Improving search experience on distributed leisure events

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

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

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

ABSTRACT

This paper examines how simple changes to a search system can influence the user's experience when using the system. In previous work, we evaluated user behaviour with a search tool designed to help people discover events distributed over a city of interest to them personally. We established, contrary to our expectations, that users mostly searched for events they already knew about, made several spelling errors and often achieved poor search performance. Taking these findings as inspiration, we made changes to how the system works. In this paper, we describe and motivate the changes and present a naturalistic log-based study (n=860) to examine the effect on user search behaviour.

1. INTRODUCTION AND MOTIVATION

When studying search behaviour most published work focuses on analysis in the context of work tasks. Such tasks are not necessarily related to work but rather involve people performing a sequence of activities in order to accomplish a goal [5]. A work task has a recognisable start and end, may consist of a series of sub-tasks, and results in a meaningful product [2]. Thus models developed tend to assume that people look for information to close a gap in knowledge [1] which prevents them from completing their current task.

In contrast we want to do search analysis in context of leisure activities where no clear focus on a concrete working task is given. Elsweiler and colleagues [4] 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 are not focused on accomplishing a task but rather aim to be entertained or to pass time. These needs are influenced by emotional state, physical state or the social context in which they live. Additionally such needs differ from work tasks by weighting the emotions induced by the found content or even the search process itself more than the raw informational content.

Schaller et al. analysed in [7, 8] mobile search behaviour in the context of a distributed event and compared search characteristics and performance of a naturalistic user study to those of mobile web search. A main finding is that the analysed queries were much shorter than those of mobile web search. Also, most queries – in contrast to web search – were for known-items: predominantly events names. Users made a huge amount of spelling mistakes perhaps due to the environmental context (e,g, typing on a bumpy bus) or due to the unfamiliarity with the correct spelling of the known-items. This is probably one of the causes of the poor search performance with over 40% of searches being unsuccessful. A major conclusion of the paper is that people used the system as a tool for filtering and not for searching for new things. The paper gives suggestions for better tailoring the search system towards the observed user behaviour.

In this paper we build upon these results and explore how the proposed changes to the system influenced search characteristics and performance with the overall aim of improving search experience. We first describe the design of a study to answer this question. We report analyses of interaction logs of a variant of the search system including the proposed changes which was tested on a similar event as was used in [8]: The Long Night of Music 2012 opposed to the Long Night of Museums 2011, both located in Munich. We admit possible doubts as to the comparability of these Nights due to the different topics: however we are able to address these doubts by showing that user behaviour is similar enough to make valid comparisons of system changes. We then show differences and similarities in search behaviour between the original and the improved search system and finally draw conclusions as to the effectiveness of the analysed changes.

2. DISTRIBUTED EVENTS

A distributed event is a collection of single events over the same time period and having the same general theme. One such event is the Long Night of Munich Museums[1] (LNMuseums), an annual cultural event organised in the city of Munich[2], which was the context of the study performed in [8]. 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 for one evening in October. Many venues organise special activities and exhibitions not otherwise available. A similar distributed event is the Long Night of Music (LNMusic) which also takes place in Munich. Aside from pubs, discotheques and clubs also some cultural venues like churches and museums take part which leads to some overlap between both nights regarding the provided events.

Visitors to 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 at the LNMuseums, including exhibitions, galleries and interactive events. The LNMusic is about the same size with 206 events at 123 locations and approximately 20,000 visitors. Events on both nights take place all over the city, mostly in the city centre, but some, such as the Museum of the MTU Aero Engines

[1]Name in German: Lange Nacht der Münchner Museen
[2]The event is organised by Münchner Kultur GmbH (http://www.muenchner.de/museumsnacht/)

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

Events can be discovered by means of 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 per Long Night) and can be difficult to navigate.

3. SYSTEM

An Android app was developed in [8] to help visitors of the Long Night find events of interest to them personally. Once they have found and selected 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 then the system tries to maximise the number of scheduled events by leaving out those requiring 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] as was used on the LNMuseums and LNMusic.

