Contextual evaluation of mobile search

Ourdia Bouidghaghen
IRIT, Paul Sabatier University
118, Route de Narbonne
Toulouse, France
bouidgha@irit.fr

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

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

Cécile Laffaire
IRIT, Paul Sabatier University
118, Route de Narbonne
Toulouse, France
laffaire@irit.fr

ABSTRACT

We discuss the issue of evaluating our context-based personalized mobile search approach with a methodology based on a combination of two evaluation approaches: context simulation and user study. Our personalized approach aims at exploiting some context-aware user profiles through a personalized score to re-rank initial search results obtained from a standard search system. We use Yahoo!'s open search web services platform BOSS [1] as a baseline. The context simulation allows us to simulate user locations and their related user interests. The user study involves real users who give their relevance judgments to the top 20 documents returned by yahoo and by our approach through an assessment tool available on the web platform OSIRIM[2]. The experimental results show the effectiveness of our personalized approach according to the proposed evaluation protocol.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Relevance feedback

Keywords

mobile search, context, user profile, evaluation protocol

1. INTRODUCTION

The proliferation of mobile technologies such as (PDAs and mobile phones, ...) and, with them, of mobile users, have moved the static world of classical and Web IR towards an always changing context-based world. The notion of context, roughly described as the situation the user is in, is exploited in the development of new IR systems. Starting from considering only a low number of contextual features (location, time and interests), such systems are faced to a new challenge for IR, that is how those contextual data can enhance user satisfaction. Another important issue is how to evaluate the strategies and techniques involved in these new systems. It is commonly accepted that the traditional evaluation methodologies used in TREC, CLEF and INEX campaigns are not always suitable for considering the contextual dimensions in the information access process. Indeed, laboratory-based or system oriented evaluation is challenged by the presence of contextual dimensions such as user profile or environment which significantly impact on the relevance judgments or usefulness ratings made by the end user [17]. To alleviate such limitations, contextual evaluation methodologies have been proposed to support simulated user profile through contextual simulations [16] or real evaluation scenarios through user studies [5].

As an initial approach, yet allowing meaningful observations, we present here, the evaluation protocol aiming to evaluate empirically the performance of a novel context-based personalized mobile search system. For this purpose, we compare the performance of retrieval: without personalization and with personalization. We compare our approach to the results obtained from yahoo BOSS web search service, which did not implement itself any personalization capability. This paper discusses the methodology adopted and presents the results obtained. We first briefly survey IR evaluation methodologies in mobile contexts (Sec. 2). We then presents our approach for mobile search personalization, and introduce our contextual IR evaluation protocol (Sect. 3). Finally, we conclude and give perspectives for future works.

2. EVALUATION OF IR IN MOBILE CONTEXTS

Context-awareness in mobile IR focuses on context models including user profiles and environmental data (time, location, near persons, device and networks). The state-of-the-art highlights that significative theoretical and technological progress has been achieved in this area over the last few years, encouraged by the growing interest to co-located human-human communications and large scale location-based applications ([10, 15]). In the development of an IR system for mobile environments, evaluation plays an important role, as it allows to measure the effectiveness of the system and to better understand problems from both the system and the

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

Appears in the Proceedings of The 2nd International Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (CIRSE 2010), March 28, 2010, Milton Keynes, UK.
http://www.irit.fr/CIRSE/

user interaction point of view. However, evaluation remains challenging because of the main following reasons ([4, 11]): 1) environmental data should be available and several usage scenarios should be evaluated across them, 2) evaluation, if present, concerns a specific application (eg.tourist guide), generalization to a wide range of information access applications is difficult. Both user-centered and benchmark evaluation approaches are adopted. However, as mobile IR systems are strictly related to users and their environment, the user-centered evaluation live (user studies [3, 14, 8]) or in laboratory (context-simulation framework [4, 9]) seem to be the most natural one. In [8] for example, a user-centered, iterative, and progressive evaluation has been adopted combining IR evaluation methods with human-computer interaction development techniques. The authors consider mainly the following guidelines: involve the right participants that are either current users or likely future; choose the right situations considering the different aspects of the environment; set relevant tasks that make participants seek information and are in accordance with situations that have been identified; use relevant evaluation approach and measures according to the different sub-goals (effectiveness, usability) within the overall objective evaluation. The main limitations introduced by user studies is that experiments are not repeatable and that they induce an extra costs. Within the mobile IR field, a benchmark evaluation has been used in [13, 12], they demonstrated the efficacy of the benchmark approach to evaluate an early stage of their system.

