A contextual evaluation protocol for a session-based personalized search

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

ABSTRACT

Most existing evaluation protocols in IR are laboratory-based ones. They are based on a controlled evaluation methodology and lack generally of user evidences expressing his search context in the data set and the evaluation protocol as well. This paper proposes an evaluation protocol for a session-based personalized search. It is based on an enhanced TREC HARD collection with simulated user profiles issued from simulated search sessions. The experimental results show the effectiveness of our personalized search approach according to the proposed evaluation protocol.

Keywords

Evaluation protocol, personalized search, user profile, search session, simulation

1. INTRODUCTION

The main goal of an IR system evaluation framework is to measure the corresponding retrieval performance according to efficiency and/or effectiveness. Regarding effectiveness evaluation, most existing evaluation protocols are laboratory-based ones where the user need is generally represented by the user query. Within the emergence of contextual IR, these protocols are not sufficient for evaluating the system performance under the challenge of context. Attempts to extend the laboratory model to user centred evaluation have been achieved via the TREC Interactive track [6] and HARD track [1] by integrating the user context in the data set. Despite these extensions, the overall evaluation still is restricted due to the exploitation of only specific contextual features. To alleviate such limitations, contextual evaluation methodologies have been proposed to support simulated user profile through contextual simulations [9] or real evaluation scenarios through user studies [3].

In this paper, we present an evaluation protocol devoted for a session-based personalized search. The user profile is simulated using hypothetic user interactions on documents provided by TREC and the search session is simulated by aligning generated sub-queries of a query along a sequence. The paper is organized as follows. Section 2 presents a short overview of contextual IR evaluation. Section 3 presents our approach of search personalization and a contextual IR evaluation integrating simulated user profiles and search sessions. Section 4 presents a conclusion and our perspectives for future works.

2. CONTEXTUAL IR EVALUATION: A SHORT OVERVIEW

Contextual IR evaluation aims at measuring the system performance by integrating the user context in the evaluation scenario. There are two main types of contextual evaluation: evaluation by context simulations and evaluation by user studies.

The first kind of evaluation simulates users and interactions by means of well defined retrieval scenarios (hypothesis) [10]. A contextual simulation in [9] used a document collection issued from a predefined Web ontology and simulates the user profile by a concept of the ontology. For a specific simulated concept /user profile, queries are generated automatically by the top terms representing the concept. The user profile is built using a set of documents classified under this concept, called the profile set. Other personalized approaches carry out a contextual simulation by enhancing TREC collection with simulated user profile [11, 4] represented by a TREC domain and built using the relevant documents of the queries annotated by the domain. Evaluation measures are the precision and recall in the case of extending a laboratory-based collection (TREC) or the average rank of the documents returned by the system and that are classified under a simulated concept in [9]. This evaluation method is worthwhile since (a) it is less time consuming and costly than experiments with real users and (b) allows comparative evaluation with respect to the defined scenarios [12].

The evaluations by user studies are carried out with real users to test the system performance through real user interactions with the system. There are two types of user studies adopted in the domain. The first one [8] consists of using a search interface plugged within a TREC collection where the user is asked to reformulate queries related to a predefined topic by TREC in order to define a search session. The user profile is represented by the user search history in a search session. The second kind of contextual evaluation by user studies [3, 7] is carried out using an API search interface

(like the Google API) that allows the user to perform natural search. Evaluation measures are the average rank of the clicked documents by the user [3] or the average precision and recall over the top N returned documents by the system [7, 8]. Recall is computed using all judgements given by all the users for the query. The main limitation introduced by user studies is that experiments are not repeatable.

3. A CONTEXTUAL EVALUATION FOR A SESSION-BASED PERSONALIZED SEARCH

3.1 A session-based personalized search

Our approach of search personalization [5] aims at representing and personalizing the search using a graph-based user profile issued from a predefined ontology. The user profile is built over a search session by combining graph-based query profiles; personalization is achieved re-ranking the search results of queries related to the same search session.

A query profile G_q^s is built by exploiting the documents clicked D_r^s by the user and returned with respect to the query q^s submitted at time s. First a Keyword query context K^s is calculated as the centroid of documents in D_r^s:

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

K^s is matched with each concept c_j of the ontology represented by single term vector $\vec{c_j}$ using the cosine similarity measure. The scores of the obtained concepts are propagated over the semantic links as explained in [5]. We select the most weighted graph of concepts to represent the query profile G_q^s at time s. The user profile G_u^0 is initialized by the profile of the first query submitted by the user. It is updated by combining it with the query profile G_q^{s+1} of a new related query submitted at time $s+1$.