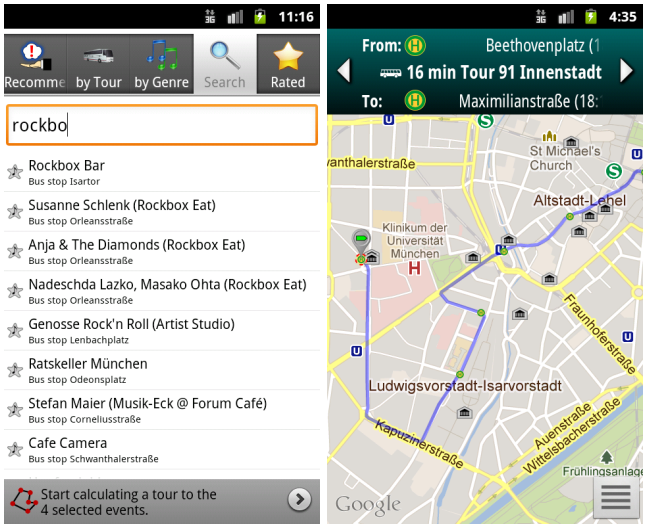

Figure 1: The search screen with a query (left) and the map screen with the planned route (right)

The user has four ways to find events he would like to visit, namely he can: receive recommendations based on a pre-defined profile and collaborative filtering algorithm built into the app; browse events by bus route; browse events by genre or type or submit free-text queries, which search over the names and descriptions of the events.

As described in [8] the search functionality was implemented in Lucene[4] and documents were represented by titles and descriptions from the Long Night booklet. Lucene was extended to perform a search based on topics. Firstly the event descriptions and titles were tokenised and stemmed then to match topically similar words, each token is mapped to one or more topic groups (these groups are taken from [3]). 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.

[3]a video demo of the application can be found on YouTube (http://www.youtube.com/watch?v=qy1F8fZbowo)

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

4. SEARCH SYSTEM CHANGES

Based on insight gained from user interactions with the original search system, as described in [8], it was determined that the following improvements could be made:

- Grep-Like Search:
  Since users used the system mainly as a filtering tool, the search-as-you-type feature might have led users to give up early: the system tried to match whole (or stemmed) words while the user faces an empty result list after typing in the first few characters but before finishing the word. In our new system – used and evaluated on the LNMusic – we extended the search system with a grep-like feature which would also match parts of words and not just complete words. For example, if a user is looking for the event "Lenbach" it is sufficient to type in just "Lenb". This means that users are not so often presented with an empty list of search results.

- Fuzzy Search:
  It was noticeable from the user interactions that a huge number of spelling errors were being made. This was presumably due to environmental factors, e.g. typing on a bumpy bus or due to the a high number of named entities, the spelling of which people are not familiar. In either case the system was adapted to better support the user by performing a fuzzy search according to [8]. Our system was improved by utilising the Lucene Fuzzy Search mechanism which uses the levenshtein edit distance to match term that differ only by a few characters. If looking for the event "Lenbach" it will be found even if the user by mistake typed "Lembach".

Both changes aim to allow users to more quickly and easily find the events they are interested in and improve the overall search experience.

As the same naturalistic study was used to analyse other parts of the system there were other small changes which are unrelated to the search system analysis: the number of tabs was increased due to the addition of a recommender tab which was beforehand combined with the list of already selected events in one tab. Secondly, the tab position of the recommender tab was tested in an A/B test. Both changes are beyond the scope of this search behaviour paper and should not have any significant influence on it as the layout of the search tab itself wasn't changed.

5. DATA COLLECTION

We examined the search behaviour of users by recording user interactions with our refined app at the LNMusic 2012 in the same manner as with the original app on the LNMuseums2011. Again the app was available for download from Google Play Store and advertised on the official Long Night of Music web page. In total the application

was downloaded approximately 1000 times and 860 users allowed us to record their interaction data (In [8] approx. 500 downloads and 391 users are reported). 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 11.79 minutes[5] with the system (median 6.46). 57.3% of users interacted for more than 5; 17.9% for more than 30.

Since queries were submitted after every typed character, it is 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.

As in [8] we manually judged queries to be intended or not. 3 assessors separately annotated all of the 12,500 queries logged as being either intended or not-intended. A very high inter-assessor agreement was found (Fleiss' kappa = 0.915, 89.8% of queries which were labeled by at least 1 assessor were also labelled by at least one other). This process resulted in a final list of 1,434 search queries, which is used in the following analyses and compared against the results reported in [8] which are based on 801 search queries.

6. IS A COMPARISON OF NIGHTS FAIR?

Undoubtedly the best way of comparing two version of a system is to run experiments under the same external conditions. Unfortunately this was not possible with our app for the LNMusic as we work together with the organisers to provide a real system for "productive" use and cannot experiment with arbitrary system variations. The alternative would be a lab study, however we consider this to be a less preferable option given that we wish to record interaction with the system in a real-world (i.e. non-simulated) setting. Therefore we looked into whether data obtained from different events could be fairly compared.