3. EVALUATION OF OUR CONTEXT-BASED PERSONALIZED SEARCH

In this section, we first introduce our context-based personalized approach for mobile search, we then present our evaluation protocol devoted for our proposed approach.

3.1 Situation-aware user profile

Our context-aware approach to personalize search results for mobile users [2] 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 user profile. A situation S^i refers to the geographical and/or temporal context of the user when submitting a query to the search engine. 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:

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

K^s is matched with each concept c_j of the ODP[3] 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 [6]. We select the most weighted graph of concepts to represent the query profile G_q^s at time s. The user profile G_i^0, 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_q^{s+1} of a new query for the same situation, submitted at time $s+1$. A case-based reasoning approach [1] is adopted for selecting a profile G^{opt} to use for personalization according to a new situation by exploiting a similarity measure between situations as explained in [2]. 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:

$$score_f\left(d_k\right) = \gamma * score_o\left(q^*, d_k\right) + (1-\gamma) * score_c\left(d_k, G^{opt}\right) \tag{2}$$

Where γ ranges from 0 to 1. Both personalized and original scores could be bounded by varying the values of γ. 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:

$$score_c\left(d_k, G^{opt}\right) = \sum_{c_j \in C^{opt}} sw\left(c_j\right) * \cos\left(\vec{d_k}, \vec{c_j}\right) \tag{3}$$

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

3.2 Evaluation of contextual personalization

In the absence of a standard evaluation framework, a formal evaluation of contextualization techniques may require a significant amount of extra feedback from users in order to measure how much better a retrieval system can perform with the proposed techniques than without them. In this case, the standard evaluation measures from the IR field require the availability of manual content ratings with respect to query relevance and specific user preference (i.e., constrained to the context of his search). For this aim we build a testbed consisting of a search space corpus, a set of queries, and a set of hypothetic context situations. A user study was conducted, participants were asked to provide ratings, in a blind test, for two retrieval scenarios: 1) top 20 documents returned by Yahoo BOSS, 2) top 20 documents returned by our personalized approach. In the following, we describe our experimental data sets and our evaluation protocol.

3.2.1 Contexts and Queries

Since the contextualization techniques are applied as the time goes, we have defined a set of *six* short use cases as part of the evaluation setup. Each use case is composed of a set of queries within a given geographical context, and a narrative describing the relevance of a document regarding a query and a geographical context. We have simulated a set of six geographical contexts defined by a location type (*zoo, music store, cinema, library, garden* and *museum*). We have created a set of totally *25* different queries, *5* queries belonging to each geographical context. Since mobile search queries are known to be short (and thus ambiguous), our queries are generally short (query length $\leq$ 3) and some of them are consequently ambiguous (eg. *jaguar*) and are tested within different geographical contexts (eg. the query *"water lilies"* is tested within the two contexts *"garden"* and *"museum"*), totalizing a number of *30* queries within the six contexts. Our goal was to verify whether the consideration

[3] The Open Directory Project (ODP): http://www.dmoz.org

of geographical contexts and user profiles can enhance the performance of the search engine to respond to such ambiguous queries. Table 1 gives an example of the use case of the context *museum*.

3.2.2 Document collection

The document collection consists of a set of about 3750 web pages retrieved from the web by yahoo BOSS as response to our set of queries. It is built by collecting the 150 first retrieved documents per query.

3.2.3 User profile

The user profiles are integrated in the evaluation strategy according to a simulation algorithm that generates them using hypothetic user interactions for each query. They are constructed based on a manual judgments of the <query, narrative, document> tuples for all the document in the collection. These, so built profiles, simulate user click-through data.

3.2.4 Evaluation protocol

Our experimental design consists of evaluating the effectiveness of our personalized approach when using the user profile in the IR model over a sequence of user contexts. In the absence of an initial score of the document results list of yahoo BOSS, the re-ranking procedure is done based only in the personalized score (ie. $\gamma = 0$ in equation 2). The evaluation scenario is based on the k-fold cross validation like in [7] explained as follows:

- for each use case, divide the query set into k equally-sized subsets, and using $k-1$ training subsets for learning the user interests and the remaining subset as a test set,
- for each query in the training set, an automatic process generates the associated profile based on its top n relevant documents listed in the manually constructed relevance judgments file.
- update the user profile concept weights across the queries in the training set and use it for re-ranking the search results of the queries in the test set.