A session boundary delimitation based on the Kendall rank correlation measure is used to compute the correlation degree between a new submitted query q^{s+1} and the user profile G_u^s. When the correlation value $\Delta I = q^{s+1} o G_u^s$ is above a predefined threshold value $\sigma*$, the search results returned by the system are re-ranked by combining their original score with their contextual score as follows:

$$S_f(d_k) = \gamma * S_i(q, d_k) + (1 - \gamma) * S_c(d_k, G_u^s) \qquad (2)$$

$$0 < \gamma < 1$$

The contextual score S_c is computed between the result d_k and the top h weighted concepts of the user profile G_u^s as follows:

$$S_c(d_k, G_u^s) = \frac{1}{h} . \sum_{j=1..h} score(c_j) * cos(\vec{d_k}, \vec{c_j}) \qquad (3)$$

3.2 Contextual evaluation protocol derived from HARD TREC

We present a contextual evaluation protocol which integrates the user profile into the evaluation process. Involved components are the query set consisting of related sub-topics of the HARD TREC topic set, the user profile built across a simulated search session defined by a sequence of related subtopics and the evaluation strategy that aims at training the session boundary delimitation and then testing the personalized search using the best system parameter.

3.2.1 Queries

Queries are the topics provided by TREC[1] 2003 HARD Track [2]. As the user profile is built over a search session (a sequence of related queries) and there is no information available in the collection concerning the correlation between the queries, we generate three subtopics per topic that define a simulated search session. The adopted strategy for generating the *sub-topics* of a topic consists of:

- extracting a document *profile set* that consists of the top r relevant documents returned by the system with respect to the topic.
- dividing the document *profile set* into equally-sized three *profile subsets*.
- creating the centroid vector of each profile subset using formula (1) by representing each document as an ordered term vector using *tf*idf* weighting scheme.
- building each *sub-topic* by selecting the top three terms of the centroid vector.

In our experiments, we excluded topics that achieve null average precision and we set $r = 9$, which implies excluding topics that have less than 9 relevant documents returned by the system. We obtain a total of 8 topics in the main data query set.

In order to validate the reliability of the generated *subtopics*. We computed the percentage of average subtopic-topic relevant document overlap and the percentage of non-overlapping documents over the Top-n results (Top-20 and Top-40) returned by the system between the *subtopics* themselves. Results in figure 1 prove that the subtopics have more than (50%) of common relevant concepts with the topic which confirms that they are related. Moreover, results in figure 2 prove that even though the *subtopics* were built from the same topic, they contains different terms which leads to do not return same documents (average non-overlapping estimated higher than 40% at Top-20). For the topic 48, we obtain a null percentage of non-overlapping documents at Top-40 because the generated subtopics contains two common terms over three. Subtopics still correlated as they are considered as different reformulations of the same topic.

3.2.2 Document collection

The main document collection used in the HARD track contains the newswire text from AQUAINT corpus and U.S. government documents.

3.2.3 User profile

The user profile is integrated in the evaluation strategy according to a simulation algorithm that generates it using hypothetic user interactions for each topic provided by TREC assessors. We consider that a topic holds an information need that represents a simulated user interest / user profile. The simulated user profile is created across related subtopics as follows:

1. Creating the ontological query profile for each *subtopic*. We notice that the query context K^s for a subtopic q^s is created using its appropriate relevant document *profile subset*.

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

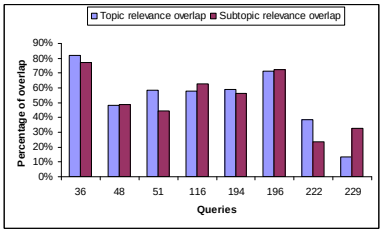

Figure 1: Percentage of relevant document overlap

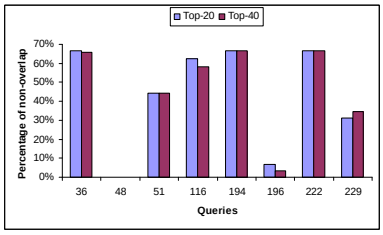

Figure 2: Percentage of non-overlapping documents

2. Initializing the user profile by the ontological profile of the first query of the session.

3. Applying the session boundary recognition mechanism using an appropriate threshold value (σ^*) when a new *subtopic* has to be processed along a query sequence. If the *subtopic* is correlated to the current user profile, this latter is then updated by combining it with the the profile of the new subtopic.