We learned from our experience with past Long Nights that user behaviour is to a huge extent independent from the actual type of Long Night. In [7] interviews with visitors of two different Long Nights revealed that beside the topic of an event other characteristics – such as novelty, the time and location of the event or the possibility to take part in the event – play a crucial role. It is precisely this that sets a system for casual leisure activities apart from a system for solving a work task [4].

In this section we want to give more insights into why our study on the LNMusic2012 and the study presented in [8] on the LNMuseums2011 are comparable. First of all both distributed events – although their topics are different – have a lot in common: They take place in the same city, are organised by the same company and hence have the same ads, booklet format, price tag and even the special bus routes provided are partially the same. We noticed that some of the events on the LNMusic were also available on the LNMuseums. We investigated further into how many events are in common between both nights. To do so we looked into how many LNMuseums events had, at the same location, an event on the LNMusic that were organised by the same museum, church, bar, etc.. Surprisingly 21.6% of the LNMuseums2011 events had a matching event on the LNMusic2012. The topic of the matched events might differ but as mentioned above the topic is only one of many relevant aspects for visitors to choose an event.

Secondly we looked into the app usage itself. Upon first start-up of the app we ask our users to fill out a short questionnaire; among others we ask for the age of the user (below 18, 18 to 29, 30 to 39, 40 to 49, 50 to 59 and above 60 years old). Answering of these questions was optional but 246 on the LNMuseums2011 and 495 users on the LNMusic2012 chose to do so. Based on these data (omitting the first and last age groups due to the small sample size) we compared the age distribution of users on both nights with a χ^2-Test revealing a χ^2 of 1.459 and $p = 0.6918$. This result states that this is no significant difference between app users of both nights.

To ascertain if the two studies are comparable it is important that system usage is similar on both nights. As described in Section 3 there were multiple ways (different tabs) of accessing the events. We looked into which of these tabs users were interested in. We therefore define a tab session to start when a user switches to a tab and to end when he switches to another tab. Table 1 shows the number of tab sessions. Based on this data, a χ^2-Test shows a χ^2 of 5.387 and $p = 0.1456$, again indicating no significant difference between users of both nights.

	by Tour	by Genre	Search	Rec.+Rated
LNMuseums	28.0%	14.5%	15.4%	42.0%
LNMusic	26.6%	15.6%	15.1%	42.7%

Table 1: Number of tab sessions

Lastly we considered properties of the search behaviour itself that should be invariant to our changes to the search system. One of the main findings of [8] was the huge number of named-entity searches and we compared the reported numbers to those of our own system. Using the same method described in [8] we instructed 3 human accessors to label all search queries into one of three categories: specific event name, not a specific event name or indeterminate. For 82.0% of all queries at least two of the assessors were able to agree on one of the three categories (Fleiss Kappa of 0.32). 84.5% (LNMuseums2011: 59.4%) of the agreed on queries were marked as clearly named entities and 8.2% (34.6%) that might be named entities. Only 7.2% (6.0%) were labeled as non named-entity searches. It is notable that the low number of non named entity searches is similar to what is described in [8].

In [7] it is reported that the same system used on the LNMuseums2011 was also evaluated on the Long Night of Science in Erlangen-Nuremberg (LNScience2011), which also is a distributed event but dedicated to science. We looked into how search characteristics differ if the same system is used on different nights. We compared query length with respect to the number of characters and the number of terms per query by performing a (non-parametric) Kruskal-Wallis Rank Sum Test. No significant difference between the usage

[5]These figures were calculated by summing the time periods for which a user was active, discounting times where the system reported no interactions for more than 15 seconds. We further discounted any interaction sequence that contains gaps of non-interaction longer than 30 minutes as these are likely due to logging problems caused by running out of power, connection problems, app crashes, etc.

on the LNMuseums2011 and the LNScience2011 could be found (for characters: $p = 0.1169$; for terms: $p = 0.6039$). When performing the same test between queries on the LNMuseums2011 and the LNMusic2011 a highly significant difference can be found with $p \ll 0.01$ for both characters and terms. Thus changes to the search system have an influence on search behaviour but changes to the overall setting of the distributed event have not.

In conclusion we believe a comparison of the app user behaviour on both nights is appropriate, given the circumstances and difficulty of obtaining real-world user data of such apps.

7. INFLUENCES OF THE CHANGES

In [8] many statistics on query characteristics and query performance are given based on analysis of search logs, a common technique in the literature [6]. In this section we recalculate these statistics based on the data logged with the new system and compare behaviour with both app variants to determine what user behaviour changes the search system modifications caused. To do so we consider a number of different indicators of an improved search experience.

