Towards Zero-Click Mobile IR Evaluation: Knowing What and Knowing When

Tetsuya Sakai
Microsoft Research Asia, Beijing, P.R.C.
tetsuyasakai@acm.org

ABSTRACT

In this poster, we propose two evaluation tasks for mobile information access. The first task evaluates the system's ability to guess what the user's query should be given a context ("Knowing What"). The second task evaluates the system's ability to decide when to proactively deploy a given query ("Knowing When"). We conduct a preliminary manual analysis of a mobile query log to limit the space of possible queries so as to design feasible and practical evaluation tasks.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation

Keywords

evaluation, metric, mobile, query log, test collection

1. INTRODUCTION

In February 2012, forty-five IR researchers gathered in Lorne, Australia, to discuss important research areas in IR for the near future. This SWIRL (Strategic Workshop on Information Retrieval in Lorne) workshop[1] identified six important topics, three of which were "Mobile IR", "Zero Query", and "Querying by Walking Around". The latter two were basically the same idea proposed by different discussion groups: in a mobile environment where the user does not (or cannot) necessarily issue a query, can the system provide the user with the right information at the right time? Even though similar ideas have existed for many years, the IR community still lacks an established evaluation methodology for them.

In response to the SWIRL outcomes, we propose a few concrete ideas for evaluating mobile information access. More specifically, we propose two tasks: *Knowing What*, which evaluates the system's ability to guess what the user's query should be given a context, and *Knowing When*, which evaluates the system's ability to decide when to proactively deploy a given query. We conduct a preliminary manual analysis of a mobile query log to limit the space of possible queries so as to design feasible and practical evaluation tasks.

[1] http://www.cs.rmit.edu.au/swirl12/index.php

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

2. RELATED WORK

Proactive IR is not a new idea. For example, over a decade ago, Rhodes and Maes [4] described *Just-in-Time IR Agents*, "that proactively present information based on a person's context in an easily accessible and non-intrusive manner." In the context of mobile IR, Jones and Brown [2] described *Context-Aware Retrieval*, where the system can be "interactive" (waits for the user to request information) or "proactive" (information to be searched contains a *triggering condition*, and is retrieved automatically when the user's context satisfies the condition). For evaluating such systems, they mention the traditional approach of precision and recall.

More recently, Coppola *et al.* [1] described the *Context-Aware Browser* for mobile IR, where the trigger for the search engine may be events such as "User arrives at Heathrow." Moreover, they built a test collection with ten topics, which resembles TREC ad hoc test collections: the key difference is that each topic contains a short text that describes the context of the user rather than an explicit query. As in traditional IR, nDCG was used for evaluating ranked lists of web pages. The TREC 2012 Context Suggestion Track[2] is currently underway to pursue similar ideas.

Sakai, Kato and Song [5] proposed an information access evaluation framework primarily designed for mobile environments. In their *One Click Access* task, systems are required to return a short multi-document summary instead of a ranked list of web pages. The task is called "One Click" because it aims to satisfy the user immediately after her click on the SEARCH button.

In the present study, we go one step further and address the problem of *Zero Click Access*: instead of waiting for the user to enter a query and click SEARCH, the system decides when and what to provide to the user. In the aforementioned SWIRL workshop, "Zero Query" referred to situations where the user does not provide a query but clicks SEARCH; "Less than Zero" referred to proactive cases where the system does not wait for a search button click.

3. PROPOSED TASKS

We break down the Zero Click Acces problem and propose two tasks: *Knowing What*, which evaluates the system's ability to guess what the user's query should be given a context, and *Knowing When*, which evaluates the system's ability to decide when to proactively deploy a given query. The key here is to isolate these two problems from the problem of finding relevant information for a given a query.

For evaluating Knowing What and Knowing When, we propose to use real mobile query log data, which is of the form:

```
<timestamp> <query> <context>.
```

[2] http://sites.google.com/site/treccontext/

Here, we define *context* generically as "any piece of information besides a query string that may help identify the user's information need." This could include the user's profile, location, sensor data that reflect the objects and the environment that surround the user, mobile apps the user is using and so on. Clearly, several obstacles exist if we are to release such data to the research community, but here we assume that they can be overcome.

An alternative approach to evaluating mobile IR would be to utilise *clicked URLs* in mobile query logs. However, a study on desktop and mobile query logs by Li, Huffman and Tokuda [3] showed that users are often satisfied without actually clicking and visiting a particular URL: the search engine result page itself suffices for obtaining the desired information; the aforementioned Once Click Access aims exactly at freeing the user from the burden of clicking several times. Moreover, a clicked URL may only partially reflect the user's information need (if relevant at all).

Below, we define two new tasks for Zero-Click mobile IR. using real mobile query logs obtained in the above format.

3.1 Knowing What

Knowing What evaluates the system's ability to guess what the user's query should be given a context. The task is defined as: *Given a mobile query log whose timestamps span* $[1, T]$ *where all queries in the* $[t, T]$ *range have been masked, and a context for timestamp* t' *(*$t \leq t' \leq T$*), recover the query string at timestamp* t'. Thus, the system is required to guess the query by observing previous data and the current context. A natural evaluation metric for this task would be *accuracy* (whether each predicted query string is right or wrong), or perhaps *weighted accuracy* which allows partial matches between the true and predicted queries.

