Integration of User Profile in Search Process according to the Bayesian Approach

https://doi.org/10.3991/ijes.v6i4.9716

Farida Achemoukh(✉), Rachid Ahmed-Ouamer
University of Mouloud Mammeri, Tizi-Ouzou, Algeria
achemoukh.farida@yahoo.fr

Abstract—Most information retrieval system (IRS) rely on the so called system-centered approach, behaves as a black box, which produces the same answer to the same query, independently on the user's specific information needs. Without considering the user, it is hard to know which sense refers to in a query. To satisfy user needs, personalization is an appropriate solution to improve the IRS usability. Modeling the user profile can be the first step towards personalization of information search. The user profile refers to his/her interests built across his/her interactions with the retrieval system. In this paper, we present a personalized information retrieval approach for building and exploiting the user profile in search process, based on Bayesian network. The theoretical framework provided by these networks allows better capturing the relationships between different information. Experiments carried out on TREC-1 ad hoc and TREC 2011 Track collections show that our approach achieves significant improvements over a personalized search approach described in the state of the art and also to a baseline search information process that do not consider the user profile.

Keywords—Information retrieval, Personalized Search, User modeling, User interest, User profile, Bayesian Network

1 Introduction

Personalization is an appropriate solution to find information adapted to the user's needs. Modeling the user can be the first step towards personalization of information retrieval. However, more personalized information retrieval approaches focused on the user profile construction in order to better identify his information needs. User profile can be deduced explicitly by asking users questions [Ma et al., 2007] or implicitly, by observing their activities [Gauch et al., 2003] [Speretta et al., 2005], [Liu et al., 2010], [Srinivasa et al., 2016], [Zhou et al., 2016]. It can be represented by a simple structure based on keywords [Shen et al., 2012], or by concept hierarchy issued from the user's documents of interests [Kim et al., 2003], [Speretta et al., 2005]. Or by using external domain ontology as an additional evidence to model the user profile as a set of concepts issued from predefined ontology [Gauch et al., 2003], [Daoud et al., 2009].

In this paper, we present a personalized information retrieval relies on building and exploiting a user profile in search process, based on Bayesian network. The notion of user profile considered here is modeled by his/ her interests represented as weighted vectors of terms and built across his/her interactions with the retrieval system. Therefore, for each submitted query, we consider the relevant documents selected by the user at his/her interaction with the retrieval system as the data source involved to build his/her interest.

User profile building is intended to improve the relevance of search results that match the user's information needs. Therefore, we propose a variant of bayesian network approach for search personalization performed by integrating the user profile in the retrieval process. In particular, we extend the Bayesian belief network model proposed in [Ribeiro-Neto et al., 1996] to provide a structure for representing a user's interaction and we define the matching measure that integrates the user profile in the retrieval process by interpreting the query-document-user profile relevance as a belief in a document and in a user profile with respect to a query.

Unlike previously cited work, our approach could be distinguished by several features. First, the user profile is modeled by his/her general interests represented as weighted vectors of terms. We consider the relevant documents selected by the user at his/her interactions with the retrieval system as the data source involved to build his/her interest. To estimate the relevance of document we use a bayesian approach for the matching measure by integrating the user profile as a separate component in the relevance retrieval function. The user profile is represented by a list of concepts issued from an external data source that is domain ontology in [Gauch et al., 2003] and then exploited in the ranking search results by combining the original score between the document and the query with the score between the document and the user profile [Daoud et al., 2009].

The rest of this paper is organized as follows: section 2 gives an overview of related work. In Section 3, we describe our personalized search approach for user profile exploitation based on Bayesian network. Section 4 presents the experimental evaluation and results. In the last section, we present our conclusion and the future work.

2 Related Work

Various approaches have been proposed to personalize the search results to a given user. Personalization consists of user modeling to build a powerful user profile and then its exploitation in the search process. User profile refers to the user interest built across his/her interaction with the retrieval system. In the next sections, we present related work to the personalization process, namely the user modeling and user profile exploitation in the retrieval process.

2.1 User modeling

A user model describes data that characterizes a user, such data related to user's preferences, goals and interests [Sieg et al., 2007], [Micarelli et al., 2007], [Shen e al., 2012], [Jenifer et al., 2015], [Srinivasa et al., 2016]. Most of user model approaches represent user profile as one or more vectors of terms [Gowan 2003], [Shen et al., 2005], [Tan et al., 2006]. Others organize user profile as hierarchical concepts structure representing the interest's domains [Gauch et al., 2003], [Kim et al., 2003], [Speretta et al., 2005] or with a structured model of predefined dimensions (personal data, interests, preferences etc). Works presented in [Micarilli et al., 2007] describe the user profile with two dimensions represented by the interactions history with search system and the user information needs based on his/her interests. Other approaches use external domain ontology as an additional evidence to model user profile as a set of concepts issued from predefined ontology [Gauch et al., 2003], [Daoud et al., 2009].

