Exploiting Multi-Evidence from Multiple User's Interests to Personalizing Information Retrieval

Lynda Tamine-Lechani
lechani@irit.fr

Mohand Boughanem
bougha@irit.fr

Nesrine Zemirli
nzemirli@irit.fr
Institut de Recherche en Informatique de Toulouse
IRIT 118 Route de Narbonne Toulouse CEDEX 09

Abstract

The goal of personalization in information retrieval is to tailor the search engine results to the specific goals, preferences and general interests of the users. We propose a novel model that considers the user's interests as sources of evidence in order to tune the accuracy of documents returned in response to the user query. The model's fundation comes from influence diagrams which are extension of Bayesian graphs, dedicated to decision-making problems. Hence, query evaluation is carried out as an inference process that aims to computing an aggregated utility of a document by considering its relevance to the query but also the corresponding utility with regard to the user's topics of interest. Experimental results using enhanced TREC collections indicate that our personalized retrieval model is effective.

1 Introduction

Numerous theoretical and empirical works [3, 4] suggest 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 these findings, information personalization is an active research area that consists mainly in enhancing an information retrieval process with the user's context with the aim of providing accurate results. There are two kinds of contexts; the short-term context, which is the surrounding information which emerges from the current user's information need in a single session. The other kind of context is long-term context which refers generally to the user's interests that have been inferred across the past user sessions.

Recent studies suggest furthermore that user's searches may have multiple goals or topics of interest and occur within the broader context of their information-seeking behaviors [12]. Research studies also indicate that IR researches often include such multiple topics, during a single session or multitasking search [11]. They found that multitasking information seeking is a common behaviour as many IR system users conduct information seeking on related or unrelated topics. The objective of the contribution reported in this paper is to highlight the prevalence and the usefuleness of the evidence extracted from multiple user's interests in order to tune the accuracy of the results presented in response to the query. The main research questions addressed are: (1) how to model a personalized retrieval task within a broad variety of topics? (2) how to pool such topics in order to measure the relevance of a document in response to a specific user query?

The inspiration and foundation of the present work come from Bayesian theory. More precisely, we claim that personalized retrieval is a decision-making problem as we shall make decisions about what information is relevant using both the probability of relevance of a document, and also the usefulness of the document to be presented according to each user interest (topic of interest). For this reason, the theoretical retrieval model we attempt to specify is based on an influence diagram (ID) [8] which is an extension of Bayesian networks to decision-making problems.

The remainder of this paper is structured as follows. After reviewing related work in section 2, we formally present in section 3, the main problem addressed in the paper. In section 4 we describe the qualitative and the quantitative components of our influence-diagram based personalized retrieval model and then detail the query evaluation process. In section 5, we present first, an evaluation framework based on a user simulation process that aims to enhancing TREC data collections with user's multi-topic ressources. It is followed by experimental results that show the effectiveness of our model. Section 5 concludes the paper.

2 Related work

The key idea behind personalizing IR is to customize search based on specific user's interests. Therefore, as a personalized search engine is intended for a wide variety of users with different goals, preferences and interests, it has to learn the user model first and then to exploit it in the retrieval task in order to provide more accurate results. Numerous works in IR address the critical question of user modelling, particularly using implicit feedback techniques [5, 14, 10]. This paper focuses on the second critical question related to the ranking model that considers the learned interests of the user when computing the relevance of a document. In the following, we particularly report work on information personalization within multiple user's interests. In [10], 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 [9] 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 [1] 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 [5] user profiles are used to represent the 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 to the query. Recently, extensions of the Page Rank algorithm [7, 2] 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. Our approach for personalizing document ranking is different from those previously cited. Our approach attempts to exploit the user profile as an explicit part of the formal retrieval model and not as a source of evidence to re-rank the documents or adapt a basic relevance estimation function. For this aim, we explore the use of ID which are Bayesian probabilistic tools dedicated to decision-making problems. Our goal is to show how user's interests can be explicitly integrated into an unified model and pooled in order to evaluate the global utility of the decisions related to the relevance of documents within a query. Our contribution in this paper focuses on the personalized retrieval model. We assume that the user's profile, representing a set of long-term interests, has already been built using an appropriate methodology based on the previous retrieval sessions [13]. Each user's interest is represented using a term-weighted vector where each term represents a dominant keyword that emerged from the user's search history. This offers flexibility to plug our personalized retrieval model to various user models.

