Out and About on Museums Night: Investigating Mobile Search Behaviour for Leisure Events

Richard Schaller
Computer Science (i8)
Uni of Erlangen-Nuremberg
richard.schaller@cs.fau.de

Morgan Harvey
Computer Science (i8)
Uni of Erlangen-Nuremberg
morgan.harvey@cs.fau.de

David Elsweiler
I:IMSK
University of Regensburg
david@elsweiler.co.uk

ABSTRACT

When search behaviour is studied in information retrieval it is nearly always studied with respect to work tasks. Recent research, however, has indicated that search tasks people perform in leisure situations can be quite different. In leisure contexts needs tend to be more hedonistic in nature and often don't require specific information to be found. Instead, information is sought that can lead to a specific emotional or physical response from the user, such as feelings of being stimulated or entertained. In this paper we investigate how people behave to meet such needs in one particular leisure context. We analyse search log data collected from a large-scale (n=391), naturalistic study of behavior with a mobile search tool designed to help people find events of interest to them at the Long Night of Museums, Munich. We examine the queries submitted, establish performance metrics and investigate how spoken queries differ from those typed via the keyboard on a mobile device. The findings provide insight into how users behave in one specific casual-leisure context and lead to several open questions for future research.

1. INTRODUCTION AND MOTIVATION

Search behaviour has traditionally been studied in the context of people completing work tasks. Despite its name, a work task need not be work-related. It is simply a sequence of activities a person has to perform in order to accomplish a goal [8]. A work task has a recognisable beginning and end, it may consist of a series of sub-tasks, and results in a meaningful product [3]. Correspondingly, the models we have of information seeking behaviour tend to assume that people look for information in response to a lack of understanding or the recognition of a gap in knowledge [2] preventing the completion of the task at hand.

Based on two investigative studies, one examining information needs in the context of television viewing and the other analysing broader information behaviour reported on twitter, Elsweiler and colleagues [7] proposed a model for what they refer to as casual leisure search, which deviates from standard work-based models. According to their model, in casual-leisure situations users seek information not in response to a knowledge gap, but with the aim of being entertained or passing time. Such needs tend to be directly related to mood, physical state or the surrounding social context. A further defining characteristic of such needs is that the informational content found by users is often less important than the feelings induced by the found content and/or the search process itself.

Beyond these two studies, very little literature explicitly focuses on information seeking behaviour in casual-leisure situations. Exceptions include studies of finding fiction [12] and non-goal oriented newspaper reading [14]. To our knowledge no other naturalistic studies of information behaviour in casual-leisure contexts exist. We believe that transactional studies, such as those that have provided a rich understanding of web search behaviour [9] would be particularly beneficial, as they would provide concrete insight into how people behave to resolve such needs. If the model proposed by Elsweiler et al. is correct and people do not care what information content is about, but rather are concerned primarily with the emotional or physical response to such content then what do queries in casual-leisure situations look like? What do people try to describe with queries and how much effort do they expend in doing this? Are queries long and descriptive and are users willing to look through lots of results to find something suitable?

In this paper we describe a study designed to answer these kinds of questions. We report analyses of interaction logs for a search system supporting one specific leisure situation - the Long Night of Munich Museums, 2011. While we do not claim that the logs are representative of all casual-leisure search behaviour, they do provide an insight into how users behave in one specific casual-leisure context and a situation where the user has a high-level, hedonistic goal. Our findings represent a good starting point from which to investigate search behaviour more generally in casual-leisure situations.

2. DISTRIBUTED EVENTS

A distributed event is a collection of single events occurring at approximately the same time and conforming to the same general theme. One such event is the Long Night of Munich Museums (Lange Nacht der Münchner Museen), an annual cultural event organised in the city of Munich, Germany[1]. In addition to a diverse range of small and large museums, other cultural venues, such as the Hofbräuhaus and the botanical garden open their doors during one evening in October. Many venues organise special activities and exhibitions not otherwise available.

Visitors to the Long Night include both locals and tourists and represent a broad range of age groups and social backgrounds. In 2011 an estimated 20,000 people visited a total of 176 events at 91 distinct locations, including exhibitions, galleries and interactive events. Events take place all over the city, mostly in the city centre, but some, such as the Mu-

[1]The event is organised by Münchner Kultur GmbH (http://www.muenchner.de/museumsnacht/)

seum of the MTU Aero Engines and the Potato Museum, are located in suburbs. Special bus tours are set up to transport visitors between events.

