A Session Based Personalized Search Using An Ontological User Profile

Mariam Daoud
IRIT, Paul Sabatier University
118, Route Narbonne
Toulouse, France
daoud@irit.fr

Lynda Tamine-Lechani
IRIT, Paul Sabatier University
118, Route Narbonne
Toulouse, France
lechani@irit.fr

Mohand Boughanem
IRIT, Paul Sabatier University
118, Route Narbonne
Toulouse, France
bougha@irit.fr

Bilal Chebaro
Lebanese universiy
Faculty of sciences
Beirut, lebanon
bchebaro@ul.edu.lb

ABSTRACT

Within the information overload on the web and the diversity of the user interests, it is increasingly difficult for search engines to satisfy the user information needs. Personalized search tackles this problem by considering the user profile during the search. This paper describes a personalized search approach involving a semantic graph-based user profile issued from ontology. User profile refers to the user interest in a specific search session defined as a sequence of related queries. It is built using a score propagation that activates a set of semantically related concepts and maintained in the same search session using a graph-based merging scheme. We also define a session boundary recognition mechanism based on tracking changes in the dominant concepts held by the user profile relatively to a new submitted query using the *Kendall* rank correlation measure. Then, personalization is achieved by re-ranking the search results of related queries using the user profile. Our experimental evaluation is carried out using the HARD 2003 TREC collection and shows that our approach is effective.

Categories and Subject Descriptors

IAR [**Information Access and Retrieval**]

General Terms

Design, Experimentation, Performance

Keywords

User profile, ontology, personalization, session boundary recognition

SAC'09 March 8-12, 2009, Honolulu, Hawaii,U.S.A.
Copyright 2009 ACM 978-1-60558-166-8/09/03 ...$5.00.

1. INTRODUCTION

Most existing search engines are classical content-based systems characterized by ”one size fits all” approaches, where the information retrieval (IR) process is based on the query-document matching pattern. Such pattern provides the same results for the same keyword queries even though these latter are submitted by different users with different intentions. For example, the keyword query *python* may refer to *python* as a snake as well as the *python* programming language.
Personalized IR has become a promising area for disambiguating the web search and therefore improving retrieval effectiveness by modeling the user profile by his interests and preferences. User profile could be inferred from the whole search history to model long term user interests [7] or from the recent search history to model short term user interests [6]. Mining short term user interests in a personalized retrieval task requires a session boundary mechanism that allows grouping related queries together. The UCAIR system [6] identifies session boundaries using a semantic similarity measure between successive queries using term relations. Concerning the representation model, user interests could be represented as a set of keyword vectors or class vectors [7], a set of concepts [3] or an instance of predefined ontology [2,5]. Indeed, ontology provides a highly expressive ground for describing user interests and a rich variety of interrelations among them and allows encountering new topics of interests.
In this paper, we describe a personalized search approach that represents the user profile as a weighted graph of semantically related concepts of predefined ontology, namely the ODP[1]. The user profile is built by accumulating graph based query profiles in the same search session. We define also a session boundary recognition mechanism that allows using the appropriate user profile to re-rank search results of queries allocated in the same search session.
Unlike previously cited work, our approach has several new features. First, the user profile is represented as a graph of the most relevant concepts of ontology in a specific search session and not as an instance of the entire ontology [2]. This allows using the most suitable user interest to personalize the search. Second, we build a single user interest across

[1] http://www.dmoz.org

related queries using a session boundary recognition mechanism based on a topical dependant similarity measure. The user interest is built for a single query in [5] where the issue of related queries is not tackled in real web environment.

The remaining of this paper is organized as follows. In section 2, we describe our approach for building and maintaining the ontological user profile and setting a session boundary recognition mechanism. In section 3, we present our search personalization method. The experimental evaluation and results are presented in section 4. In the final section, we present our conclusion and plan for future work.

2. BUILDING THE USER PROFILE OVER A SEARCH SESSION

We build and maintain the user profile by aggregating graph based query profiles in the same search session. We define a search session S at time s as a sequence of related search activities performed by queries $\{q^0, .., q^{s-1}, q^s\}$ submitted respectively at time $\{0, .., s-1, s\}$. The user profile has a hierarchical (tree) component made of "is-a" links, and a non hierarchical component made of cross links of different types predefined in the ontology, namely the ODP, defined as a directed graph G=(V,E). Figure 1 illustrates an example of a user profile inferred from the ODP ontology and corresponding to the *computer language programming* interest. Building the user profile could be detailed in three main stages: (1) building the query profile over a search activity, (2) initializing and maintaining the user profile over a search session, (3) detecting a possible session boundary when a new query is submitted.

