Pa(Q) \cap R(Pa(Q))) = \oslash \\ \frac{1-\prod_{T_i \in R(Pa(Q))} nidf(T_i)}{1-\prod_{T_i \in Pa(Q)} nidf(T_i)} \; otherwise \end{cases} \quad (20)$$

Computing $p(t_i/d_j)$ and $p(t_i/c_k)$. In each term node T_i, a probability function $p(t_i/d_j, c_k)$ is stored. Assuming the independency hypothesis between the document and each of the user's interests, $p(t_i/d_j, c_k)$ is computed as: $p(t_i/d_j, c_k) = p(t_i/d_j) * p(t_i/c_k)$. The probability that a term accurately describes the content of a document and user's interest can be estimated in several ways. We propose:

$$p(t_i/d_j) = \begin{cases} \frac{wtd(i,j)}{\sum_{T_l \in \tau(D_j)} wtd(l,j)} \; if \; T_i \in \tau(D_j) \\ \delta_d \; otherwise \end{cases} \quad (21)$$

$$p(t_i/c_k) = \begin{cases} \frac{wtc(i,k)}{\sum_{T_l \in \tau(C_k)} wtc(l,j)} \; if \; T_i \in \tau(C_k) \\ \delta_c \; otherwise \end{cases} \quad (22)$$

where $wtd(i,j)$ and $wtc(i,k)$ are respectively the weights of term T_i in document D_j and user's interest C_k, $\tau(D_j)$and $\tau(C_k)$ are respectively the index terms of document D_j and the user's interest C_k, δ_d and δ_c constant values ($0 \leq \delta_d, \delta_c \leq 1$) expressing the default probability value.

4.4 Preliminary experimental validation

In order to evaluate the effectiveness of our model, we need the following three datasets: (1) a document collection (2) query topics and relevant judgments and (3) user's interests. We used a *TREC* data set from disk 1 and disk 2 of the ad hoc task containing 741670 documents issued from journals like *Associate Press (AP)* and *Wall Street Journal (WJS)* which provides the requirements (1) and (2). We particularly tested the queries among $q_{51} - q_{100}$ because they are enhanced by the domain meta data that gives the query domain of interest. The collection contains queries addressing 12 domains of interest. We choosed randomly four among them: *Environment, Law & Government, International Relations and Military*. We exploited the domain meta data in order to achieve the requirement (3) related to the user's interests. In order to map the query domains to realistic and dynamic user's interests, we applied the OKAPI algorithm that allows us to built a user's interest vector according to the formula:

$$wtc(i,k) = \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R-r+0.5)}$$

where R is the number of relevant documents to the queries belonging to C_k, r the number of relevant documents containing the term T_i, n the number of documents containing the term T_i, N is the total number of documents in the collection. For each specific domain tested addressed with n queries, we built n different user's interests. Furthermore, in order to validate our personalized retrieval model, we compared its performances to a naive bayesian model (Turtle and Croft, 1990). Table (1) presents the retrieval performance measures expressed using the well known *P@5*, *P@10* and *MAP* metrics on each of queries related to the four domains experimented. We can notice that our personalized information retrieval model is effective and achieve significant performance improvements over the traditional bayesian model for all the domains. The degree of improvement varies however from a query to another. This is probably depending, in one hand, on the relatedness between the simulated user's interests and the query domain (expressed in our model using a utility measure) and in the other hand, on the performance level of the baseline.

In the second series of experiments, we focus on the choice of a suitable aggregation operator. Tables (2) and (3) present the average results obtained for a pair of related domains (International Relations and Law&Gov) and quite unrelated ones (Environement and Military) using the sum and the max aggregation operators.

The experimental results presented above reveal that the sum operator is outperformed by the max operator in the case of both related and unrelated domains. Further interesting work will consist on ex-

	Baseline			Our model		
Environment	P@5	P@10	MAP	P@5	P@10	MAP
59	0,40	0,40	0,01	0,80	0,80	0,05
77	0,80	0,70	0,39	1,00	1,00	0,25
78	1,00	1,00	0,75	1,000	1,00	0,35
80	0,00	0,10	0,03	0,40	0,20	0,01
Intern. Rel	P@5	P@10	MAP	P@5	P@10	MAP
64	0,20	0,20	0,18	0,80	0,60	0,24
67	0,00	0,10	0,00	0,40	0,30	0,01
69	0,20	0,20	0,08	1,00	1,00	0,47
79	0,00	0,00	0,00	1,00	0,60	0,08
Law &Gov	P@5	P@10	MAP	P@5	P@10	MAP
70	0,60	0,60	0,42	1	1	0,65
76	0,60	0,70	0,08	0,6	0,3	0,09
85	0,60	0,80	0,21	0,60	0,70	0,16
87	0,20	0,20	0	1	0,6	0,05
Military	P@5	P@10	MAP	P@5	P@10	MAP
62	0,20	0,40	0,33	0,80	0,80	0,80
71	1,00	1,00	0,80	0,20	0,20	0,20
91	0,00	0,00	0,00	0,80	0,60	0,60
92	0,00	0,00	0,00	0,80	0,60	0,60

Table 1: Experimental results per domain

	$\sum$ **Operator**			*Max* **Operator**		
Domains	P@5	P@10	MAP	P@5	P@10	MAP
Environment	0,30	0,25	0,06	0,8	0,75	0,17
Military	0,28	0,40	0,04	0,43	0,50	0,10

Table 2: aggregation of unrelated domains of interest

ploring the common data distributions between relevant documents related to the queries issued from these specific domains. A good correlation would be an effective indicator of relatdeness that can be really exploited to tune the aggregation operator.