From interviews (n=25) we conducted with people attending the evening we know that on average each visitor attends 4 events meaning that approximately 80,000 visits took place in 2011. The standard way to discover events on offer is to use the booklet that is distributed for free by the organisers and contains descriptions of all events in the order they lie along the bus tours. This booklet is necessarily large (110 A6 pages) and can be difficult to navigate.

Only a few of our interviewees reported having specific events they would like to visit. Instead, most described having the same kinds of high-level, hedonistic needs as reported in the literature [6, 15]. i.e. "to have a pleasant evening", "to enjoy time with friends", "to extend or diversify their general knowledge" etc. We will report on the interview results in detail in a future publication, but the findings seem to substantiate Elsweiler et al.'s model.

Here we want to establish how visitors to the Long Night of Museums query a search system to address these kinds of needs. We also want to know how successful they are, and identify noteworthy behaviours, problems and any potential solutions. The long-term goals of our work are to learn about behaviour in order to understand how to build better search tools and to augment existing theoretical models of casual-leisure search. We present the results of initial analyses that lead to more detailed future research questions.

3. SYSTEM

An Android app was developed to help visitors of the Long Night find events of interest to them personally. Once they have found and indicated the events they would most like to visit, the system can create a time plan for the evening, taking into account constraints such as start and end times of events, time to travel between events and public transport routes and schedules. If the user chooses more events than would fit into the available time[2], then the system tries to maximise the number of scheduled events by leaving out those that require long travel time. It is also possible for the user to manually customise the plans by adding, removing and re-ordering events to be visited. Based on the created plan, the application can lead the user between chosen events using a map display and textual instructions. Figure 1 provides some screenshots of the app[3].

The user has four ways to find events he would like to visit, namely he can: Browse events by bus route; browse events by event type (e.g. exhibitions, guided tours, interactive event, etc.); submit free-text queries, which search over the names and descriptions of the events; receive recommendations based on a pre-defined profile and collaborative filtering algorithm built into the app.

In this paper, in line with the research aims as outlined above, we focus on the way the search features were used. The search functionality was implemented in Lucene[4] and documents were represented by titles and descriptions from the Long Night booklet. Based on interviews conducted, we expected visitors to search for topics or for other high level needs not accessible for a full text search. Therefore we extended Lucene to perform a search based on topics. In a first step the event descriptions and titles were tokenised and stemmed. To match topically similar words we then map every token to one or more topic groups (these groups are taken from [4]). This way terms such as "dinner" and "food" are mapped to the same groups, thus event descriptions containing one of these words could be found by the other. To speed up interaction with the system, queries were submitted after each typed character (search-as-you-type). The presented result list contains the name and nearest bus stop for each of the retrieved events.

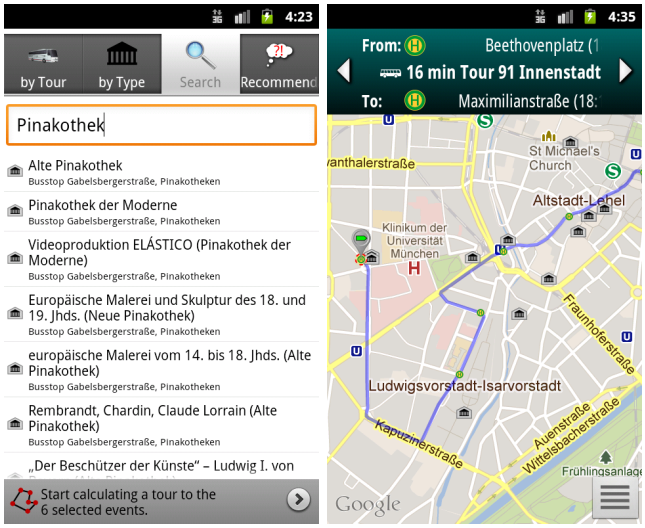

Figure 1: The search screen with a query (left) and the map screen with the planned route (right)

4. METHOD

We examined user search behaviour by recording user interactions with our app at the 2011 Long Night. The app was available for download from the Android Market and advertised on the official Long Night of Museums web page. In total the application was downloaded approximately 500 times and 391 users allowed us to record their interaction data. We recorded all interactions with the application including submitted queries, result click-throughs, all interactions with browsing and recommendation interfaces, tours generated, modifications to tours, as well as all ratings submitted for events. Users interacted on average for 45.26 minutes[5] with the system (median 19.31). 80.1% of users interacted for more than 5; 38.4% for more than 30.