The average length of a search query on the LNMusic2012 was 5.6 characters ($\sigma = 3.36$) and 1.14 terms ($\sigma = 0.41$). This is much shorter than what is reported [8] for the LNMuseums2011: 8.9 characters ($\sigma = 5.31$) and 1.21 terms ($\sigma = 0.52$). A Z-test performed between both nights reveals that this difference is highly significant for both metrics (i.e. $p \ll 0.01$ in both cases). It seems that the grep-like searching – which matches also partial words – has influenced people to stop typing much earlier. We assume that there are two main causes that users stopped typing: either they have found what they were looking for (successful search) or they gave up on the search because they couldn't find what they were looking for (unsuccessful search). Users of our system had three options to interact with the entries in the result list: they could view details of an event, mark an event as a candidate for tour inclusion or add the event to an pre-existing tour. We consider any of the three as an indicator for search success and the lack thereof as an indicator of an unsuccessful search. Good abandonment wasn't considered since the result list contains no information beyond the event name and nearest bus stop. The length of successful queries on the LNMusic2012 was 5.39 characters ($\sigma = 3.27$) and 1.12 terms ($\sigma = 0.38$) which is highly significantly ($p \ll 0.01$) shorter than reported in [8]: 9.90 characters ($\sigma = 5.42$) and 1.26 terms ($\sigma = 0.57$). This means users have to type on average 45% less to find the events they are interested in. On the other hand the query length of unsuccessful queries was slightly reduced with 6.39 characters ($\sigma = 3.56$) opposed to 7.47 characters ($\sigma = 4.80$) but slightly longer with regard to terms: 1.20 ($\sigma = 0.49$) as opposed to 1.13 ($\sigma = 0.42$).

Of the 1,434 queries entered on the LNMusic 76.7% resulted in an interaction of the user with an event, meaning they were successful. 23.3% were unsuccessful, a much better conversion rate compared to the 40.3% unsuccessful searches in [8]. This decrease is highly significant ($p \ll 0.01$) and demonstrates that the improved search system was able to assist users in finding events they were looking for.

In [8] a large ratio of 59.75% of unsuccessful search queries had an empty result list. With the improved search system this was only the case in 12.57% of unsuccessful queries. But how successful were those "added" entries? We looked at all 2,157 interactions with events (viewing, rating, selecting) on the result list of the LNMusic2012 app (created by the improved search system). We then ran the corresponding search queries through the old search system and counted whether these events would be on the result list had the original search system been used. Only 938 event interactions would be possible with the previous search system, meaning that 56.5% of the interactions performed by users wouldn't have been possible, simply because the events wouldn't be in the results.

This analysis has revealed indicators of an improved search experience which means that the changes proposed in [8] are useful in the context of distributed events assistance.

8. DISCUSSION AND CONCLUSIONS

In this paper we analysed the changes in query behaviour of users due to modifications of a search system used on distributed events. We first describe two studies performed during two such events, including a description of the system used and what changes to the search system were tested. As the LNMuseums and LNMusic have different topics we then showed that a comparison of user behaviour between both nights is sensible and worthwhile. With this preparatory work we then analysed users' search behaviour by comparing search characteristics and search performance. Overall, users typed much shorter search queries, especially in the case of a successful search. Also comparing query performance revealed a much higher success rate with the ratio of unsuccessful searches being almost halved. Finally we presented a comparison of both search systems running on the same search queries which showed that only half of the interactions with events would have been possible with the old system.

The search system as it is now is designed for users to find events they already know of in advance. But how can users be assisted in finding events that are new to them? How can we better support the discovery of serendipitous events? Since the users seldom used the search system for that purpose, a second tool like a recommender is necessary. The user could then decide if he wants to look for a concrete event he already knows of or if he would rather be inspired by the system. If such a split into two "orthogonal" tools is understood and accepted by users then it is worth investigating and would point the way to vastly better distributed events assistance systems.

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

9. REFERENCES

[1] 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.

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

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

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

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

[6] B. J. Jansen and A. Spink How are we searching the world wide web?: a comparison of nine search engine transaction logs In *IPM*, (1,42) pp. 248–26, 2006.

[7] R. Schaller, M. Harvey, and D. Elsweiler. Entertainment on the go: Finding things to do and see while visiting distributed events. In *Proceedings of IIiX*, 2012.

[8] R. Schaller, M. Harvey, and D. Elsweiler. Out and about on museums night: Investigating mobile search behaviour for leisure events. In *Proc. of Searching4Fun Wksp, ECIR*, 2012.