In order to evaluate the performance of our proposed approach, a user study is conducted to compare the 20 top ranking output of our approach and of Yahoo BOSS. Using an assessment tool available on the web platform OSIRIM, *six* users who participated to the experiment were asked to judge each tuple <query, document, narrative> within the 20 top ranking output of both our approach and of Yahoo BOSS. Participants were unaware of the system they judge. Relevance judgments have been made using a three level relevance scale: relevant, partially relevant, or not relevant.

3.3 Results and Discussion

We evaluate the effectiveness of the personalized search over the six use cases and we compare the obtained results to the initial ones from Yahoo BOSS. To better estimate the quality of the search results at the top of the ranked list (since mobile users are unlikely to scroll long lists of retrieved items), we estimate the DCG@10 for all the queries.

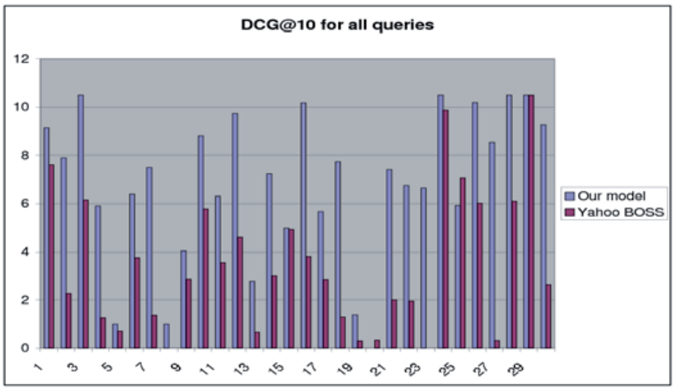

Figure 1: DCG@10 comparison between our personalized search and Yahoo BOSS over all queries

Table 2: Average Top-n precision comparison between our personalized search and Yahoo BOSS over all queries

	Average precision over all queries at:			
	P@5	P@10	P@15	P@20
Yahoo BOSS	0,37	0,39	0,38	0,36
Our model	0,70	0,64	0,59	0,55
Improvement	**87,50%**	**63,56%**	**53,49%**	**50,92%**

Figure 1 compares the effectiveness obtained by the initial yahoo search lists and the re-ranked ones obtained by our approach over all the queries. We observe that in general, our approach enhances the initial DCG@10 obtained by the standard search and improve the quality of the top search results lists. We have also computed the percentage of improvement of personalized search comparatively to the standard search computed at different cut-off points P@5, P@10, P@15 and P@20 averaged over all the queries. Results are presented in Table 2. Results prove that personalized search achieves higher retrieval precision of almost the queries in the six simulated contexts. Best performance are achieved by the personalized search in terms of average precision at different cut-off points achieving an improvement of 87,50% at P@5, 63,56% at P@10, 53,49% at P@15 and 50,92% at P@20 comparatively to Yahoo BOSS. However, precision improvement varies between queries, Figure 2 gives an example of this improvement variation between the queries of the context *museum*. This is probably due to the difference between the degree of ambiguity of the queries, which can not be explained only by the difference in query length. In fact, it depends also on the contents of the documents present in the collection.

4. CONCLUSION

In this paper we have presented our evaluation protocol of a context-aware personalization approach for mobile search. It is based on a combination of context simulation and user study. More precisely, we exploit context simulation to create user contexts and profiles in one hand. On the other hand, we exploit Yahoo's BOSS web search service and real user judgments, through a user study, to evaluate the search effectiveness of our approach comparatively to a standard search. We evaluated our approach according to the pro-

Table 1: an example of the use case *"museum"*

Context	QueryID	Query terms	Narrative
museum	M17	da Vinci	A document is relevant if it speaks about *da Vinci* painter and or his paintings
	M23	sunflowers	A document is relevant if it speaks about the painting *sunflowers* and or its painter *Van Gogh* and or his paintings
	M24	woman with a parasol	A document is relevant if it speaks about the painting *woman with a parasol* and or its painter *Claude Monet* and or his paintings
	M25	Edgar Degas	A document is relevant if it speaks about painter *Edgar Degas* and or his paintings
	M21	water lilies	A document is relevant if it speaks about the painting *water lilies* and or its painter *Claude Monet* and or his paintings