A short questionnaire provided us with demographic information. 51% of the app users were first-time visitors to the Long Night of Museums, 22% were second-time visitors and 27% had attended more than twice previously. 4% of users were 17 years of age or younger, 39% were between 18 and 29, 30% 30-39, 18% 40-49, 8% 50-59 and 1% above 60 years old. These demographics are very similar to those reported by event organisers for previous Long Nights [1] suggesting that our sample of users should reflect well the visitors as a whole. Comparing both age distributions with Fisher's exact test reveals a p-value of 0.29; thus it is highly

[2]most events are open between 7pm and 2am

[3]a video demo of the application can be found on YouTube (http://www.youtube.com/watch?v=woVjpivxtMc)

[4]Lucene version 3.1. (http://lucene.apache.org)

[5]discounting times where no user interaction was recorded for more than 15 seconds

unlikely that the counts are drawn from different underlying distributions.

Since queries were submitted after every typed character, it was necessary to pre-process the recorded queries to establish those that the users actually intended to submit. For example, if the user wanted to search for "food", the system logged "f", "fo", "foo", as well as "food". Furthermore, should the user wish to submit a new query, then he must first remove the old search terms from the search box, resulting again in all prefixes but this time in decreasing length.

Automatically extracting the intended query proved difficult due to spelling errors and automatic correction. We therefore manually judged queries to be intended or not. 3 assessors separately annotated all of the approx. 10,000 queries logged as being either intended or not-intended. A high inter-assessor agreement was found (Fleiss' kappa = 0.872, 86.2% of queries which were labeled by at least 1 assessor were also labelled by at least one other assessor). This process resulted in a final list of 801 search queries, which is used in the following analyses.

5. QUERY CHARACTERISTICS

Overall the search queries were short, having a mean length of 1.21 terms ($\sigma = 0.52$) and 8.9 characters ($\sigma = 5.31$). These values are much shorter than those reported for similar mobile-like devices for web search. [10] report lengths of 2.3 terms for older mobile phones and new research suggests even longer queries (2.9 terms and 18.25 characters) for modern phones similar to those used in our study [11].

It was very apparent while analysing the queries that many represented searches for named entities, in particular the names of specific museums. Again 3 human assessors were asked to assign queries into categories: specific event name, not a specific event name or indeterminate. The third category was necessary as some queries were short and it was not possible to definitively claim that the term referred to a specific event. For example "deutsches" is likely to be a reference to the "deutsches Museum" but it is not possible to say for certain. For 87.3% of all queries at least two of the assessors were able to agree on one of the three categories (Fleiss Kappa of 0.43).

59.4% of the agreed on queries were marked as clearly named entities and 34.6% that might be named entities. Only 6.0% were labeled as non named entity searches. These remaining searches were often queries for non-museum locations, e.g. 18.2% of these are names of bus stops.

Notably absent from the logs were queries describing topical content of events e.g. "art history", "engineering", "modern art", etc. There were also no queries referring to properties of events e.g. "interactive", "talks", "discussions" and no evidence of high-level, hedonistic qualities an event might bring about e.g. "fun", "exciting", "entertainment", etc.

In line with previous query analysis papers, we analysed the diversity of submitted queries. The cleaned query set contained 417 unique queries. As expected the distribution looks rather Zipf-like with the top 2 queries being "deutsches" and "deutsches Museum". The top 50 unique queries amount to 43.1% of all queries, the top 10 amount to 16.6% and the most common search term was used in 2.5% of all searches. The entropy of the unique search terms is 2.44 bits. The queries submitted were therefore far less diverse than web search queries on desktop or mobile devices. This can be partially explained by the fact that our collection is much smaller and much more specific than the web. Another explanation for the more homogenous queries is the fact that most queries are event names which are usually only one or two words long. This reduces the possibilities for searching for these names when compared with the possibilities to express interest, constraints or needs in general.