The construction of the user profile consists of collecting information representing the user. It can be done in two ways; explicitly or implicitly [Micarelli et al., 2007], [Jenifer et al., 2015], [Zhou et al., 2016]. In the explicit approach, the user is asked to be proactive and to directly communicate to the system his/her data and preferences [Ma et al., 2007]. However, an explicit request of information to the user implies to burden the user, and to rely on the user's willingness to specify the required information. To overcome this problem, several techniques have been proposed in the literature to automatically capture the user interests by implicit feedback techniques; this is done by monitoring the user's actions in the user system interaction, and by inferring from them the user's preferences. The proposed techniques range from click through data analysis, query log analysis, desktop information analysis, document display time [Speretta et al., 2005], [Agichtein et al., 2006], [Srinivasa et al., 2016].

2.2 Personalization process

User profile can be exploited before search to reformulate the query or after a search by re-rank the initial results [Micarelli et al., 2007].Query reformulation consists of initial query expanding with the user profile terms [Koutrika et al., 2005],[Joachims et al., 2007] [Gan et al., 2008]. In [Qiu et al., 2006] user profile is incorporated in the query-document matching model. It consists of computing the document score by considering its relevance to the query and to the user profile. Most of personalization approaches are based on initial results re-ranking by combining either original rank or score between the document and the query with the rank or score between the document and the user profile [Gowan 2003], [Liu and al., 2010], [Teevan and al., 2011], [Cai et al., 2017].

3 User profile representation and exploitation approach

In personalized search, one of the main issues is how to infer user profile and how to exploit it in search process.

To address these issues, our general approach for search personalization relies on building and using this user profile in retrieval process. First, the user profile is modeled by his/her general interests learned across his/her interactions with the retrieval system including queries.

User interest is built from returned documents judged relevant by the user for a query. It is represented as a vector of weighted terms. The building user profile is used to improve relevant results that match the user information needs. We propose a variant of bayesian network approach for search personalization performed by integrating the user profile in retrieval process. In particular, we extend the Bayesian belief network model proposed in [Ribeiro-Neto et al., 1996] to provide a structure for representing a user interaction and interpreting the query-document-user profile relevance as a belief in a document and in a user profile with respect to a query.

We summarize below the terminology and notations used in our contribution, then we detail our approach.

3.1 Terminology and notations

User's Interaction: A user's interaction with the search system, noted in, includes a query submitted by the user, the returned documents and the subset documents judged relevant implicitly by the user.

User profile: User profile refers to the user interests learned across his/her interactions with the retrieval system. A user interest is issued from the relevant documents selected by the user at his/her interaction. It is also represented as a vector of weighted terms,

noted $c_k = \{(t_1,w_{1k})^{,}(t_2,w_{2k})^{,\cdots\cdot}(t_i,w_{ik})\}$,

where w_{ik} denotes the weight of term t_i in user interest c_k. The weighting term value w_{ik} will be detailed below.

3.2 Building a keyword user interest

Building the user interest starts by collecting a set of relevant documents Dr returned with respect to a query q related to a user's interaction. Each relevant document is represented as a vector of weighted terms, where the weight wij of term ti in document dj is computed using the TF-IDF weighting scheme:

$$w_{ij} = tf_{ij} \times \log \frac{N}{n_i} \tag{1}$$

Where tf_{ij} is the frequency of term t_i in document d_j, N is the total number of documents and n_i is the number of document that contain term t_i.

The user interest c_k is also represented as a weighted vector of the most relevant terms occurring in the relevant documents judged by the user. The weight w_{ik} of term t_i in user interest c_k is computed as follows:

$$w_{ik} = \frac{1}{|D_r|}\sum_{dj \in Dr} w_{ij}.\log\frac{(r+0.5)(N-R-n+r+0.5)}{(n-r+0.5)(R-r+0.5)} \tag{2}$$

Where, N and R the total number of documents and the number of relevant documents to the query belonging to user interest c_k, respectively. r is the number of relevant documents that contain term t_i, in the number of documents that contain term t_i.

3.3 Bayesian belief network for search personalization

To improve relevant results that match the user information needs, we present a personalized information retrieval approach integrating the user profile in the retrieval process. Let us consider a submitted query q related to the user's interaction. Let $D=\{d_1,\ldots d_j,\ldots d_n\}$ the set of documents in the collection, $C\_I= \{c_1,\ldots c_k,\ldots c_m\}$ the set of user interests, and $T= \{t_1...,t_i ,\ldots t_p\}$ the set of index terms used to index these documents and user interests . Furthermore, documents, user interest and query are modeled identically.

The relationship between user interests, documents and query can be modeled as a Bayesian belief network that provide an effective and flexible framework for modeling distinct sources of evidence in support of a ranking. We propose to extend the Bayesian belief network model proposed in [Ribeiro-Neto et al., 1996] by integrating the user profile to provide a structure for representing a user's interaction and interpreting the query-document-user profile relevance as a belief in a document and in a user profile with respect to a query.