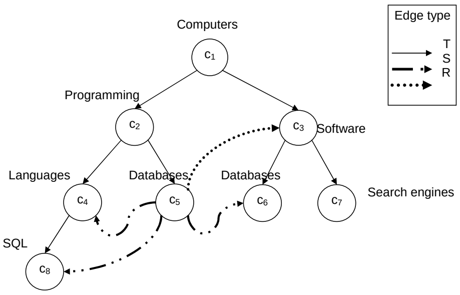

Figure 1: A Portion of an ontological user profile

2.1 Building the query profile over a search activity

Our query profiling begins by collecting the user's documents of interest D^s returned with respect to query q^s submitted at time s. We assume that a document retrieved by the search engine with respect to a query is relevant if it generates some observable user behaviors (page dwell time, click through, saving, printing etc). We create the query context K^s as being the centroid of the documents in D^s. The weight of term t in K^s is computed as follows:

$$K^s(t) = \frac{1}{|D^s|} \sum_{d \in D^s} w_{td} \tag{1}$$

where w_{td} is the weight of term t in document d computed as: $w_{td} = tf * log(n/n_t)$ where tf is the frequency term in document d, n is the total number of documents in the test collection and n_t is the number of documents containing term t.

Then, we map the query context K^s on the ODP ontology using the cosine similarity measure. Each concept c_j of the ontology is represented as a single term vector $\vec{c_j}$ issued from the web pages classified under this concept as described in previous work [1]. Given a concept c_j its similarity score with $\vec{K^s}$ is computed as follows:

$$score(c_j) = cos(\vec{c_j}, \vec{K^s}) \tag{2}$$

We obtain an initial weighted concept set, called θ^s. Based on this set, we infer the query profile using one-hop score propagation. We present in Algorithm 1 the process of inferring the graph-based query profile. Our intuition behind this algorithm is to represent the user interest as a graph of semantically related concepts located in different portions of the ontology. Score propagation activates semantically related concepts through both hierarchical and cross links of the ontology in order to induce a single graph or disconnected concept-based graphs. We re-use the edge weight setting adopted in [4] in our score propagation as follows: $w_{ij} = \alpha_S$ for $e_{ij} \in S \cup T$, $w_{ij} = \alpha_R$ for $e_{ij} \in R$, where

Algorithm 1 Building a graph-based ontological query profile using a score propagation strategy

```
input: $\theta^s$ an initial set of activated concepts
output: $G^s_q = (Vs_q, Es_q)$ the query profile
$\theta^s = \{c_1, c_2, .., c_n\}$, $ListGraphs = \emptyset$
for each concept $c_i \in \theta^s$ do
  $Queue_i = \{c_i\}$
  //initialization of the graph induced by concept $c_i$
  $G_i = (V_i, E_i), V_i = V_i \cup \{c_i\}, E_i = \emptyset, w(G_i) = score(c_i)$
  while $Queue_i.HasElement()$ do
    $c_j = Queue_i.PopElement()$
    //extract all linked concepts (is-a, symbolic, related)
    $\ell_j = GetLinkedConcepts(c_j)$
    for each concept $c_k \in \ell_j$ do
      if $e_{jk} \in S$ the edge type is symbolic then
        $\alpha = \alpha_S$
      else if $e_{jk} \in R$ the edge type is related then
        $\alpha = \alpha_R$
      end if
      //score propagation pattern for all linked concepts
      $score(c_k) = (\alpha * score(c_j) + score(c_k))/(\alpha + 1)$
      $V_i = V_i \cup c_k, E_i = E_i \cup e_{jk}, w(G_i) = w(G_i) + score(c_k)$
      if $c_k \in \theta^s$ then
        $\theta^s = \theta^s - \{c_k\}$
        $Queue.PushElement(c_k)$
      end if
    end for
  end while
  $ListGraphs = ListGraphs \cup \{G_i\}$
end for
//if there are graphs $G_m$, $G_n$ having common concepts
for each $G_m, G_n \in ListGraphs$ do
  if $V_m \cap V_n \neq \emptyset$ then
    $E_m = E_m \cup E_n, V_m = V_m \cup V_n, w(G_m) = w(G_m) + w(G_n)$ // merge the graphs together
  end if
end for
$G^s_q = argmax_{G_i \in ListGraphs}(w(G_i))$;
```