In summary, our main observation is that the queries submitted to the search system did not reflect the information needs described in the pre-study interviews. It seems as if the users did not use the search engine to discover new events, but rather used the feature to filter to events they already knew existed. Reflecting this, our queries have similar properties to those reported for known-item searches in web, email and desktop search, which have also been shown to be very short and contain a high percentage of named-entities [5, 13].

6. QUERY PERFORMANCE

We wanted to understand how successful queries were. With this in mind we defined three success metrics based on the user's interaction with search results. The first refers to whether the user selected a returned result to read a detailed description of the event. This metric is our equivalent to click-through data. 58.4% of all searches resulted in a click-through with an average of 0.73 clicks per query ($\sigma = 0.93$) and 5.95 results on average ($\sigma = 9.10$). We didn't consider good abandonment since the result list contains no information beyond name and nearest bus stop.

Two further, more explicit, definitions of success were if the user marked a returned event as a candidate for tour inclusion (38.0% of all searches) or the user added the event to an preexisting tour (15.6% of all searches). These searches were performed at different stages of application use. Reflecting this we derived a general success metric: in 59.7% of all searches at least one of these three actions was performed. Of the remaining 40.3% unsuccessful queries 59.8% were using a search term which resulted in an empty result list, in most cases a miss-spelled or only partial written named entity. The huge number of spelling errors underlines the need for fuzzy search methods in this application context.

As the queries that were submitted were very short, we wanted to investigate if the length of the query had any impact on the success of the search. Searches defined as successful were on average longer with a mean of 1.26 terms ($\sigma = 0.57$) compared to unsuccessful searches with a mean of 1.13 terms ($\sigma = 0.42$); a highly significant difference ($p \ll 0.01$). Likewise the number of characters per query was significantly ($p \ll 0.01$) longer with the successful searches having on average 9.90 characters ($\sigma = 5.42$) and the unsuccessful searches having just 7.47 characters ($\sigma = 4.80$). We implemented a search-as-you-type system which searches for whole words, however the evidence suggests that users used the system as a means to filter to events they already knew about. Therefore while entering the search term the result list is empty till you entered the complete word. This might have led users to the conclusion that their queries will be unsuccessful and abandon the search early. This would be one explanation for the shorter query length in unsuccessful searches.

7. TYPED VS SPOKEN QUERIES

An additional feature our app offers is the possibility to submit spoken queries. Rather than typing search terms

in using the keyboard, the user speaks the query into the phone. The system uses Google Speech Recognition to identify the query terms and the user selects the queries based on a list. This is familiar to android users as it is a standard feature for web search on Android phones. We wanted to establish how this feature was used, if queries submitted in this way differed from typed queries and whether there was a notable difference in performance between spoken and typed queries.

In total 22 app users submitted 68 spoken queries, which equates to 8.5% of all search queries. Of these 6 users used it more than three times. When comparing the length of the search queries we discovered that voice searches tend to be considerably longer than typed searches: 1.8 ($\sigma = 0.65$) vs. 1.2 ($\sigma = 0.46$) terms and 14.9 ($\sigma = 8.1$) vs. 8.4 ($\sigma = 4.6$) characters. Both comparisons[6] are significant ($p \ll 0.01$). It seems it is easier to create long queries with the voice interface than typing. The success rate is also significantly higher: 75% success for speech queries compared to 58.3% (p-value[7]: 0.01) success for typed queries.

It could be that the complicated input method when typing combined with the expectation of a filtering system might have tempted people to give up early, whereas spoken queries are always full words. This would explain the ratio of empty result list where 11.8% of the voice searches have an empty result list compared to 25.2% of non-voice searches; a difference which is significant (p-value[7]: 0.013). In summary, there is evidence to suggest that voice search can be an effective tool for entering search queries on a mobile device in leisure situations. There are, however, issues such as background noise and user self-consciousness that may explain why only a limited set of users used this functionality.

8. DISCUSSION AND CONCLUSIONS

In this paper we analysed the query behaviour of users in a specific casual-leisure situation: a mobile application to assist users at a distributed event. It was apparent when analysing the queries that there was a mismatch between the queries people submitted to the search system and what we anticipated based on the needs reported in the interviews. The overwhelming majority of queries were partial or complete event names, where the user was trying to filter to a specific event. There were very few queries relating to topics that the user may be interested in e.g. "art", "history", etc. Furthermore there were no references to descriptors of events that people noted they wanted in interviews e.g. "interactive", "talks", "discussions". Likewise there was no evidence of the high-level, hedonistic qualities an event might bring about e.g. "fun", "entertainment", etc.