Bayesian belief network is represented by a directed acyclic graph G (V, E), where nodes $V = T \cup D \cup C\_I \cup q$ correspond to the set of random variables and the set of arcs $A = V \times V$ represents conditional dependencies among them.

Figure (Fig.1) shows the topology of our belief network model for user's interaction where the terms nodes represent the network roots.

Each term in the index terms, $t_i \in T$, is modeled by a random variable $t_i \in \{0, 1\}$. The event of "observing term t_i" is noted $t_i = 1$ or shortly t_i. The complement event that "term t_i is not observed", is noted $t_i = 0$ or shortly $\overline{t_i}$.Let p be the number of index terms present in the set of terms T. It exists 2^P possible term configurations represented by the set θ. A term configuration may represent a query, a document, or a user interest. It is represented by a vector of random variable $\vec{t} = (t_1,t_2,\ldots,t_p)$ where each variable indicates if the corresponding term is observed .For example, an index of 2 terms t_1 and t_2 presents $2^2 = \ = 4$ possible term configurations represented by the set $\theta = \{(t_1, t_2), (t_1,\overline{t2}), (\overline{t1}, t_2), (\overline{t1},\overline{t2})\}$. The event of observing a particular configuration $\vec{t} = \{t_1,t_2,\ldots,t_p\}$ is noted $\vec{t}$. Each document $dj \in D$ is modeled by a random variable $d_j \in \{0, 1\}$ with two possible values 0 or 1. The event $d_j = 1$, simplified with $d_{j,}$ denotes that the document d_j is observed. The event $d_j = 1$, simplified with $d_{j,}$ denotes that document d_j is not observed. A document d_j is represented as a term configuration $d_j = (t_1,t_2,\ldots,t_p)$ with t_i is a random variable indicating if either term t_i is present in the document or not. Obviously, observing a document in a retrieval process means that this document is relevant to the query.

Each user interest $c_k \in C\_I$ is modeled by a random variable $c_k \in \{0, 1\}$. The event $c_k = 1$, simplified with c_k, denotes that the user interest c_k is observed. The complement event that "user interest c_k is not observed", is noted $c_k = 0$ or shortly $\overline{c_k}$. A user interest c_k is represented as a term configuration $c_k = (t_1,t_2,\ldots,t_p)$ with t_i is a random variable indicating if either term t_i is present in the user interest or not. Obviously, observing a user interest means that this user interest is related to the query q.

- A user query q is represented by a random variable $q \in \{0, 1\}$. The two events of observing the query ($q = 1$) or not observing the query ($q = 0$) are noted q and $\overline{q}$, respectively. In our case, we interest only to a positive instantiation of q. In the same way as documents and user interests, a query is represented as a term configuration $q = (t_1,t_2,\ldots,t_p)$ with t_i is a random variable indicating if either term t_i is present in the query or not.

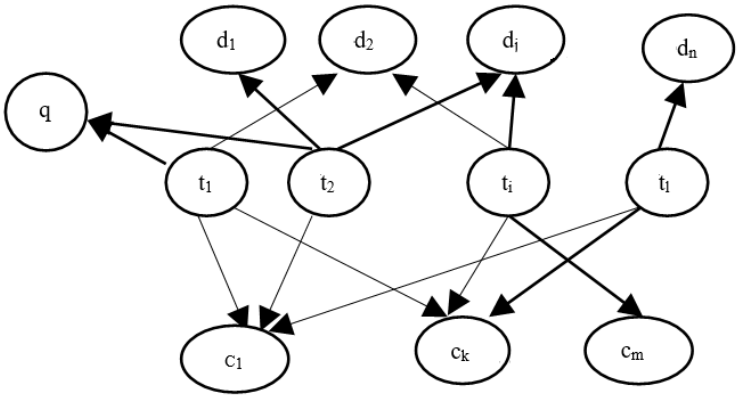


Fig. 1. Belief network model for given user's interaction.

To express conditional dependencies between random variables, three types of arcs are identified in the inference network model for search personalization: (1) Term to document: Arcs joining term node $t_i \in T$ to document node $d_j \in D$, (2) Term to user interest: Arcs joining term node $t_i \in T$ to user interest node $c_k \in C\_I$, (3) Term to query: Arcs joining term node $t_i \in T$ to user's query node q. Whenever term t_i belongs to document d_j, to user interest c_k and to a query q.

We detail in what follows the query evaluation process for the proposed belief network.