e_{ij} is the edge linking the concept i to the concept j. We set $\alpha_S = 1$ because *symbolic* links seem to be treated as first-class link "is-a" in the ODP web interface, and we set $\alpha_R = 0.5$ because *related* links are treated differently on the ODP web interface, labeled as "see also" topics. Each concept c_i propagates its weight to all its linked concepts c_k and interrelated concepts are grouped together in order to induce a single weighted graph G_i. The query profile is represented finally by the most relevant graph among the created ones. Indeed, we assume that the most relevant graph has greater number of related concepts, initially weighted or activated according to the ontology. For this aim, we define a graph relevance estimation based on summing the weights of the concept nodes of the graph.

2.2 Initializing and maintaining the user profile over a search session

The user profile G_u^0 is initialized by the graph-based ontological profile G_q^0 of the first query submitted in the search session. We maintain the user profile in the same search session when new submitted query q^{s+1} is correlated to the current user profile G_u^s. User profile maintaining process is based on merging G_u^s and G_q^{s+1} as follows:

- accumulating the weights of possible common concepts which allows bringing them to the top of the user profile representation.
- adding nodes and edges to the user profile, which allows taking into account all possible concepts in which the user has shown interest in the search session.

2.3 Detecting session boundary

We propose a session boundary recognition method using the *Kendall* rank correlation measure that quantifies the conceptual correlation ΔI between the user profile G_u^s and the query q^{s+1}. We choose a threshold σ and believe the queries are from the same session if the similarity is above the threshold.

Here, the term based query vector $\vec{q_t^{s+1}}$ (where terms are weighted according to their frequency in the query) is mapped onto the ontology in order to represent the concept vector $\vec{q_c^{s+1}}$. We adopt a context-sensitive query weighting scheme by introducing the query frequency (QF) in the current search session, in order to rank concepts in the top of $\vec{q_c^{s+1}}$ when they are close to the user profile. Indeed, query vector $\vec{q_c^{s+1}} =< w_1, w_2, .., w_i, ... >$ is computed as follows:

$$w_i = CW(q^{s+1}, c_i) * QF(c_i) \qquad (3)$$

where the CW and QF are defined as:

$$QF(c_i) = \frac{|\vec{q}|^S}{< |\vec{q}|^S, c_i >}, CW(q^{s+1}, c_i) = cos(\vec{q_t^{s+1}}, \vec{c_i}) \qquad (4)$$

$|\vec{q}|^S$ is the total number of related queries submitted in search session S, $< |\vec{q}|^S, c_i >$ is the number of user profiles built in the session S and containing concept c_i.

Thus, the similarity ΔI is computed as follows.

$$\Delta I = (\vec{q_c^{s+1}} o \vec{G_u^s}) = \frac{\sum_{c_i}\sum_{c_j} S_{c_ic_j}(\vec{q_c^{s+1}}) \times S_{c_ic_j}(\vec{G_u^s})}{\sqrt{\sum_{c_i}\sum_{c_j} S^2_{c_ic_j}(\vec{q_c^{s+1}}) \times \sum_{c_i}\sum_{c_j} S^2_{c_ic_j}(\vec{G_u^s})}} \qquad (5)$$

$$S_{c_ic_j}(\vec{v}) = sign(\vec{v}(c_i) - \vec{v}(c_j)) = \frac{\vec{v}(c_i) - \vec{v}(c_j)}{|\vec{v}(c_i) - \vec{v}(c_j)|}$$

Where, c_i and c_j are two concepts issued from both the query and the user profile, $\vec{q_c^{s+1}}(c_i)$ (resp. $\vec{G_u^s}(c_i)$) is the weight of the concept c_i in $\vec{q_c^{s+1}}$ (resp. $\vec{G_u^s}$). The correlation values ΔI are in the range of [-1 1], where a value closer to -1 means that the query and the user profile are not similar, and a value closer to 1 means that they are very related.

