A Diary Study-Based Evaluation Framework for Mobile Information Retrieval

Ourdia Bouidghaghen, Lynda Tamine, and Mohand Boughanem

IRIT-University Paul Sabatier, 118 Route de Narbonne, 31062, Toulouse, cedex 09
{bouidgha,tamine,bougha}@irit.fr

Abstract. In this poster, we propose an evaluation framework that investigates the integration of the user context (interests, location and time) into the evaluation process of mobile IR. Our approach is based on a diary study where users are asked to log their queries annotated by their location and time. Users' interests are explicitly acquired or implicitly learned based on users' relevance judgments for the retrieved documents answering their queries. We propose two evaluation protocols namely training/test in chronological order and k-fold cross validation. We exploit this framework in order to evaluate the performance of our context-based personalized mobile search approach. Experimental results show the stability performance of our approach according to the proposed evaluation protocols and demonstrate the viability of the diary approach as a means to capture context in evaluation.

Keywords: Experimental evaluation, evaluation framework in mobile context, diary study, location, time, user's interests.

1 Introduction

Within the emerging mobile IR environment, the focus is over context models including user's interests and environmental data (time, location, near persons, activity, device and networks) [1]. Contextual IR evaluation in this environment aims at measuring the system performance by integrating the user context in the evaluation process [2]. We can classify evaluation methodologies within mobile contextual IR, to two main types: 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). Contextual simulation frameworks allow systems to be evaluated, according to a formative view, with less regard for constraints that arise from using sensor technologies, and several social and personal differences of users in interaction with the system. The contextual simulation framework proposed in [3] is based on hypothetic user search context and queries. User context is represented by a set of possible locations and users' interests are integrated in the evaluation strategy according to a simulation algorithm that generates them using hypothetic user interactions for each query.

P.-J. Cheng et al. (Eds.): AIRS 2010, LNCS 6458, pp. 389–398, 2010.

In [4], authors propose a contextual simulation framework based on a set of simulated context descriptors that include location, time and user activities. User's queries are automatically formulated from the context descriptors using different techniques. Context simulation based evaluation method is worthwhile since it is less time consuming and costly than experiments with real users. However, the method has still areas of uncertainty, for example the choice of assumptions underlying the major scenarios is open to criticism for its lack of realism.

The evaluation by user studies is carried out with real users, called participants, to test the system performance through real user's interactions with the system. To evaluate the performance of contextualized search, each participant is required to issue a certain number of test queries and determine whether each result is relevant in its context. There are two types of user studies adopted in the domain. The first one [5] is based on the evaluation framework proposed in [6] which makes use of "simulated work task situations" and where users are assigned a set of predefined tasks to perform in predefined situations. This kind of user studies is criticized because it still rely on artificial information needs and may be confounded by inter-subject and order effects. The second kind of contextual evaluation by user studies [7] is carried out in realistic use settings. In these latter, users are free to use the system as they would wish to use it and for only as long as they want, submitting their own queries arising from their natural information needs within real and natural situations, rather than asking them to perform some predefined series of tasks. The advantage of user studies based evaluation is that they are conducted with real users and thus the relevance can be explicitly specified by them. The main limitation is that experiments are not repeatable, the extra cost they induce and they may be of little use if the system is not fully developed.

In the absence of a standard evaluation benchmark for a mobile contextual IR task, we propose in this poster an evaluation framework based on a diary study. Our evaluation framework keeps up the benefits of user study based evaluation by allowing evaluation with real users and real contexts and alleviates its requirement that the system be fully developed by allowing the evaluation of an early stage development system; moreover we estimate our framework to be easily extensible to include any other contextual aspects from the mobile environment (eg. near persons, activity, ...). Our approach is based on a diary study where mobile users are asked to log their queries annotated by their search context, here location and time. User's interests are explicitly acquired or implicitly learned based on their expressed relevance judgments for the retrieved documents for their queries. Two evaluation protocols training/test in chronological order and cross validation are experienced within this framework.

This poster is organized as follows: we first present our evaluation framework and introduce our experimental design in Section 2. We then present our approach for mobile search personalization, and its performance evaluation using the proposed evaluation framework in Section 3. Finally, we conclude and give perspectives for future work.

2 Proposition of an Evaluation Framework Based on a Diary Study

In our previous work [3] we have proposed an evaluation framework based on context simulation. The contribution of this poster is twofold: first we proposed a new evaluation framework based on a diary study as a tool that enables evaluation with real users in real contexts, second we compared evaluation results obtained using the two evaluation protocols.