Evaluation process: Intuitively, we can express the personalization retrieval problem as follows: Given a query q, the search personalization consists in ranking documents according to the information need and user interest. In the network of (Fig.1), the ranking computation is based on interpreting the similarity between a document d_j, a user interest c_k and the query q as an intersection between dj ,c_k and q. To quantify the degree of intersection of the document d_j, the user interest c_k given the query q, we use the probability (dj, c_k and q). Thus, to compute a ranking, we use Bayes' law and the rule of total probabilities, as follows:

$$P(d_j,c_k|\ q)=\frac{P(d_j,c_k,q)}{P(q)} \tag{3}$$

As the denominator P (q) is a constant, we can use only the numerator in order to estimate the probability $P(d_j,c_k|q)$. Thus, the formula (5) is computed as:

$$P(d_j,c_k,q)=\sum_{\vec{t}\in\Theta} P(d_j,c_k,q|\ \vec{t})\times P(\vec{t}) \tag{4}$$

The probability P ($\vec{t}$) corresponds to the likelihood of observing term configuration $\vec{t}$. We assume that all the configurations are independent and have an equal probability to be observed. Therefore, the probability P (d_j ,c_k, q) is then approximated with:

$$P(d_j,c_k,q)=\sum_{\vec{t}\in\Theta} P(d_j,c_k,q|\ \vec{t}) \tag{5}$$

In the network of (Fig.1), instantiation of the root nodes separates the document nodes, the user interest's nodes and the query node, making them mutually independent, which allows writing:

$$P(d_j,c_k,q|\ \vec{t})=P(d_j|\ \vec{t})P(c_k|\ \vec{t})P(q|\ \vec{t}\ ) \tag{6}$$

By substituting $P(d_j,c_k,q|\ \vec{t})$ in formula 4, the probability P (d_j ,c_k, q) is estimated as:

$$P(d_j,c_k,q)=\sum_{\rightarrow} P(d_j|\ \vec{t}\ )P(c_k|\ \vec{t}\ )P(q|\ \vec{t}\ ) \tag{7}$$

Probabilistic inference in a Bayesian network is NP-Hard [Turtle et al., 1991]. To simplify the computation of probability P(d_j ,c_k, q), only instantiated terms in the query q are considered in term configuration and other terms are assumed not effective for document and user interest relevance.

We detail in what follows the computation of the conditional probabilities in formula (7).

- **Probability** $P(q|\vec{t})$

$P(q|\vec{t})$ determines the probability of generating the query q from term configuration $\vec{t}$. As proposed in [Turtle et al., 1991] the probability $P(q|\vec{t})$ is computed using the And–combination:

$$\begin{cases} P(q|\vec{t}) = 1 \text{ if } \vec{t} = q \\ 0 \text{ otherwise.¶} \end{cases} \qquad (8)$$

Probability $P(d_j|\vec{t})$

The probability $P(d_j|\vec{t})$ that document d_j is generated by term configuration $\vec{t}$ is estimated as the similarity between the document d_j and term configuration $\vec{t}$. As described in [Ribeiro et al 1996], a Bayesian network can be used to represent the rankings generated by any of the classic models. For instance, a Bayesian network can be used to compute the vector space model ranking. So, the similarity between the document d_j and term configuration $\vec{t}$ is interpreted as an intersection between document d_j and terms configuration $\vec{t}$ .Then P(dj| $\vec{t}$)is computed as follows:

$$P(d_j|\vec{t}) = \frac{\sum_{ti \in dj \wedge \vec{t}} w_{ij} \times w_{it}}{\sqrt{\sum_{ti \in dj} w^2{}_{ij}} \times \sqrt{\sum_{ti \in \vec{t}} w^2{}_{it}}} \qquad (9)$$

w_{ij}, and w_{it} denote respectively, the weight of term t_i in document d_j and in term configuration $\vec{t}$.

- **Probability** $P(c_k|\vec{t})$

Analogously, the similarity between user interest c_k and term configuration $\vec{t}$ is interpreted as the similarity between the user interest c_k and term configuration $\vec{t}$. Then the probability $P(c_k| \vec{t})$ is computed as follows:

$$P(c_k| \vec{t}) = \frac{\sum_{ti\in ck \wedge \vec{t}} w_{ik} \times w_{it}}{\sqrt{\sum_{ti\in ck} w^2{}_{ik}} \times \sqrt{\sum_{ti\in \vec{t}} w^2{}_{it}}} \tag{10}$$

w_{ik}, denotes the weight of term ti in user interest c_k .