If we allow arbitrary queries, this task would be very challenging. We believe that a feasible and practical task setting would be to impose some restrictions on the type of queries and thereby limit the possible space of queries. We shall discuss this in Section 4.

3.2 Knowing When

Knowing When evaluates the system's ability to decide when to proactively deploy a given query. Using the same mobile query log data, this task can be defined as: *Given a mobile query log whose timestamps span* $[1, T]$ *where all queries in the* $[t, T]$ *range have been masked, and a query string sampled from* $[t, T]$*, recover the query's timestamp (i.e. when it was issued by the user).* Thus, if a system is effective in both Knowing What and Knowing When, it should be able to automatically issue an appropriate query for the mobile user at the right moment.

The output of a Knowing When system is a timestamp (or a set of timestamps). Thus, suppose the absolute difference between the true and the predicted timestamps is d. Then, for an allowance parameter Δ, the *timestamp accuracy* could be computed for example as $1 - \min(1, d/\Delta)$. It is one when the predicted timestamp is perfectly accurate; it is zero if the difference exceeds Δ; the system is partially rewarded otherwise for being "close."

4. INITIAL QUERY LOG ANALYSIS

For both Knowing What and Knowing When, we probably need to limit the space of possible queries for designing feasible and practical tasks. For this purpose, we manually examined 5,000 most frequent mobile queries (approximately 472 million queries by volume) sampled between November and December 2011 from the US English market, and manually classified them into major query types. These query types were defined in a bottom-up manner and were revised repeatedly as the author carefully went through the queries.

Table 1: Query types for 5,000 head mobile queries.

type	example	%unique	%volume
WEBSITE	`facebook`	10.74	30.25
VIDEO	`youtube`	17.98	15.11
MUSIC	`romantic song`	19.03	10.99
PERSON	`Obama`	12.35	10.34
GEO	`chicago, IL`	4.92	3.92
SHOP	`starbucks`	5.99	2.99
SPORTS	`nfl scores`	2.59	1.56
ANIMAL	`squirrels animal`	1.54	1.05
WEATHER	`weather`	0.42	0.93
FOOD	`pizza`	0.46	0.63
FACILITY	`Beijing The Egg`	0.66	0.50
MAP	`mapquest`	0.51	0.45
GAME	`games`	0.93	0.40
IMAGE	`funny pictures`	0.56	0.36
NEWS	`msnbc`	0.81	0.32
PRODUCT	`Windows Phone 7`	0.49	0.16
BANKING	`wellsfargo.com`	0.32	0.16
AIRLINE	`southwest airlines`	0.29	0.12
STOCK	`quote GOOG`	0.24	0.10
OTHER	-	19.15	19.66

Table 1 shows the query type distribution after removing pornographic queries (18% of the unique queries; 14% by volume). It can be observed that WEBSITE, VIDEO, MUSIC and PERSON query types constitute about 66% of the head queries by volume. In particular, most of the WEBSITE queries are homepage finding queries such as `facebook` and `bing`. Thus, even though previous work in mobile IR evaluation discussed ranking webpages and evaluating with nDCG etc., perhaps a more important question for mobile users is whether the system can provide the one correct URL to the user at the right time. As a first step to evaluating Knowing What and Knowing When, limiting the query space to homepage finding queries is probably feasible and practical.

Leveraging query types such as the ones shown in Table 1 for *evaluating* Knowing What may also be a good idea, as this may separate the task of collaborative recommendation from that of personalisation. For example, suppose that a "Knowing When and What" system learns from the training data that a certain group of users tend to issue MUSIC queries right after they get off a train. But just because User X likes `romantic songs` (Table 1) does not necessarily imply that User Y, who behaved similarly to X, wants to listen to the same kind of music. Personalisation can be used orthogonally to this Knowing When and What system, so that (say) `techno beats` can be provided to Y.

5. SUMMARY

We defined mobile IR tasks called Knowing What and Knowing When, and identified popular query types in a mobile query log to make the tasks both feasible and practical. We hope to deploy a pilot task at evaluation venues such as NTCIR.

6. REFERENCES

[1] P. Coppola, V. D. Mea, L. D. Gaspero, D. Menegon, D. Mischis, S. Mizzaro, I. Scagnetto, and L. Vassena. The context-aware browser. *IEEE Intelligent Systems*, 25(1):38–47, 2010.

[2] G. J. F. Jones and P. J. Brown. Context-aware retrieval for ubiquitous computing environments. In *Mobile and Ubiquituos Information Access (LNCS 2954)*, pages 227–243, 2004.

[3] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and PC internet search. In *Proceedings of ACM SIGIR 2009*, pages 43–50, 2009.

[4] B. J. Rhodes and P. Maes. Just-in-time information retrieval agents. *IBM Systems Journal*, 39(685-704), 2000.

[5] T. Sakai, M. P. Kato, and Y.-I. Song. Click the search button and be happy: Evaluating direct and immediate information access. In *Proceedings of ACM CIKM 2011*, pages 621–630, 2011.