Diary study is a method that has its roots in both psychological and anthropological research. In its simplest form, it consists of a representative sample of subjects recording information about their daily lives in situ for a given period. The data captured can then be analyzed in a variety of ways depending upon the nature of the data. Diary studies are presented in an early work by Rieman [8] as a workplace-oriented tool to guide laboratory efforts in the HCI field, they are exploited in [9] to analyze mobile information needs. In our work, we propose to undertake a diary study as a basis for collecting mobile information queries together with their external context namely time and location in situ. The diary entries are used as building blocks that compose the evaluation framework datasets.

2.1 Methodology

The focus of our framework is the evaluation of the effectiveness of a context-aware personalization technique for mobile search, in an early stage development. Such techniques involve the consideration of mobile search user's contexts namely interests, location and time in the development and the evaluation processes. The general process we adopted to build our framework is shown in Figure 1.

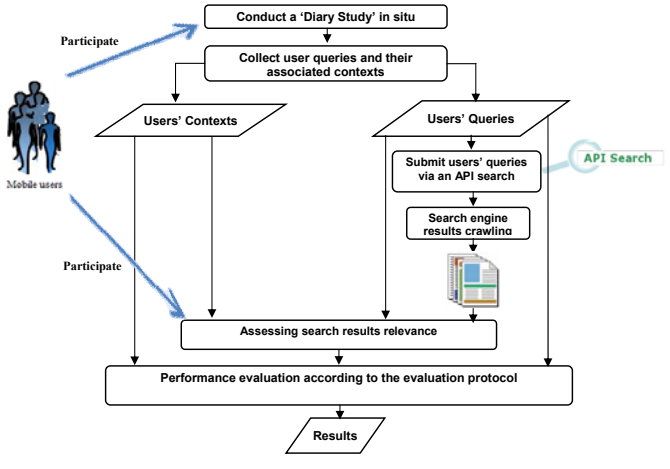

Fig. 1. Our diary study based evaluation framework

First, a diary study is conducted in situ, were real users are asked to log their queries together with their context whenever and wherever it occurs. The entire resulting diary entries are processed to extract user queries and contexts. Then, users queries are submitted to a standard search engine via an API. After, the resulting top N search engine documents are crawled, users are asked to judge these documents according to their queries and contexts. Finally, user's queries and contexts are integrated in the evaluation protocols. The general guidelines for conducting the diary study are: (1) Set the number of participants and the time of the diary study. (2) Assure that all the participants already have experience with using search engines on the web, using a PC or a mobile phone. (3) Set a description of the recording activities your are asking for, namely: recording the date, the time, the location, and the query the user have while he is mobile. (4) To avoid participants forgetting to record entries, send periodic reminders in order to keep participants on track. In what follows, we describe our datasets and evaluation protocols.

2.2 Datasets

Contextual Query Set. The diary study entries constitute a set of contextual queries. While many contextual information can be recorded, in this paper we only focused on the spatio-temporal context and users's interests. A contextual query can then be represented by: $Q_i^u =< q_i^u, l_i^u, t_i^u, g_i^u >$, where q_i^u (resp. l_i^u, t_i^u, and g_i^u) represents the i^{th} query (resp. time, location and interests) of the diary entries of a user u. Each contextual query is annotated with a description of its associated information needs and a narrative about what would be a relevant document belonging to it. Location (l_i^u) and time (t_i^u) information can be expressed as low level data or using semantic concepts depending on the application needs. The user interests (g_i^u) can be manually specified by participants themselves or automatically learned from the user manual judgments of returned documents for their past queries.

Ground Truth in Context. The document collection is to be built by collecting the top N results retrieved from a publicly available Web search API for each query blind of context. The relevance assessments for the documents are to be collected through an assessment tool (available on line). To do, each user who submitted a query (in the diary study), is asked to judge whether a document from the set of top N retrieved results as response to his query was relevant or not according to his query and its context. Relevance judgments are to be made using a three level relevance scale: relevant, partially relevant, or not relevant.

2.3 Evaluation Protocols

In order to evaluate contextualized techniques for mobile search, the set of queries is to be divided into two sets: a training set for learning the parameters of the underlying contextualization technique, and a testing set to evaluate the effectiveness of this technique. Having a set of K contextual queries by user, that

contains time information, two evaluation protocols are possible: training/test in chronological order and K-fold cross validation, to be applied on each users' set of queries. The only recommendation to be observed is to respect a *minimum* of 25 testings queries [10] in order to make the evaluation process outcomes significant. These two evaluation strategies are described as follows:

1. **Training/Test in Chronological Order:** this strategy keeps queries in their temporal order of emission, uses $Q_1^u \cdot\cdot Q_{i-1}^u$ past queries as the training set for the learning step, and tests with the following queries $Q_i^u \cdot\cdot Q_K^u$. This strategy is the simplest and more natural one, however effectiveness evaluation may depend heavily on which data points end up in the training set and which end up in the test set.

