

Influence diagrams for contextual information retrieval

Lynda Tamine-Lechani, Mohand Boughanem

Université de Paul Sabatier, Laboratoire IRIT,
118 Route de Narbonne, 31400 Toulouse CEDEX 06, France

Abstract. The purpose of contextual information retrieval is to make some exploration towards designing user specific search engines that are able to adapt the retrieval model to the variety of differences on user's contexts. In this paper we propose an influence diagram based retrieval model which is able to incorporate contexts, viewed as user's long-term interests into the retrieval process.

Keywords: personalized information access, influence diagrams, user context

1 Introduction

A key challenge in information retrieval is the use of contextual evidence within the ad-hoc retrieval. Several approaches explored in contextual retrieval area, techniques for building the user's profile [3, 6, 2] and using it in the retrieval process [2, 7, 5, 8]. Most of these approaches employ implicit user feedback to model the related long-term interests as contexts represented by word vectors [2], classes of concepts [3] or a hierarchy of concepts [6]. Since the contexts are modeled, they are exploited in order to refine the query [2, 8], re-rank the documents [7] or model the whole retrieval process [5, 8]. The latest goal is precisely our own one. Our contribution is particularly based on the belief that contextual retrieval is a decision-making problem. For this reason we propose to apply influence diagrams which are an extension of Bayesian networks to such problems, in order to solve the hard problem of user's relevance estimation.

2 The influence diagram based model

An influence diagram [4] is a graphic model used to represent and resolve a decision-making problem. Our interest in influence diagrams is namely to model the decision problem of document relevancy by taking into account the influence of both user's long-term interests and the query submitted.

2.1 Diagram topology

Figure 1 illustrates the qualitative component of our influence diagram based retrieval model. The set of nodes V is composed of four different types of nodes $V = D \cup T \cup P \cup C$. The set $D = \{D_1, D_2, \dots, D_n\}$ represent the set of documents in the collection, $T = \{T_1, T_2, \dots, T_m\}$ represent the set of terms used to index these documents, $P = \{P_1, P_2, \dots, P_n\}$ represent the decisions to state that these documents are relevant and $C = \{C_1, C_2, \dots, C_u\}$ represent the set of a specific user's contexts expressing his long-term interests. These different types of nodes are described below:

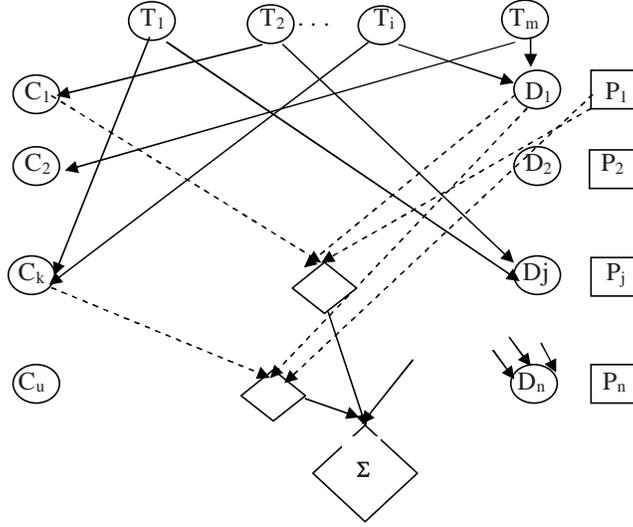


Fig. 1. Influence diagram-based retrieval model

- *Chance nodes.* There are three types of chance nodes: documents, terms and contexts. Each document node D , represents a binary random variable taking values in the set $\{d, \bar{d}\}$, where d represents 'the document d is relevant for a given query', and \bar{d} represents 'the document d is not relevant for a given query'. Each term node T represents a binary random variable taking values in the set $\{t, \bar{t}\}$, where t represents 'the term t is representative for a given query' and \bar{t} represents 'the term t is not representative for a given query'. Each context node C represents a binary random variable taking values in the set $\{c, \bar{c}\}$, where c represents 'the context c is relevant for a given query' and \bar{c} represents 'the context c is not relevant for a given query'.
- *Utility nodes.* There is an utility node corresponding to each decision node.

- *Decision nodes.* Each decision node P represents a binary random variable p , taking values in the set $\{p, \bar{p}\}$. These values corresponds to each pair $\{d, \bar{d}\}$ corresponding to a document node.

Influence arcs join each node term $T_i \in \tau(D_j)$ and each document node $(D_j) \in D$. Similarly there are influence arcs joining each node term $T_i \in \tau(C_k)$ and each context node $C_k \in C$. We note $Pa(\cdot)$ the parent sets for each node in the network: $\forall T_i \in T, Pa(T_i) = \emptyset, \forall D_j \in D, Pa(D_j) = \tau(D_j), \forall C_k \in C, Pa(C_k) = \tau(C_k)$

The informative arcs point to utility nodes for which ordered numerical values are assigned.