3 The problem

Intuitively, the problem of personalizing IR may be expressed basically as follows:
Given a query **Q**, the IR system's problem is to identify those documents **D** that are relevant to the information need of the user **U**. From the probabilistic point of view, the IR system's 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,

$$p(d/q, u) = \frac{p(q/d, u)p(d/u)}{p(q/u)} \tag{1}$$

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:

$$RSV_U(\mathbf{Q}, \mathbf{D}) = p(q/d, u)p(d/u) \tag{2}$$

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, in contrast, is query independent, highlighting the usefulness of the document to the user. This may express the suitability of the document to the whole interests 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:

$$RSV_U(q, d) = p(q/d, c_1, c_2, ..., c_n)p(d/c_1, c_2, ..., c_n) \tag{3}$$

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,

[1] Relevance Status Value

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_i expresses the utilty of a fixed document **D** to the user interest C_i

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:

$$RSV_U(Q, D) = \Psi_{j=1..u}(\mu(d, c_j)p(q/d, c_1, c_2, ..., c_u)) \quad (4)$$

The following section gives formal details of our personalized IR model based on the above specification.

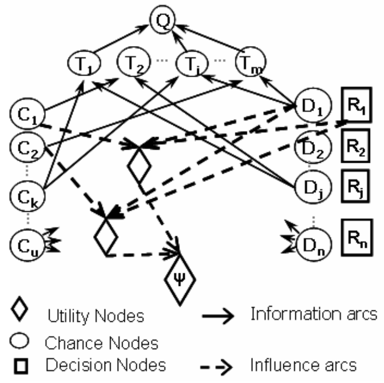
4 The personalized retrieval model

4.1 The diagram qualitative component

The proposed network architecture appears on Figure (1). From a qualitative point of view, nodes in the graphical component represent different kinds of information expressed by three types of nodes : the *chance nodes*, the *utility nodes* and the *decision nodes*.

The ***chance nodes*** represent the whole of binary random variables used in our model expressed by the set $V = Q \cup D \cup C \cup T$ where the set $D = \{D_1, D_2, \ldots, D_n\}$ represent the set of documents in the collection, $C = \{C_1, C_2, \ldots, C_u\}$ represent the set of specific user's long-term interests, $T = \{T_1, T_2, \ldots, T_m\}$ represent the set of terms used to index these documents and user's interests and Q represents the user's query. Each chance node X in the set V takes values in a binary domain $\{x, \overline{x}\}$ which indicates the positive and the negative instantiation respectively. More precisely, for each document node in D, d_i traduces, as in the Turtle model [15], that the document D_i has been observed and so introduces evidence in the diagram, all the remaining documents nodes are set to $\overline{d_i}$ alternatively to compute the posterior relevance. Similarly, c_k and $\overline{c_k}$ express respectively that the context C_k is observed or not observed. For each term node, t_i expresses that the term T_i is relevant for a given query, and t_i that the term T_i is not relevant for a given query. The relevance of a term means its closeness to the semantic content of a document. In the domain value of the query $\{q, \overline{q}\}$, q means that the query is satisfied and $\overline{q}$ that it is not satisfied. As only the positive query instantiation is of interest, we consider $Q = q$ only.

Figure 1. The influence diagram topology



A ***utility node*** is attached to each decision node related to present the document by taking into account the user's interest. So, we associate for each document D_j and each user interest C_k one utility node. All the values given by the pair (D_j, C_k) are used by a specific utility node in order to compute the global utility attached to the decision to return this document D_j according to the whole user's interests.