Given this latter probabilities, the formula (7) becomes:

$$P(d_j, c_k, q) = \frac{1}{\sum_{ti\in q} w_{iq}{}^2} \times \frac{\sum_{ti\in q\wedge ck} w_{ik} \times w_{iq} \cdot \sum_{ti\in q\wedge dj} w_{id} \times w_{iq}}{\sqrt{\sum_{ti\in ck} w_{ik}{}^2} \times \sqrt{\sum_{ti\in dj} w_{ij}{}^2}} \tag{11}$$

$\frac{1}{\sum_{ti\in q} w_{iq}{}^2}$ is a constant for a given document and user interest. Ignoring it, formula (11) is rewritten as follows:

$$P(d_j, c_k, q) = \frac{\sum_{ti\in q\wedge ck} w_{ik} \times w_{iq} \times \sum_{ti\in q\wedge dj} w_{id} \times w_{iq}}{\sqrt{\sum_{ti\in ck} w_{ik}{}^2} \times \sqrt{\sum_{ti\in dj} w_{ij}{}^2}} \tag{12}$$

We can use an m×n matrix X, noted $X_{m,n}$ (m and n indicate respectively the number of user interests and documents) to represent resulting probabilities for each instantiations of document d_j and user interest c_k .It is defined as:

$$X_{m,n} = \begin{array}{c} c_1 \\ c_2 \\ c_k \\ c_m \end{array} \begin{pmatrix} P_{11} & P_{12} & ...P_{1j} & P_{1n} \\ P_{21} & P_{22} & ...P_{2j} & P_{2n} \\ P_{k1} & P_{k2} & ...P_{kj} & P_{kn} \\ P_{m1} & P_{m2} & ...P_{mj} & P_{mn} \end{pmatrix} \qquad d_1 \quad d_2 \quad d_j$$

P_{kj} denotes the probability $P(d_j,c_k |q)$ of relevance of user interest c_k and document d_j for a given query q.

We consider that the most likely user interest, noted $\widehat{c}$, given a query q, is selected as follows:

$$\widehat{c} = \operatorname{Arg\,max}_{\forall c_k \in C\_I} \frac{1}{n} \sum_{j=1}^{n} P_{kj} \tag{13}$$

Where $\frac{1}{n} \sum_{j=1}^{n} P_{kj}$ represents the arithmetic mean of the set { P_{k1},…, P_{kj},… P_{kn} } for given user interest c_k.

Therefore, for a given query q and user interest $\widehat{c}$, the probabilities P_{kj} presented in matrix $X_{m,n}$ are used to output a ranking list of documents

4 Experimental Evaluation

Our experiments have two main objectives. The first one is to compare the performance of our search personalization approach to the personalized approach proposed in [Daoud et al., 2009]. The second one is to evaluate the impact of user profile on the search results by comparing our personalized search approach to a baseline search information process that does not consider the user profile.

4.1 Evaluating the effectiveness of our personalized approach

Our purpose is to compare the performance of our search personalization approach to the approach proposed in [Daoud et al., 2009]. We recall that in our approach, the user profile is integrated in the retrieval process by interpreting the query-document-user profile relevance as a belief in a document and in a user profile with respect to a query. In [Daoud et al. 2009] approach, personalization consists of re-ranking the search results by combining query-document score and profile-document score.

The experiments have been handled in TREC data set from disk 1& 2 of the TREC ad hoc collections AP88 (Associated Press News, 1988) and WSJ90-92(Wall Street Journal, 1990-92). Collections contain 741670 documents, queries and relevant judgments. We particularly tested the queries among q51 − q100.

The choice of this test collection is due to the availability of a manually annotated domain for each query. This allows us, to simulate user interests changing over different domains of TREC. We used the same domain categorization than [Daoud et al 2009] approach. Table 1 shows six domains of TREC including 25 queries provided by TREC collection.

Table 1. TREC domains used for simulating user interests

Domains	Queries
Environment	59 77 78 83
International Politics	61 74 80 93 99
International Relations	64 67 69 79 100
Law and Government	70 76 85 87
Military	62 71 91 92
US Economics	57 72 84

Experimental design and results: The evaluation is based by simulating user interest's process based on N-fold cross validation strategy [Mitchell 1997] explained as follows:

For each TREC domain, divide the query set into N subsets. We repeat experiments N times, each time using a different subset as the test set and the remaining N−1 subsets as the training set.

For each query in the training set, the 1000 top documents are first returned by BM25 Model provided by terrier-3.5 platform then an automatic process uses the returned top documents which are listed in the assessment File (qrels) provided by TREC collections , to generate the user interest vector of weighted terms, using formula (2).

Then for each query in the test set, an automatic evaluation process (cf. section 3.3.1) generates the matrix given the relevance scores of documents and user interests.

Table 2 shows the percentage of improvement of our approach compared to [Daoud et al., 2009] approach computed at P5, P10 and MAP (Mean Average Precision) and averaged over the queries belonging to the same domain.

Table 2. Performance comparisons of the two personalized approach

Domains	Approach [Daoud et al]	Our Approach	Approach [Daoud et al 2009]	Our Approach	Approach [Daoud et al 2009]	Our Approach
	P5	***P5***	***P10***	***P10***	***MAP***	***MAP***
Environment	0,35	0,80	0,37	0,70	0,19	0,19
Inter. Politics	0,20	0,40	0,16	0,36	0,07	0,10
Inter. Relations	0,16	0,40	0,16	0,32	0,02	0,04
Law and Gov.	0,50	0,20	0,45	0,20	0,14	0,19
Military	0,35	0,45	0,32	0,40	0,07	0,12
US Economics	0,33	0,33	0,36	0,40	0,10	0,17

We notice that our approach gives higher performance than Daoud et al. (2009) approach for most of the queries in the all domains at P5, P10 and mean average precision (MAP). Based on the overall evaluation results, the conclusion we can made is that the integration of user profile in the matching model of retrieval process as computing the query-document-user profile relevance can better improve the search that the re-ranking of search results for a given query using the user profile as done in [Daoud et al 2009].