3. SEARCH PERSONALIZATION

We personalized search results of query q^{s+1} in the same search session using the user profile G_u^s by combining for each retrieved result d_k, the initial score S_i and a contextual score S_c as follows:

$$S_f(d_k) = \gamma * S_i(q, d_k) + (1 - \gamma) * S_c(d_k, G^s) \qquad (6)$$

$$0 < \gamma < 1$$

The contextual score S_c of result d_k is computed using the cosine similarity measure between the result d_k and the concepts of the user profile G_u^s as follows:

$$S_c(d_k, G_u^s) = \frac{1}{h} \sum_{j=1..h} score(c_j) * cos(\vec{d_k}, \vec{c_j}) \qquad (7)$$

Where c_j is a concept in the user profile, $score(c_j)$ is the weight of concept c_j in the user profile G_u^s.

4. EXPERIMENTAL EVALUATION

4.1 Experimental Data sets

Since the queries of average web users tend to be short and ambiguous, we used query topics provided by TREC[2] 2003 HARD Track. As our approach requires a session based evaluation setup, we simulate a search session by a single topic and related queries by creating three subtopics of the main topic. For each topic we selected randomly a relevant document set called the profile set provided by TREC data. This set is divided into equally-sized three subsets. We created the centroid vector of each profile subset using formula (3) and built the *sub-topics* from the top three terms of each centroid vector. In our search experiments, the relevant documents of a subtopic are those of its associated topic and the profile sets of all the topics are excluded from evaluation.

In order to validate the reliability of the generated *subtopics*, we computed first the percentage of relevant overlapping documents returned by the system with respect to the *subtopics* comparatively to their main topic and proved that *subtopics* contain significant terms able to return common relevant documents with the main topic. Moreover, we compute the percentage of average non-overlapping documents over the Top-n results (Top-20 and Top-50) returned by the system between the *subtopics* themselves and proved that even though the *subtopics* were built from the same topic, they do not return same documents (average non-overlapping estimated higher than 40% at Top-20).

4.2 Experimental design and results

In order to evaluate the search personalization effectiveness independently of the session boundary recognition, we

[2] Text REtrieval Conference:http://trec.nist.gov

divided the HARD topics into two topic sets. The first one is a *training topic set* used to identify the optimal session boundary threshold value σ. The second one is a *test topic set* used for testing our personalized approach.

4.2.1 Evaluating the session boundary recognition mechanism

The goal of this experiment is to evaluate the accuracy of the session boundary recognition mechanism and identify the best threshold value σ. For this aim, we define a real evaluation scenario that consists of simulating search sessions by aligning successively 15 main topics of the HARD track along a *training subtopic sequence*. We define a critical *subtopic sequence* by aligning successively the most correlated topics in order to test our system using a threshold value issued from the most difficult query sequence.

We computed the *Kendall* correlations between each subtopic and the user profile built across subtopics related the same topic and obtained values in $[-0.6\ \ +0.01]$. Figure 2 shows the *Kendall* correlation values presented on the Y-axis computed along the *training subtopic sequence* presented on the X-axis. *Subtopics* are assigned by the number of the topic dotted by the number of the *subtopic* $\{1, 2, 3\}$.

A correct session boundary is marked by a vertical line. We notice that when the correlation degree of a *subtopic* tend to be very low, the probability of detecting a session boundary becomes high.

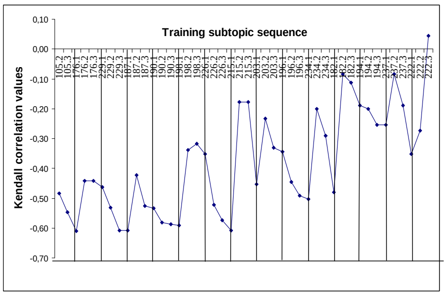

Figure 2: Kendall correlation values computed across a *subtopic sequence*

To evaluate the session boundary recognition accuracy, we define two measures: the precision $P_{intra}(\sigma)$ of allocating related queries into the same search session and the precision $P_{inter}(\sigma)$ of detecting correct session boundaries using a threshold value σ. They are defined as follows:

$$P_{intra}(\sigma) = \frac{|RQ|}{|TRQ|}, P_{inter}(\sigma) = \frac{|BQ|}{|TBQ|} \tag{8}$$