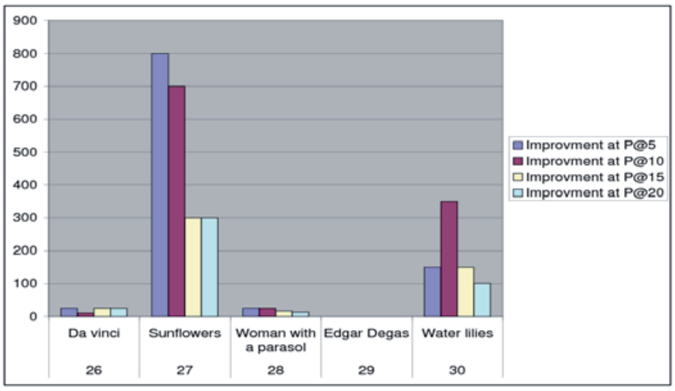

Figure 2: Improvement at P@5, P@10, P@15 and P@20 for the queries of the context "museum"

posed evaluation protocol and show that it 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 real mobile search contexts and click-through data available in the log file.

5. ACKNOWLEDGMENTS

The authors acknowledge the support of the project QUAERO, directed by OSEO agency, France, and thank PhD students at IRIT for their participation in the experiment.

6. REFERENCES

[1] A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1), 1994.

[2] O. Bouidghaghen, L. Tamine-Lechani, and M. Boughanem. Dynamically personalizing search results for mobile users. In *Proc. of Flexible Query Answering Systems*, pages 293–298, 2009.

[3] N. O. Bouvin, B. G. Christensen, K. Grønbæk, and F. A. Hansen. Hycon: a framework for context-aware mobile hypermedia. *Hypermedia*, 9(1):59–88, 2003.

[4] M. Bylund and F. Espinoza. Testing and demonstrating context-aware services with quake iii arena. *Communications of the ACM*, 45(1), 2002.

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

[6] 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)*, pages 1031–1035, 2009.

[7] M. Daoud, L. Tamine-Lechani, and M. Boughanem. Using a concept-based user context for search personalization. In *Proc. of the 2008 Internat. Conf. of Data Mining and Knowledge Engineering*, 2008.

[8] A. Göker and H. I. Myrhaug. Evaluation of a mobile information system in context. *Information Processing and Management*, 44(1):39–65, 2008.

[9] F. Gui, M. Adjouadi, and N. Rishe. A contextualized and personalized approach for mobile search. In *2009 Internat. Conf. on Advanced Information Networking and Applications Workshops*, pages 966–971.

[10] R. Iqbal, J. Sturm, O. Kulyk, J. Wang, and J. Terken. User-centred design and evaluation of ubiquitous services. In *Proc. of the 23rd annual internat. conf. on Design of communication*, pages 138–145, 2005.

[11] J. Kjeldskov and C. Graham. A review of mobile hci research method. In *Human-Computer Interaction with Mobile Devices and Services-5th Internat. Symposium, Mobile HCI 2003 proceedings*, 2003.

[12] D. Menegon, S. Mizzaro, E. Nazzi, and L. Vassena. Benchmark evaluation of context-aware web search. In *Proc. of ECIR 2009 Workshop on Contextual Information Access, Seeking and Retrieval Evaluation.*

[13] S. Mizzaro, E. Nazzi, and L. Vassena. Retrieval of context-aware applications on mobile devices: how to evaluate? In *Proc. of IIiX'08*, pages 65–71, 2008.

[14] C. Panayiotou, M. Andreou, G. Samaras, and A. Pitsillides. Time based personalization for the moving user. In *Proc. of the International Conference on Mobile Business (ICMB'05)*, pages 128–136, 2005.

[15] W. Schwinger, C. Grün, B. Pröll, W. Retschitzegger, and A. Schauerhuber. *Context-awarness in mobile tourism guides- a comprehensive survey.* Technical Report,Johannes Kepler University Linz, IFS/TK, 2005.

[16] A. Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. In *Proc. of the 16th ACM conference on information and knowledge management*, pages 525–534, 2007.

[17] L. Tamine-Lechani, M. Boughanem, and M. Daoud. Evaluation of contextual information retrieval effectiveness: Overview of issues and research. *Knowledge and Information Systems, Springer*, 2009.