3.2.4 Evaluation protocol

The evaluation protocol is designed to tune the session boundary parameter in a training stage and evaluate the effectiveness of the personalized search in the testing stage. For this purpose, we divided the HARD topics into two topic sets: a *training topic set* and a *testing topic set*.

A. Training stage

In this stage, we tune the optimal correlation value of the session boundary delimitation that achieves the best performance. To do so, we set a training query sequence created by aligning successively subtopics of the training topic set given by the HARD track.

First we computed *subtopic-profile* correlations values according to the Kendall measure between each *subtopic* on the training sequence and the user profile built across previous and related *subtopics* (related to the same topic). Once the correlation values are computed, we proceed to tune an average correlation value that maximize the precision of detecting correct session boundaries $P_{intra}(\sigma)$ and correct correlated queries $P_{inter}(\sigma)$ defined as follows:

$$P_{intra}(\sigma) = \frac{|RQ|}{|TRQ|}, P_{inter}(\sigma) = \frac{|BQ|}{|TBQ|} \quad (4)$$

Where $|RQ|$ is the number of *subtopics* identified as correctly correlated according to the *subtopic sequence*, and $|TRQ|$ is the total number of *subtopics* that should be identified as correlated, $|BQ|$ is the number of *subtopics* indicating correct session boundaries, and $|TBQ|$ is the total number of session boundaries in the *subtopic sequence*. The optimal session boundary threshold value is identified when both measures ($P_{intra}(\sigma)$ and $P_{inter}(\sigma)$) reach the highest accuracy.

$$\sigma* = argmax_{\sigma}(P_{intra}(\sigma) * P_{inter}(\sigma)) \quad (5)$$

B. Testing stage

In this stage, we evaluate the search personalization along a testing query sequence (different from the training query sequence). The evaluation is based on comparing the typical search performed using only the query to the personalized search using the query and the correlated user profile.

This stage could be explained by the following steps:

- Creating the testing query sequence by aligning the subtopics of the testing topic set.
- Along this sequence, we used the optimal threshold value in the session boundary recognition mechanism to build the user profile across related testing subtopics. For each subtopic having a correlation value greater than the optimal threshold value $\sigma*$, we proceed to:
  - perform the personalized search on this subtopic by re-ranking its search results using the correlated user profile,
  - update the user profile by combining it with the query profile of the subtopic being processed.

3.3 Experimental Results

The evaluation protocol presented above is used to evaluate our search personalization approach. A total of 8 topics are divided equally into training topics and testing topics.

3.3.1 Evaluating the session boundary delimitation

According to the training stage, we align a total of 4 training topics along a sequence of 12 subtopics. We computed the subtopic-profile correlation along the training sequence. The optimal session boundary threshold is identified at $\sigma^* = -0.41$ achieving significant precision of identifying correct session boundaries (66%) and correct correlated queries (75%).

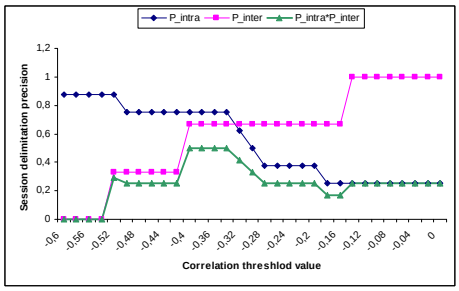

Figure 3: Kendall correlation values computed across the training *subtopic sequence*

3.3.2 Evaluating the personalized search performance

According to the testing stage, we evaluate the effectiveness of the personalized search by comparing it to the typical search. A total of 4 testing topics generates 12 subtopics along a testing sequence. We used the optimal session boundary identification ($\sigma^* = -0.41$) to detect related subtopics.