Where $|RQ|$ is the number of queries identified as correctly correlated and $|TRQ|$ is the total number of correlated queries. $|BQ|$ is the number of the correctly identified session boundaries and $|TBQ|$ is the total number of session boundaries. The optimal threshold value is identified for the maximum value of the product of both measures that maximizes their accuracy measures as follows:

$$\sigma = argmax_{\sigma}(P_{intra}(\sigma) * P_{inter}(\sigma)) \tag{9}$$

We show in figure 3 the accuracy of the *Kendall* correlation measure with varying σ in $[-0.6\ \ +0.01]$ range values. In our *training subtopic sequence*, there are 14 session boundaries (TBQ =14) and 30 correlated *subtopics* (TRQ=30, where two correlated *subtopics* must be encountered per topic). Results show that -0.34 is the best σ value maximizing P_{intra} at 53.33% and P_{inter} at 85.71%.

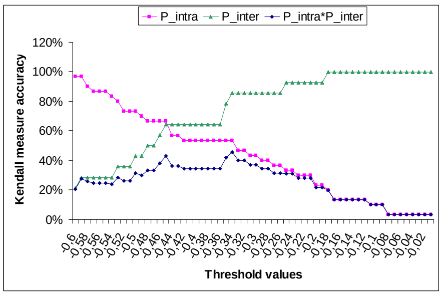

Figure 3: Kendall correlation accuracy as a result of varying the threshold value

4.2.2 Evaluating search personalization

We evaluate our personalized approach on a *test subtopic sequence* on one hand, and on the *test topic set* on the other hand. We compare the standard search (using only the query ignoring any user profile) to the personalized search (performed using the query and the user profile). Our baseline search is based on the BM25 relevance scoring formula given as follows:

$$w_{td} = tf_d \times \frac{log(\frac{n-n_t+0.5}{n_t+0.5})}{K_1 \times ((1-b) + b \times \frac{dl}{avgdl}) + tf} \tag{10}$$

where tf_d is the frequency of term t in document d, n is the total number of documents in the test document collection and n_t is the number of documents containing term t, $K_1 = 2$ and $b = 0.75$. We fix γ at the value 0.3 in the formula (6) and $h = 3$ in formula (7). We have shown in preliminary experiments that this value is the best value among the range of values [0.1 0.9], achieving the higest retrieval performance. Evaluation measures are the Top-n precision and Top-n recall computed as follows:

$$Top{-}nRecall = \frac{RelDoc_n}{RelDoc_{total}}, Top{-}nPrecision = \frac{RelDoc_n}{n} \tag{11}$$

where $RelDoc_n$ is the number of relevant documents within the top n search results, $RelDoc_{total}$ is the total number of relevant documents for the given *subtopic*, by excluding the profile set of its topic.

(A) Evaluating search personalization on the subtopics
With respect to the ordered numbers of the topics done by HARD TREC, we aligned the associated *subtopics* randomly in the *test subtopic sequence*. The evaluation scenario consists of using the value ($\sigma = -0.34$) in the session boundary recognition and create the user profile by merging the query profiles in the same search session. We obtained a precision of allocating related search activities in the same search session equal to 71.42% and a precision of detecting session boundaries on the *test subtopic sequence* equal to 40.47%. Figure 4 shows the average Top-n precision and Top-n recall achieved by the personalized search comparatively to the standard one on the *test subtopic sequence*.

We see that personalized search achieves higher retrieval precision and recall comparatively to the standard search.

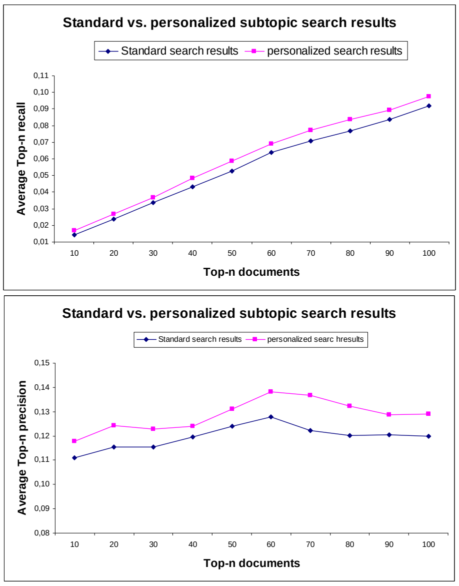

Figure 4: Average Top-n recall and Top-n precision comparaison between the personalized search and the standard search on the subtopic sequence