4.2 Evaluating the impact of user profile on the search results

The goal of this experiment is to evaluate the system performance by introducing the user profile in search process. We compare our approach to the baseline BM25 Model [Robertson et al 1998] provided by terrier-3.5 platform, using only the query ignoring any user profile.

We use a TREC 2011 Track collection. It consists of clueweb09_English1 collection of documents and includes relevance judgments, 61 main queries (topics). Each topic has a number of subtopics distributed as follows: 202 interactions queries and 75 currents queries. Interactions queries and current query are a sequence of reformulations of the main query. Table 3 shows the statistics data characteristics of the test collection.

Table 3. Statistics data of the test collection

Number of documents	**about 50.000.000 documents**
Number of Main queries (Topics)	61 queries
Number of Interaction queries	202 queries
Number of Currents queries	75 queries
Total queries	338 queries

Experimental design and results: The evaluation scenario we adopted is the following:

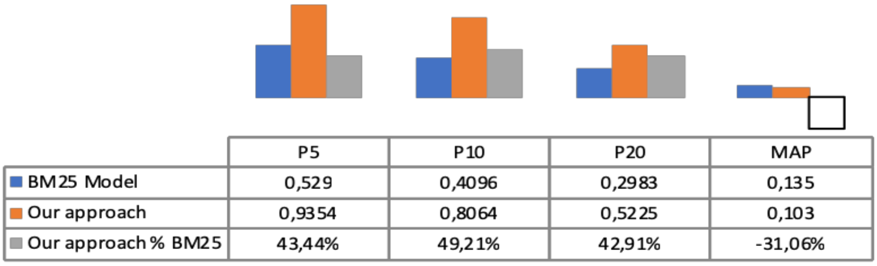
Inferring user profile: For each main query, the 1000 top documents are first returned by BM25 Model provided by terrier-3.5 platform then an automatic process uses the returned top documents which are listed in the assessment File (qrels) provided by TREC to generate the user interest vector of weighted terms, using formula (2). The vector represents the user interest

Personalization process: It consists of ranking the search results of a current query by using the user profile. We present in table 4 the precision improvement obtained by our approach introduced the user profile compared to the baseline BM25 Model [Robertson et al 1998] using only the query ignoring any user profile, at P5, P10, P20 and MAP averaged over the current queries

Table 4. Query by query comparison results between the baseline BM25 and our approach

N°query	Our Model			BM25 Model		
	P5	***P10***	***P20***	***P5***	***P10***	***P20***
1	0,6	0,4	0,2	0	0	0,05
2	1	0,8	0,4	0,8	0,4	0,3
3	0,8	0,7	0,35	0,6	0,5	0,35
4	0,8	0,4	0,25	0,2	0,3	0,25
5	1	1	1	0,8	0,7	0,65
6	0,8	0,6	0,3	0,4	0,2	0,15
7	1	1	1	0,8	0,9	0,6
8	1	0,7	0,35	0	0	0
9	1	1	1	1	0,8	0,7
12	0,8	0,6	0,3	0,8	0,4	0,25
13	1	0,8	0,4	0,6	0,6	0,35
17	1	1	0,65	1	0,9	0,55
19	1	0,6	0,4	0,2	0,1	0,25
21	1	0,7	0,35	0,6	0,3	0,2
23	0,8	0,7	0,4	0	0,1	0,15
29	1	0,8	0,4	0,4	0,3	0,2
30	0,8	0,4	0,25	0	0,2	0,15
33	0,8	0,8	0,4	0,2	0,1	0,15
36	1	1	0,8	0,6	0,5	0,3
39	1	1	0,9	1	0,7	0,6
40	1	1	0,55	0,8	0,5	0,3
43	1	1	0,55	0,6	0,3	0,25
44	1	1	0,55	0,8	0,4	0,2
47	1	0,8	0,4	0,8	0,7	0,35
49	1	0,7	0,4	0,6	0,3	0,15
51	1	1	0,75	0,8	0,7	0,45
52	1	0,8	0,4	0,2	0,3	0,15
54	1	1	0,5	0,2	0,3	0,2
57	1	1	0,65	0,2	0,2	0,2
58	1	1	0,85	0,8	0,6	0,45
60	0,8	0,7	0,5	0,6	0,4	0,35

We present in figure 2 the percentage of improvement of our model comparatively to baseline BM 25 Model [Robertson et al 1998] computed at P5, P10, P20 and MAP and averaged over the currents queries.