This poses the question: why are people using the search system in this way? Are people conditioned to do so, i.e. do they have a preconceived notion about how search engines work and only use the system in ways that reflects this? Or is it because the app has other features, such as browsing by tour or genre that might be better suited for tasks other than known-item search? To answer these questions we are currently analysing the log data for the other features of the system. A comparison with other casual-leisure search would also complement our understanding of this issue. Are there similar trends for search on YouTube, Wikipedia or the web?

[6]Wilcoxon sign rank test

[7]Two-Tailed Test of Population Proportion

Our analysis of query performance showed that a high number of spelling mistakes were made. We wonder if this is caused by environmental factors, e.g. typing on a bumpy bus or if it is caused by a high number of named entities, the spelling of which people are not familiar? Further research would be needed to differentiate between the two, however a fuzzy search feature would certainly help people who struggle with the query input. A grep-style search would further reduce this problem since users would only need to enter a few characters as opposed to whole terms. In the comparison of spoken vs. typed queries we have seen that although not used much it provides a more successful way of querying the system.

We also believe that voice-queries deserve further research. The reason behind the decision for typing or speaking a query is difficult to analyse based on the logged data. Perhaps users are shy of speaking to their smartphone in the public. Further studies would be necessary to gain a proper insight into this behaviour. The information obtained from this early study points to a number of potential avenues for further research. One plan we have is to look at different usage patterns with the system and see how they correlate with the outcomes of the evening e.g. number of events visited, the ratings of visit events, the geographical coverage of the user etc. This would provide insight into how the features of our system support casual-leisure needs.

Acknowledgments This work was supported by the Embedded Systems Initiative (http://www.esi-anwendungszentrum.de).

9. REFERENCES

[1] *Die Lange Nacht der Musik Besucherbefragung 2010.* Münchner Kultur GmbH, 2010.

[2] N. J. Belkin, R. N. Oddy, and H. M. Brooks. ASK for information retrieval: Part I. Background and theory. *Journal of Documentation*, 38(2):61–71, 1982.

[3] K. Byström. *Task complexity, information types and information sources. Examination of relationships.* PhD thesis, University of Tampere, Dep. of Inf. Studies, 1999.

[4] F. Dornseiff. *Der deutsche Wortschatz nach Sachgruppen.* DeGruyter, Berlin, New York, 2004.

[5] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff I've seen: a system for personal information retrieval and re-use. SIGIR '03, pages 72–79, NY, 2003. ACM.

[6] D. Elsweiler, S. Mandl, and B. Kirkegaard Lunn. Understanding casual-leisure information needs: a diary study in the context of television viewing. IIiX '10, pages 25–34, NY, 2010. ACM.

[7] D. Elsweiler, M. L. Wilson, and B. Kirkegaard Lunn. *New Directions in Information Behaviour*, chapter Understanding Casual-leisure Information Behaviour. Emerald Pub., 2011.

[8] P. Hansen. User interface design for IR interaction. a task-oriented approach. In *CoLIS 3*, pages 191–205, 1999.

[9] B. J. Jansen and A. Spink. How are we searching the world wide web?: a comparison of nine search engine transaction logs. *IPM*, 42(1):248–263, 2006.

[10] M. Kamvar and S. Baluja. A large scale study of wireless search behavior: Google mobile search. In *CHI 2006*, 2006.

[11] M. Kamvar, M. Kellar, R. Patel, and Y. Xu. Computers and iphones and mobile phones, oh my!: a logs-based comparison of search users on different devices. WWW '09, pages 801–810, NY, 2009. ACM.

[12] C. S. Ross. Finding without seeking: The information encounter in the context of reading for pleasure. *IPM*, 35(6):783–799, 1999.

[13] J. Teevan, E. Adar, R. Jones, and M. A. S. Potts. Information re-retrieval: repeat queries in Yahoo's logs. SIGIR '07, pages 151–158, NY, 2007. ACM.

[14] E. Toms. Understanding and facilitating the browsing of electronic text. *J. of Human-Comp. Studies*, 52:423–452, 2000.

[15] M. L. Wilson and D. Elsweiler. Casual-leisure Searching: the Exploratory Search scenarios that break our current models. HCIR '10, Aug 2010. New Brunswick, NJ.