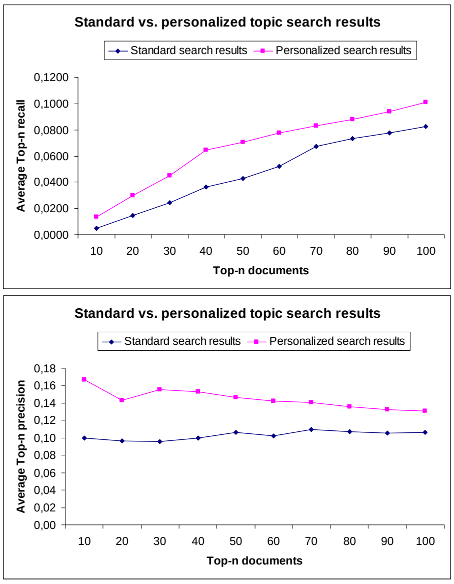

Figure 5: Average Top-n recall and Top-n precision comparaison between the personalized search and the standard search on the HARD topics

The improvement in both Top-n precision and Top-n recall is significant at 5% level. Maximum improvements are 11.95% and 23.62% achieved respectively at average top-70 precision and at average top-10 recall. Results proved that personalizing search using the ontological user profile built across related search activities improves the search accuracy of queries allocated in the same search session.

(B) Evaluating search personalization on the topics
We evaluate the effectiveness of search personalization on a topic using the user profile created across its subtopics. We present in figures 5 the Top-n precision and Top-n recall achieved by the personalized search comparatively to the standard search averaged over the topic set. The percentage of improvement of personalized search is 66.7% and 188.24% achieved respectively at average Top-10 precision and average Top-10 recall. Our evaluation results confirm that difficult queries that were intentionally selected from the HARD topics of TREC, perform better retrieval precision when the search results are re-ranked using the related user profile, especially in the top ranked documents.

5. CONCLUSION AND OUTLOOK

We presented in this paper an approach for personalizing search using a graph based user profile. The user profile refers to the user interest in a particular search session and is built by aggregating related graph-based query profiles. We also defined a session boundary recognition mechanism that allows re-ranking the search results of queries allocated in the same search session using the user profile. Our experimental evaluation is carried out using the TREC 2003 HARD TRACK. Evaluation results show that our session boundary recognition mechanism achieves significant accuracy. Furthermore, our personalized search approach achieves significant precision improvement compared to the standard search. In future work, we plan to learn a user profile including diversity of user interests and evaluate the accuracy of the user profile. Moreover, we plan to define an adaptive session boundary threshold value and evaluate our session boundary recognition measures using real user data provided by web search engine log file.

6. REFERENCES

[1] M. Daoud, L. Tamine-Lechani, and M. Boughanem. Learning user interests for a session-based personalized search. In *Proceedings of the the Second International IIiX Symposium (IIIX'08)*, pages 57–64, London, U.K., 2008.

[2] S. Gauch, J. Chaffee, and A. Pretschner. Ontology-based personalized search and browsing. *Web Intelli. and Agent Sys.*, 1(3-4):219–234, 2003.

[3] F. Liu, C. Yu, and W. Meng. Personalized web search for improving retrieval effectiveness. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):28–40, 2004.

[4] A. G. Maguitman, F. Menczer, H. Roinestad, and A. Vespignani. Algorithmic detection of semantic similarity. In *Proceedings of the 14th international conference on World Wide Web*, pages 107–116, New York, NY, USA, 2005. ACM.

[5] A. Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. In *CIKM'07: Proceedings of the ACM Conference on information and knowledge management*, pages 525–534, New York, NY, USA, 2007. ACM.

[6] S. Sriram, X. Shen, and C. Zhai. A session-based search engine. In *Proceedings of the International ACM SIGIR Conference*, 2004.

[7] L. Tamine-Lechani, M. Boughanem, and N. Zemirli. Personalized document ranking: exploiting evidence from multiple user interests for profiling and retrieval. *In Journal of Digital Information Management*, 2008.