	P5	P10	P20	MAP
🟦 BM25 Model	0,529	0,4096	0,2983	0,135
🟧 Our approach	0,9354	0,8064	0,5225	0,103
◼ Our approach % BM25	43,44%	49,21%	42,91%	-31,06%

Fig. 2. Performance comparisons between BM25 and our model

We notice that our approach gives higher performance than BM25 Model at P5, P10 and P20. More particularly, our approach brings an improvement of 43.44% in

P5, 49.21% in P10 and 42.91% in P20, but there is a decrease in the mean average precision (MAP). However, these results are acceptable given the values of P5 P10 and P20.

5 Discussion

Our research work relies on how to build and how to exploit a user profile in the search process to produce better result rankings. Our intuition was based on the assumption that the search system provides the probability that a document is relevant to a user query, the goal is to estimate this probability by taking into account the user profile. For this purpose, our user profile is modeled by his/her general interest learned across his/her interaction with the retrieval system. Following this general view, our approach could be distinguished by several features in the personalized search community. The first one concerns the user profile construction and the second one concerns the user profile integration in the search process.

In our approach the user profile is modeled by his/her, interests represented as weighted vectors of terms. We consider the relevant documents selected by the user at his/her interactions with the retrieval system as the data source involved to build his/her interest. Then to estimate the relevance of document we use a bayesian approach for the matching measure by integrating the user profile as a separate component in the relevance retrieval function. While in [Gauch et al., 2003] and [Daoud et al., 2009], a user profile is represented by a list of concepts issued from an external data source that is domain ontology and original score between the document and the query with the score between the document and the user profile [Daoud et al., 2009]. The main assumption behind this representation is that we aim at representing the user profile as weighted vector of terms and incorporate it in the query-document matching model both represented as vector of terms using probabilistic approach.

6 Conclusion

In this paper, we have explored our approach for the user profile representation and its integration in personalized search. It consists of two basic steps: (1) inferring user interest at user's interaction (2) incorporating the user profile in the matching model of retrieval process. The user profile refers to the user interests built across his/her user's interactions. To integrate the user interest in the search process, we use a Bayesian networks to represent the user's interaction.

To evaluate the performance of our approach, we have conducted two experiments, based on using standard test collections in order to allow accurate comparative evaluation. First, to evaluate the effectiveness of our personalized search approach, we use TREC ad hoc collections. We compared our approach to Daoud et al., (2009) approach. In our approach we integrate the user profile in the matching model by interpreting the query-document-user profile relevance as a belief in a document and in a user profile with respect to a query. In Daoud et al., (2009) approach, personalization consists of re-ranking the search results of a given query using the user profile. Moreover, our experimental evaluation shows an improvement of personalized retrieval effectiveness compared to Daoud et al., (2009) approach. Second, to evaluate the user profile impact on the search results, we use clueweb09_English1 test collection and we compared our approach to baseline BM25 Model of the Terrier-3.5 platform, using only the query ignoring any user profile. The obtained results show that our approach gives higher performance than BM25 Model.

As future work, we plan to use user profile evolution in to improve the system performance for a recurring query and then undergo experiments in order to evaluate the impact of introducing the user profile in personalizing search results by comparing our approach to another personalized approach.

7 References

[1] Agichtein E., Brill E., Dumais S., Ragno R., «Learning user interaction models for predicting Web search preferences». In Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp: 3–10, 2006. https://doi.org/10.1145/1148170.1148175

[2] Cai F.,ShuaiqiangW., Maarten R., « Behavior-Based Personalization in Web Search». Journal Of The Association For Information Science And Technology, pp: 855–868, 2017. https://doi.org/10.1002/asi.23735

[3] Daoud, M. Tamine-Lechani L., Boughanem M., and Chebaro B., « A session based personalized search using an ontological user profile ». In: SAC '09: Proceedings of the ACM symposium on Applied Computing, New York, pp: 1732–1736, 2009. https://doi.org/10.1145/1529282.1529670

[4] Devarakonda R., Palanisamy G., Gil K I.S , « Framework for Building Collaborative Research Environment », International Journal of Recent Contributions from Engineering, Science & IT (iJES) , pp; 11-15 October 2014.

[5] Gan.I, Wang S., Wang M., Xie Z., Zhang L., and Shun Z., «Query Expansion based on Concept Clique for Markov Network Information Retrieval Model». Proceedings of the 5th International Conference on Fuzzy Systems and Knowledge Discovery, pp.29-33, 2008. https://doi.org/10.1109/FSKD.2008.648

[6] Gauch S., Gao J., Yuan W., Deng X. Li, and Nie. J.Y., «Smoothing clickthrough data for web search ranking». In SIGIR ACM, pp: 355–362, 2009.

[7] Chaffee J., and Pretschner A., « Ontology-based personalized search and browsing". Web Intelligence and Agent Systems, pp: 219– 234, 2003.