2. **K-fold Cross Validation:** this strategy divides the query set into *k* equally sized subsets, then uses *k-1* training subsets for learning and the remaining subset as a test set. The holdout method is repeated *k* times, each time, one of the *k* subsets is used as the test set and the other *k-1* subsets are put together to form a training set. The advantage of this method over the first protocol, is that all the queries are used for both training and testing and avoid consequently the bias on the choice of the training set.

We expect that the two protocols are applicable, and despite the difference between the two protocol strategies and the number of queries they allow to test, the evaluation results are expected to be consistent between them.

3 Evaluation Framework Application and Results

We have deployed our proposed evaluation framework and exploited it to validate the performance of our spatio-temporal personalization approach for mobile users [11]. The main objectives of the experimental evaluation are 1) showing the feasibility of our evaluation framework within a real testing scenario, 2) measuring the consistency of results using the two evaluation protocols. In what follows we first give an overview of our personalization approach, describe the framework evaluation in a real diary study and then present a comparative evaluation of the two protocols.

3.1 Our Approach for Personalizing Mobile Search Using a Spatio-Temporal User Profile

Here we give an overview of our approach for personalizing mobile search developed in our previous work [11]. It will serve as a testing scenario for our evaluation framework. Our personalization technique aims to adapt search results according to user's interests in a certain situation. A user U is represented by a set of situations with their corresponding user profiles (interests), denoted: $U = \{(S^i, G^i)\}$, where S^i is a situation and G^i its corresponding conceptual

graph user profile. A situation S^i refers to the geographical and temporal context of the user when submitting a query to the search engine. Each situation can be represented by an aggregation of four dimensions:

- Location type: refers to a class name (such as beach, school, ...) extracted from a classification category of location types (like ADL feature type thesaurus[1]),
- Season: refers to one of the year's seasons,
- Day of the week: refers either to workday, weekend or holiday,
- Time of the day: refers to time zone of the day: morning, midday, afternoon, evening and night.

User profiles are built over each identified situation by combining graph-based query profiles. A query profile G_q^s is built by exploiting clicked documents 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:

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

K^s is matched with each concept c_j of the ODP[2] ontology represented by single term vector $\overrightarrow{c_j}$ using the cosine similarity measure. The scores of the obtained concepts are propagated over the semantic links as explained in [12]. The user profile G_i, within each identified situation S^i, is initialized by the profile of the first query submitted by the user at the situation S^i. It is updated by combining it with the query profile G^* of a new query for the same situation as follows:

$$sw_{c^i}(c_j) = \begin{cases} \eta * sw_{c^i}(c_j) + (1-\eta) * sw_{c^*}(c_j) \\ \quad if\ c_j \in G^i \\ \\ \eta * sw_{c^*}(c_j)\ \ otherwise \end{cases} \tag{2}$$

where $sw_{c^i}(c_j)$ is the weight of concept c_j in the profile G^i and $sw_{c^*}(c_j)$ is the weight of concept c_j in the profile G^*. A case-based reasoning approach is adopted for selecting the most similar profile G^{opt} to use for personalization according to a new situation by exploiting a similarity measure between situations as explained in [11]. Personalization is achieved by re-ranking the search results of queries related to the same search situation. The search results are re-ranked by combining for each retrieved document d_k, the original score returned by the system $score_o(q^*, d_k)$ and a personalized score $score_c(d_k, G^{opt})$ obtaining a final $score_f(d_k)$ as follows:

$$score_f(d_k) = (1-\gamma) * score_o(q, d_k) + \gamma * score_c\left(d_k, G^{opt}\right) \tag{3}$$

[1] http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver100301/
[2] The Open Directory Project (ODP): http://www.dmoz.org

Where γ ranges from 0 to 1. The personalized score $score_c(d_k, G^{opt})$ is computed using the cosine similarity measure between the result d_k and the top ranked concepts of the user profile C^{opt} as follows:

$$score_c\left(d_k, G^{opt}\right) = \sum_{c_j \in C^{opt}} sw\left(c_j\right) * \cos\left(\vec{d_k}, \vec{c_j}\right) \quad (4)$$

Where $sw\left(c_j\right)$ is the similarity weight of the concept c_j in the user profile G^{opt}.