2.2 Probability distributions

The estimation of the probability distributions stored in chance and decision nodes is carried out in the following ways:

- *Term node:* $p(t_i/pa(t_i)) = p(t_i)$ as $Pa(T_i) = \emptyset$. We assume that $p(t_i) = \alpha, p(\bar{t}_i) = (1 - \alpha) \forall t_i \in T$ ($0 \leq \alpha \leq 1$)
- *Document node:* $p(d_j/pa(d_j)) = \sum_{t_i \in rel(pa(d_j))} Wtd(i, j)$ where $rel(pa(d_j)) = \{T_i \in Pa(D_j)/t_i \in pa(d_j)\}$, $Wtd(i, j) = \frac{wtd_{ij}}{\sum_{t_i \in \tau(D_j)} w_{lj}}$ wtd_{ij} is the weight of the term T_i in the document D_j
- *Context node:* $p(c_k/pa(c_k)) = \sum_{t_i \in rel(pa(c_k))} Wtc(i, k)$ where $rel(pa(c_k)) = \{T_i \in Pa(C_k)/t_i \in pa(c_k)\}$, $Wtc(i, k) = \frac{wtc_{ik}}{\sum_{t_i \in \tau(C_k)} w_{lk}}$ wtc_{ik} is the weight of the term T_i in the context C_k

3 Query evaluation

The query evaluation consists in the propagation of new evidence through the diagram, like in Bayesian networks [1], in order to maximize a re-ranking utility measure. More precisely, given a query Q represented by a set of positive terms ($\tau(Q) = \{T_1, T_2, \dots, T_r\}$), the retrieval process starts placing the evidence in the term nodes (marginally independent): $p(t_i/Q) = 1$ if $T_i \in \tau(Q)$ and $p(t_i/Q) = \alpha$ if $T_i \notin \tau(Q)$. Then, the inference process is run by maximizing the re-ranking utility measure $\frac{EU(p/Q)}{EU(\bar{p}/Q)}$ computed as follows: (we assume that documents are independent given the query and context)

$$EU(p/Q) = \sum_{c_k \in \{c, \bar{c}\}, d_j \in \{d, \bar{d}\}} u(p/c_k, d_j) p(c_k/Q) p(d_j/Q) \quad (1)$$

$$EU(\bar{p}/Q) = \sum_{c_k \in \{c, \bar{c}\}, d_j \in \{d, \bar{d}\}} u(\bar{p}/c_k, d_j) p(c_k/Q) p(d_j/Q) \quad (2)$$

When using the probability functions used respectively for document and context nodes, we compute respectively $p(d_j/Q)$ and $p(c_k/Q)$ as follows:

$$p(d_j/Q) = \alpha + (1 - \alpha) \sum_{T_i \in (\tau(D_j) \cap \tau(Q))} Wtd(i, j) \quad (3)$$

$$p(c_k/Q) = \alpha + (1 - \alpha) \sum_{T_i \in (\tau(C_k) \cap \tau(Q))} Wtc(i, k) \quad (4)$$

4 Conclusion

We proposed in this poster, an influence diagram based model for contextual information retrieval. This model allows to make inferences about the user's search intention and to take ideal actions based on probability query term distributions over the document collection and the user's contexts. We are currently experimenting with the graph representation, identification of user's contexts and parameters to be used for query evaluation.

Acknowledgments

This research was partially supported by the French Ministry of Research and New Technologies under the ACI program devoted to Data Masses (ACI-MD), project MD-33.

References

1. S. Acid, L.M. De Campos, J.M, Fernandez-Luna J.F. An information retrieval model based on simple Bayesian networks. *International Journal of Intelligent Systems*, 18, pages, 251-265, 2003
2. J. Budzik, K.J Hammond, Users interactions with everyday applications as context for just-in-time information access. In *Proceedings of the 5th international conference on intelligent user interfaces*, pages 44-51, 2000
3. J.P Mc Gowan , A multiple model approach to personalised information access. Master Thesis in computer science, Faculty of science, University College Dublin, February 2003
4. F. Jensen : *Bayesian networks and decision graphs*. Berlin: Springer Verlag, 2001
5. Lin C., Xue G.R., Zeng H.J., Yu Y., Using probabilistic latent semantic analysis for personalized Web search. In *Proceedings of the APWeb Conference*, Springer Verlag Eds, pages 707-711, 2005
6. F. Liu, C. Yu, Personalized Web search for improving retrieval effectiveness, *IEEE Transactions on knowledge and data engineering*, 16(1), pages 28-40, 2004
7. M. Speretta, S. Gauch, Personalizing search based on user search histories, In *30th Conference on information retrieval and management, CIKM*, 2004
8. J.R Wen, N. Lao, W.Y. Ma, Probabilistic model for contextual retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and development in information retrieval*, pages 57-63, Sheffield, 2004