5 CONCLUSION

The paper investigates how to exploit theoretical foundations of graphical models to address the problem of contextual IR. In particular, we formalize the qualitative expression of a specific user's information need using CP-Nets. More precisely, the UCP-Nets predominance property allow us tu generate automatically accurate term queries according to the user's preferences. An other factor of user's context is considered in the retrieval model via influence diagrams. An inference process based on an extension of the bayesian joint law makes possible to personalize the relevance statement of documents.
What we learned from our investigation?

	$\sum$ **Operator**			*Max* **Operator**		
Domains	P@5	P@10	MAP	P@5	P@10	MAP
Intern. Rel	0,50	0,55	0,18	0,8	0,62	0,20
Law &Gov	0,6	0,5	0,18	0,8	0,65	0,23

Table 3: aggregation of related domains of interest

1. graphical models constitute opportunities to exploit for dealing with uncertainty embedded particulary within context features: preferences, interests, user's relevance statement etc,
2. context aware IR can be viewed as a constraint based inference process than can be involved naturally within a graphical model. The constraints represent the available conditions supporting useful information for a specific user,
3. the theoretical support of graphical models allows facilities for integrating various aspects of context in IR.

Future interesting work is to model contextual retrieval via a unified graphical formalism supporting user profiling in one hand (eliciting queries, building user's interests etc.) and contextual relevance measurement in another hand.

REFERENCES

Acid, S., Campos, L. D., Fernandez, J. M., and Huete, J. F. (1988). An information retrieval model based on simple bayesian networks. *International Journal of Intelligent Systems*, 18:251–265.

Bordogna, G., Carrara, C., and Pasi, G. (1991). Query term weights as constraints in fuzzy information retrieval. *Information Process Managment (IPM)*, 1:15–26.

Bordogna, G., Carrara, P., and G.Pasi (1986). Query term weights as constraints in fuzzy information retrieval. In *Information Processing and Management*, page 1526.

Bordogna, G. and Pasi, G. (1993). A fuzzy linguistic approach generalizing boolean information retrieval: a model and its evaluation. *Journal of Ameriacn Society Information Science*, 44:70–82.

Boubekeur, F., Boughanem, M., and L.Tamine-Lechani (2007). Semantic information retrieval based on cp-nets. In *Proceedings of IEEE International Conference on Fuzzy Systems*.

Boutilier, C., Bacchus, F., and Brafman, R. (2001). Ucp-networks: A directed graphical representation of conditional utilities. In *Proceedings of UAI*, page 56 64.

Boutilier, C., Brafman, R., Hoos, H., and Poole, D. (1999). Reasoning with conditional ceteris paribus preference statements. In *Proceedings of UAI*, page 71 80.

Campos, M. L. D., Fernandez-Luna, J., Huete, M., and Juan, F. (2004). Using context information in structured document retrieval: an approach based on influence diagrams. *Information Processing and Management (IPM)*, 40(5):829–847.

Crestani, F. and Ruthven, I. (2007). Introduction to special issue on contextual information retrieval systems. *Information retrieval (IR)*, 10:829–847.

Efthimiadis, E. (1996). Query expansion. *Annual review of Information Science and Technology*, 31:121–187.

Gowan, J. M. (2003). *A multiple model approach to personalised information access*. Master Thesis in computer science, Faculty of science, University College Dublin.

Haveliwala, T. (2002). Topic-sensitive page rank. In *International ACM World Wide Web conference*, pages 727–736.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive ir theory. *Journal of documentation*, 52(1):3–50.

Jansen, B., Spink, A., and T.Saracevic (2000). Real life, real users and real needs: a study and analysis for user queries on the web. *Information Processing and Management (IPM)*, 36:207–227.

Jensen, F. (2001). *Bayesian networks, decision graphs*. Springer.

Kantor, P. (1981). The logic of weighted queries. *IEEE Transactions on systems Man and Cybernetics*, 11:816–821.

Kraft, D., Bordogna, G., and Pasi, G. (1994). An extended fuzzy linguistic approach to generalize boolean information retrieval. *Information Science*, 2:119–134.

Liu, F. and Yu, C. (2004). Personalized web search for improving retrieval effectiveness. *IEEE Transactions on knowledge and data engineering*, 16(1):28–40.

Nunberg, G. (2003). As google goes, so goes the nation. *New York Times*.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible Inference*. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA. isbn: 0-934613-73-7.

Qiu, F. and Cho, J. (2006). Automatic identification of user interest for personalized search. In *International ACM World Wide Web conference*, pages 727–736.

Ruthven, I. and Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. *Knowledge engeneering review*, 18(2):95–145.

Shachter, R. (1988). Probabilistic inference and influence diagrams. *Operating Research*, 36(4):589–604.

Sieg, A., Mobasher, B., and Burke, R. (2004). Users information context: Integrating user profiles and concept hierarchies. In *Proceedings of the 2004 Meeting of the International Federation of Classification Societies*, number 1, pages 28–40.

Speretta, M. and Gauch, S. (2005). Personalized search based on user search histories. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 622–628.

Turtle, H. and Croft, W. (1990). Inference networks for document retrieval. In *Proceedings of the* 13^{th} *International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1–24.