[8] Gowan J.P."A multiple model approach to personalised information access, Thesis in computer science, University College Duplin, February 2003.

[9] Jenifer K., Yogesh Prabhu M., Gunasekaran N., « A survey on web personalization web approaches for efficient information retrieval on user interests ». J. Recent Res. Eng. Technol, pp: 2349–2260, 2015.

[10] Joachims T., Granka L., Hembrooke H., Radlinski F., and Gay G., « Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search». ACM Transactions on Information systems, 25(2): 7, April 2007. https://doi.org/10.1145/1229179.1229181

[11] Kim .H., Chan P.K. «Learning implicit user interest hierarchy for context in personalization. » In: IUI '03: Proceedings of the 8th international conference on

Intelligent user interfaces, New York, pp: 101–108, 2003. https://doi.org/10.1145/604045.604064

[12] Koutrika G., and Ioannidis Y., « A unified user profile framework for query disambiguation and personalization ». In Proceedings of Workshop on New Technologies for Personalized Information Access, pp: 44-53, July 2005.

[13] Labriji A., Charkaoui S., Abdelbaki I., Namir A., Labriji E. «Similarity Measure of Graphs» International Journal of Recent Contributions from Engineering, Science & IT (iJES) , pp: 42-56. April 2017.

[14] Liu, J. and Belkin, N.J., « Personalizing information retrieval for multi-session tasks: the roles of task stage and task type». SIGIR, pp: 26–33, 2010. https://doi.org/10.1145/1835449.1835457

[15] Ma, Z., Pant, G. and Sheng, O., « Interest-based personalized search. » ACM TOIS, 25(1), 2007.

[16] Micarelli A., Gaspaetti F., Sciarrone F., and Gauch S., « Personalized Search on the World Wide Web». Lecture Notes in Computer Science, pp.195-230, 2007. https://doi.org/10.1007/978-3-540-72079-9_6

[17] Qiu F., and Cho J., «Automatic identification of user interest for personalized search». In Proc. of WWW, pp: 727-736, May 2006. https://doi.org/10.1145/1135777.1135883

[18] Ribeiro-Neto B., and Muntz R., «A belief network model for IR ». In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp: 253-260, 1996.

[19] Rosihan A.Y Dewanto H.,Nugraha A. K., Cucuk W. B., «Data Similarity Filtering of Wartegg Personality Test Result using Cosine-Similarity». International Journal of Recent Contributions from Engineering, Science & IT (iJES) , pp:19 -28. Octobre 2018.

[20] Shen, S., Hu, B., Chen W., and Yang Q., «Personalized click model through collaborative filtering». WSDM, pp: 323–332, 2012. https://doi.org/10.1145/2124295.2124336

[21] Sieg B., Mobasher B., and Burke. R., « Web search personalization with ontological user profiles». Proceedings of the 16th ACM conference on Conference on information and knowledge management, New York, NY, USA, ACM, pp: 525–534, 2007. https://doi.org/10.1145/1321440.1321515

[22] Speretta M., and Gauch S., «Personalized search based on user search histories ». In Proc. of Int Conf on Web Intelligence, pp: 622-628, 2005. https://doi.org/10.1109/WI.2005.114

[23] Srinivasa R.K., Renuka Kalyani A., Krishnamurthy M., «A New Approach of Inferring User Interest for User Search Goals with Web Log Sessions». Journal of Web Development and Web Designing, pp: 1-15, 2016.

[24] Tan B., Shen X., and Zhai Ch., «Mining long-term search history to improve search accuracy ». KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA. ACM, pp: 718–723, 2006. https://doi.org/10.1145/1150402.1150493

[25] Teevan J., Liebling D., and Geetha G.R., « Understanding and prediction personal navigation ». Proc .WSDM, pp: 85-94, 2011.

[26] Zhou D., Lawless S., Xuan W., Zhao W., Liu J., «A study of user profile representation for personalized cross-language information retrieval » , pp:448-477 , 2016.

8 Authors

Achemoukh Farida is a Ph.D. student at the University of Mouloud Mammeri, Tizi-Ouzou, Algeria. She received a Magister in Computer Science from the

University of Tizi-Ouzou, Algeria, in 2006. Her engineer degree in computer science was obtained from the same University, in 2002. His research interests include information retrieval, statistical models, and personalized information retrieval.

Rachid Ahmed-Ouamer is a Professor at the University of Tizi-Ouzou, Algeria, where he is a team leader at the Research Laboratory of Computer Science LARI. He is the Director of LARI since 2003. His current research interests include information retrieval models, social IR, ontology engineering, semantic web services, web-based applications, and information systems interoperability. He has served as a program committee member of the annual French conference in information retrieval CORIA since 2007. He is a member of the editorial board of the French reputable journal Document Numérique.

Article submitted 15 October 2018. Resubmitted 26 November 2018. Final acceptance 27 November 2018. Final version published as submitted by the authors.