3.2 Evaluation Framework Application

We conducted a diary study, where users were asked to record the date, the time, the location, and the query they have while they are mobile (out of desk and home). Seven volunteers participated to our study (3 female and 4 male), ages ranged from 21 to 36. The diary study lasted for 4 weeks and it generated 79 diary entries, with an average of 11.28 entries per person. Table 1 shows an example of such diary entries, each diary entry represents a userid, date, time, place and the user query. From the diary study entries, we obtained a total of 79 queries expressed principally in the French language. Query length varies between 1 and 5, with an average of 2,99. The user intent behind these queries is mostly informational "*velo hauteur selle*" or transactional "*paris hotel cardinal*". From the diary study entries, we extract location and time information associated with each query. While the location information is already expressed in semantic concepts, the time entries are not. Thus, according to our personalization approach, we transformed each date time on a semantic period of the day or the week. We totally obtained 36 different situations, with an average of 5 different situations by user (min=2, max=12) and an average of 3 (min=1, max=8) queries within a same situation. We submit the total queries to Yahoo boss search API[3], and crawled the top 50 obtained results for each query. These documents are presented for relevance judgment to our diary study participants via an assessment tool available on line and developed in our lab[4]. The user interests are integrated in the evaluation protocol according to an automatic algorithm that generates them based on the users manual judgments of the documents like described in section 3.1.

Table 1. An example of some diary entries

User	Date	Hour	Place	Query
1	20-fvr	14h30	place de la concorde	"histoire obélisque"
2	27-fvr	11h10	périphérique	"parking relais bordeaux"
6	16-fvr	16h30	musée	"exposition beaubourg artistes"
7	02-mars	19h40	station bus	"tisseo horaire bus 2"

[3] http://developer.yahoo.com/search/boss/
[4] https://osirim.irit.fr developed at IRIT lab.

This first diary study allows us to verify the feasibility of our evaluation framework and its ability to provide as with the desired functionality. In what follows we present our experiment to test results consistency over the two evaluation protocols.

3.3 Measuring Results Consistency over the Two Evaluation Protocols

The goal here is to measure results consistency over the two proposed evaluation protocols. For this aim, we applied these latter for evaluating the effectiveness of our personalized approach. We mention here that the two protocols satisfy the minimum of 25 testing queries, and as it can be expected, the k-fold cross validation allows us to test more queries (68 against 29 for the training/test in chronological order protocol). We first study the effect of combining the original document's rank of Yahoo boss (corresponding to the original document score in formula 3) and the personalized document rank obtained according to our approach, on the retrieval effectiveness. Figure 2 (resp. Figure 3) shows the improvement of our personalized search in terms of P@10,P@20, nDCG@10 and nDCG@20 obtained when using the training/test in chronological order protocol (resp. when using the cross validation protocol) with varying the combination parameter γ in the interval [0 1]. Results show that the best performance is

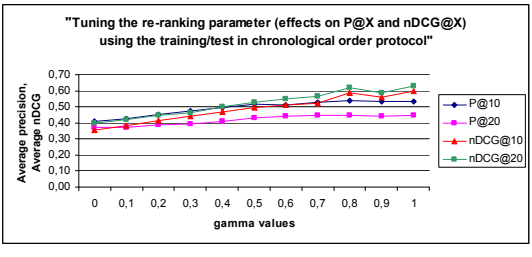

Fig. 2. Effect of the parameter gamma on Precision and nDCG in the combined ranks using the training/test in chronological order protocol

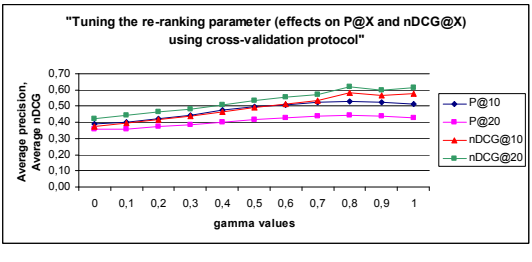

Fig. 3. Effect of the parameter gamma on Precision and nDCG in the combined ranks using the k-fold cross validation protocol

obtained when γ is between 0.8 and 1 for the two protocols. This is likely due to the fact that all the results on the top 50 match the query well and thus the distinguishing feature is how well they match the user profile.

In a second time, we compare our personalized retrieval effectiveness to the baseline search using the best γ value for each protocol. Table 2 shows the improvement of our personalized search in terms of P@10, P@20, nDCG@10 and nDCG@20 over the two protocols. Significant improvement are noted by * in table 2 according to a statistical t-test assuming the significance level fixed at $\alpha = 5\%$. Results prove that personalized search achieves higher retrieval precision of almost the queries. Moreover, our approach enhances the initial nDCG@10 and nDCG@20 obtained by the standard search and improve thus the quality of the top search results lists.