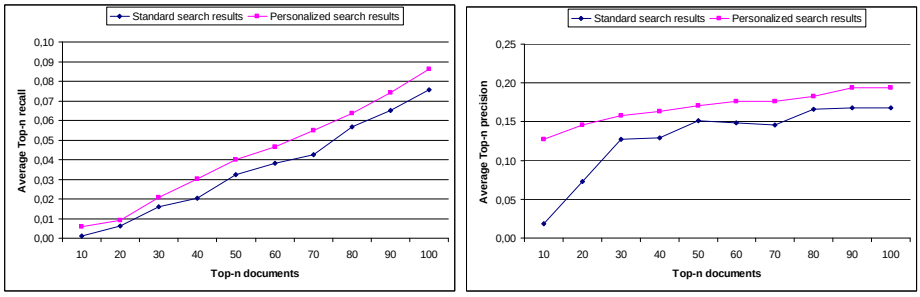

Figure 4: Average Top-n recall and Top-n precision comparison between the personalized search and the standard search on the subtopic sequence: profile built from the top ranked documents of the topic

Figure 4 shows the average Top-n precision and Top-n recall achieved by personalized search comparatively to the standard one on the *subtopic* sequence. Results prove that personalized search achieves higher retrieval precision and recall comparatively to the standard search. The best performance is achieved by the personalized search in terms of top-10 precision (12.73) and top-10 recall $(0,57)$ comparatively to the standard search having lower top-10 precision (1.82) and lower top-10 recall $(0,10$). We have conducted experiments in previous work [5] where the profile set of each topic was defined by the first 30 relevant documents listed in Qrels [5]. We notice that the improvement is much more higher when the user profile is built using the top ranked documents returned by the system with respect to the topic. This confirms that our method achieves an effective personalization as in real world search engines.

4. CONCLUSION

In this paper, we have presented an evaluation protocol devoted for a session-based personalized search. More precisely, this protocol is suitable for the evaluation of a personalized search that requires a session boundary delimitation to build the user profile. It is based on the TREC HARD collection where the user profile is simulated for each topic using a set of relevant returned documents by the system and that are previously judged by TREC assessors. We defined also a session-based evaluation scenario that integrates the search session as a sequence of subtopics generated for a specific topic. We evaluated our approach according the proposed evaluation protocol and show that our approach is effective. In future work, we plan to extend this protocol by using real user data provided from a search engine log file. Extending the protocol aims at testing the effectiveness of the personalized search based on relevance judgements given by the user who submits the query. It consists of defining a query set provided from real users and the associated relevance judgements provided from the clickthrough data available in the log file.

5. REFERENCES

[1] J. Allan. Hard track overview in trec 2003 high accuracy retrieval from documents. In *Proceedings of the* 12^{th} *text retrieval conference (TREC-12)*, pages 24–37. National Institute of Standards and Technology, NIST special publication, 2003.

[2] J. Allan. Hard track overview in trec 2003: High accuracy retrieval from documents. In *TREC*, pages 24–37, 2003.

[3] V. Challam, S. Gauch, and A. Chandramouli. Contextual search using ontology-based user profiles. In *Proceedings of RIAO 2007, Pittsburgh USA*.

[4] M. Daoud, L. Tamine, and M. Boughanem. Learning user interests for session-based personalized search. In *ACM Information Interaction in context (IIiX), London, 14/10/2008-17/10/2008*.

[5] M. Daoud, L. Tamine, M. Boughanem, and B. Chebaro. A Session Based Personalized Search Using An Ontological User Profile. In *ACM Symposium on Applied Computing (SAC), Haiwai (USA)*, pages 1031–1035. ACM, march 2009.

[6] D. Harman. Overview of the the 4^{th} text retrieval conference (trec-4). In *Proceedings of the* 4^{th} *text retrieval conference (TREC-4)*, pages 1–24. National Institute of Standards and Technology, NIST special publication, 1995.

[7] 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.

[8] X. Shen, B. Tan, and C. Zhai. Context-sensitive information retrieval using implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference*, pages 43–50, New York, NY, USA, 2005. ACM.

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

[10] L. Tamine, M. Boughanem, and M. Daoud. Evaluation of contextual information retrieval: overview of issues and research. *Knowledge and Information Systems (Kais)*, 2009.

[11] L. Tamine, M. Boughanem, and W. Zemirli. Exploiting multi-evidence from multiple user's interests to personalizing information retrieval. *IEEE International Conference on Digital Information Management(ICDIM 2007)*, 2007.

[12] R. White, I. Ruthven, J. Jose, and C. Van Rijsbergen. Evaluating implicit feedback models using searcher simulations. *ACM Transactions on Information Systems*, 23(3):325–361, 2005.