A ***decision node*** R_j is associated to each document D_j in the collection. It represents the decision to state that the document D_j is relevant. The node R_j represents a binary random variable taking values in a domain $dom(R_j) = \{r_j, \overline{r_j}\}$.

Informative arcs join each term node T_i to each document node $D_j \in D$ and each user interest node $C_k \in C$, whenever T_i belongs to D_j and C_k. This simply reflects the dependence between the relevance values of both document, user's interests and terms used to index them. There are also arcs which connect each term node with the query node. We note $Pa(.)$ the parent sets for each node in the network: $\forall T_i \in T, Pa(T_i) = \tau(D_j) \cup \tau(C_k)$, $\forall D_j \in D, Pa(D_j) = \oslash$, $\forall C_k \in C, Pa(C_k) = \oslash$, where $\tau(D_j)$ and $\tau(C_k)$ represent the index terms.

Influence arcs specify the influence degree of the variables associated with a decision. More precisely, they join in our model, the decision nodes, user's interest nodes and document nodes by using an aggregation operator specified below.

4.2 The diagram quantitative component

4.2.1 Probability distributions

We compute the posterior probability or belief associated with each node in the network as follows.

- ***Query node***. 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 [6] such as:

$$p(Q/pa(Q)) = \begin{cases} 0\, if\ \ (Pa(Q) \cap R(Pa(Q)) = \oslash \\ \frac{1 - \prod_{T_i\ \in R(Pa(Q))} nidf(T_i)}{1 - \prod_{T_i\ \in Pa(Q)} nidf(T_i)}\ otherwise \end{cases} \quad (5)$$

  where $nidf(T_i)$ is the normalised idf of the term T_i.

- ***Term node***. 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:

$$p(t_i/d_j) = \begin{cases} \frac{wtd(i,j)}{\sum_{T_l \in \tau(D_j)} wtd(l,j)}\ if\ T_i \in \tau(D_j) \\ \delta_d\ otherwise \end{cases} \quad (6)$$

$$p(t_i/c_k) = \begin{cases} \frac{wtc(i,k)}{\sum_{T_l \in \tau(C_k)} wtc(l,j)}\ if\ T_i \in \tau(C_k) \\ \delta_c\ otherwise \end{cases} \quad (7)$$

  where $wtd(i,j)$ and $wtc(i,k)$ are respectively the weights of the term T_i in the document D_j and 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.

4.2.2 The Utility value

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 $\mu(r_j/c_k)$:

$$\mu(r_j/c_k) = \frac{1 + \sum_{T_i \in D_j} nidf(T_i)}{1 + \sum_{T_i \in D_j - C_k} nidf(T_i)} \quad (8)$$

$\mu(\overline{r_j}/c_k)$ is computed as: $\mu(\overline{r_j}/c_k) = \frac{1}{\mu(r_j/c_k)}$

4.3 Relevance scoring

Following the decision theoritical support of our approach, 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:

$$RSV_U : \begin{cases} R \longrightarrow\ R \\ RSV_U(Q, D) \mapsto \frac{EU(r/D)}{EU(\overline{r}/D)} \end{cases} \quad (9)$$

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

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

$$EU(r/D) = \Psi_{k=1..u}\left[\mu(r/c_k) * p(q/d, c_k)\right] \quad (10)$$

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:

$$EU(r/D) = \Psi_{k=1..u}$$

$$\left[\mu(r/c_k) * \sum_{\theta^s \in \theta} (p(q/\theta^s) * \prod_{T_i \in Q \cap (D_j \cup c_k)} p(\theta_i^s/d_j) * p(\theta_i^s/c_k))\right] \quad (11)$$

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

$$EU(r/D) = \Psi_{k=1..u}$$

$$\left[\mu(\overline{r}/c_k) * \sum_{\theta^s \in \theta} p(q/\theta^s) \prod_{T_i \in Q \cap (D_j \cup C_k)} p(\theta_i^s/d_j) * p(\theta_i^s/c_k)\right] \quad (12)$$

where 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 for the term T_i in $pa(Q)$ and Ψ an appropriate aggregation operator specified below.