Table 2. Average Top-n precision and nDCG comparison between our personalized search and Yahoo boss over the two evaluation protocols

Evaluation protocol	System/ improvement	Average precision		Average nDCG	
		P@10	P@20	nDCG@10	nDCG@20
training/test in chronological order	Yahoo boss	0,41	0,37	0,35	0,40
	Our approach	0,53	0,45	0,59	0,63
	Improvement	31,14%*	20,72%*	67,65%*	58,80%*
k-fold cross validation	Yahoo boss	0,39	0,36	0,37	0,42
	Our approach	0,52	0,43	0,58	0,61
	Improvement	32,14%*	19,58%*	55,84%*	44,48%*

When comparing the two protocols results, we can observe that there is some difference in improvement of our approach over the two protocols. To determine whether or not an evaluation protocol might be better than another, we conducted a t-test. More precisely, we stated the null hypothesis (denoted H0) specifying that both evaluation protocols achieved similar performance levels, here evaluated between the means obtained on P@10, P@20, nDCG@10 and nDCG@20 over the common queries. This hypothesis would be rejected at the significance level fixed at $\alpha = 5\%$. We obtained a p-value of 0.434 for P@10, 0.478 for P@20, 0.387 for nDCG@10 and 0.365 for nDCG@20, wich are all greater than 0.05. We can then accept the null hypothesis and conclude that there is no significant difference between the two protocols.

4 Conclusion

In this poster we have presented a new evaluation framework for evaluating context-aware personalization techniques for mobile search. It is based on a diary study approach. More precisely, we exploit diary study entries to collect mobile queries, an API web search service and real user judgments to construct our ground truth, in context. We have deployed our proposed framework and exploit it for evaluating the search effectiveness of our personalized approach

comparatively to a standard search. We compared the two evaluation protocols training/test in chronological order and K-fold cross validation and showed the consistency of the obtained results. Our example application illustrates the feasibility and usefulness of our proposed evaluation framework. In future, we plan scaling our diary study to include more users and for more long time in order to collect more contextual search situations.

Acknowledgments

The authors acknowledge the support of the project QUAERO, directed by OSEO agency, France, and thank all persons who participated in the experiment.

References

1. Brown, P.J., Jones, G.J.F.: Context-aware retrieval: Exploring a new environment for information retrieval and information filtering. Personal Ubiquitous Computing 5(4), 253–263 (2001)
2. Tamine-Lechani, L., Boughanem, M., Daoud, M.: Evaluation of contextual information retrieval effectiveness: Overview of issues and research. KIS 24(1), 1–34 (2009)
3. Bouidghaghen, O., Tamine, L., Daoud, M., Laffaire, C.: Contextual evaluation of mobile search. In: Doan, B.-L., Jose, J., Melucci, M., Tamine, L. (eds.) Workshop on Contextual Information Access, Seeking and Retrieval Evaluation, Milton Keynes, CEUR Workshop Proceedings, vol. 569 (2010)
4. Mizzaro, S., Nazzi, E., Vassena, L.: Retrieval of context-aware applications on mobile devices: How to evaluate? In: Proceedings of IIix, pp. 65–71. ACM, NY (2008)
5. Göker, A., Myrhaug, H.: Evaluation of a mobile information system in context. Information Processing and Management 44(1), 39–65 (2008)
6. Borlund, P., Ingwersen, P.: Measures of relative relevance and ranked half-life. In: 21^{st} annual international ACM SIGIR conference on Research and development in information retrieval, pp. 324–331. ACM, NY (1998)
7. Mountain, D., MacFarlane, A.: Geographic information retrieval in a mobile environment: evaluating the needs of mobile individual. JIS 33(5), 515–530 (2007)
8. Rieman, J.: The diary study: a workplace-oriented research tool to guide laboratory efforts. In: INTERACT 1993 and CHI 1993 conference on Human factors in computing systems, pp. 321–326. ACM, NY (1993)
9. Sohn, T., Li, K.A., Griswold, W.G., Hollan, J.D.: A diary study of mobile information needs. In: CHI, pp. 433–442. ACM, New York (2008)
10. Voorhees, E.M.: The Philosophy of Information Retrieval Evaluation. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 355–370. Springer, Heidelberg (2002)
11. Bouidghaghen, O., Tamine-Lechani, L., Boughanem, M.: Dynamically personalizing search results for mobile users. In: Andreasen, T., Bulskov, H. (eds.) FQAS 2009. LNCS, vol. 5822, pp. 99–110. Springer, Heidelberg (2009)
12. Daoud, M., Tamine, L., Boughanem, M.: Towards a graph based user profile modeling for a session-based personalized search. KIS 21(3), 365–398 (2009)