4.4 Relevance aggregation

The problem addressed at this level concerns the joint utility estimation of a document according to the whole user's interests. Assuming that the query may cover one major topic or various sub-topics, we shall specify the aggregation operator Ψ on the basis of the relatedness of the user's interests.

- ***Hypothesis 1***: *User's interests are unrelated*
  In this case, the rank of a document should be high

according to the suitable user's interest and low according to the others. A Possible formulation of the aggregation operator is:

$$\Psi(z_1, ..., z_u) = Max(z_1, ..., z_u) \tag{13}$$

where $z_k = \mu(r/c_k) * p(q/d, c_k)$ according to the formula (11), u is the number of user's interests

- ***Hypothesis 2***: *User's interests are related*
  The relatedness of user interests implies a possible reinforcement of the information relevance according to the query. This could express in some cases as the presence of subtopics of a general topic as in hierarchical representations. A Possible formulation of the aggregation operator is:

$$\Psi(z_1 \ldots z_n) = \sum (z_1 \ldots z_u) \tag{14}$$

5 Evaluation

We present below our framework evaluation and then describe our experiments and discuss the results obtained.

5.1 Framework evaluation

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 (WSJ)* 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 BM25 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 [15].

5.2 Experimental results

We attempt to achieve through our experiments two main objectives: (1) evaluating the effectiveness of our model over the various simulated user's interests embedded within different domains of interest identified in the collection (2) tunning the aggregation operator according to the relatedness of the user's interests.

Table (1) presents the retrieval performance measures expressed using the well known $P@5$, $P@10$ and MAP metrics on each of the four domains experimented. We can notice that our personalized IR 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 (Environment and Military) using the sum and the max aggregation operators.

	Baseline			**Our model**		
Environement	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

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. This finding

	$\sum$ **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

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

Table 3. Aggregation of related domains of interest

doesn't match our intuition specified earlier about related domains of interest. This could be explained by a false hypothesis on the degree of effective relatedness between the two domains via the document collection. Further interesting work will consist on exploring 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 relatedness that can be really exploited to tune the aggregation operator.

6 Conclusion and further work

In this article, we have presented a new personalized IR model based on ID. The Bayesian theoritical support offers a solid fundation for representation of uncertainty about various kinds of information (documents, terms, query, user's interests) and dealing accurately, when seeking information, with preferences embedded in the user's interests. This model has been endowed with an inference propagation process that allows to perform a personalized query evaluation. The experimental evaluation results using an enhanced TREC data test collection, show that our model is effective. As future works, we plan to explore the analysis of the document collection in order to identifying other user's interests parameters to be used for query evaluation. Furthermore, we plan to study new probability estimations and other appropriate aggregation operators in order to combine more accurately the evidence issued from the broad variety of user's interests.

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

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

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

[4] P. Ingwersen and K. Jrvelin. *The Turn: Integration of information seeking and information retrieval in context*. Springer, 2005.

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

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

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

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

[9] A. Sieg, B. Mobasher, and R. Burke. 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, 2004.

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

[11] A. Spink. Multitasking information and behavior and information task switching: an exploratory study. *Journal of documentation*, 60(4):336–351, 2004.

[12] A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic. From e-sex to e-commerce: Web search changes. *IEEE Computer*, 35(3):133–135, 2004.

[13] L. Tamine, M. Boughanem, and W. N. Zemirli. Inferring the user's interests using the search history. In *Workshop on information retrieval, Learning, Knowledge and Adaptatbility (LWA), Hildesheim, Germany, Schaaf, Martin, Althoff, Klaus-Dieter*, pages 108–110, 2006.

[14] J. Teevan and S. Dumais. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28_{th} International SIGIR conference on research and development in information retrieval*, pages 449–456, 2005.

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