How people use the web in large indoor spaces

Ren, Yongli ; Tomko, Martin ; Ong, Kevin ; Sanderson, Mark

DOI: https://doi.org/10.1145/2661829.2661929

How People Use the Web in Large Indoor Spaces

Yongli Ren[1], Martin Tomko[2], Kevin Ong[1], Mark Sanderson[1]
[1]School of Computer Science and Information Technology, RMIT University, Melbourne, Australia
[2]Department of Computing and Information Systems, the University of Melbourne, Melbourne, Australia
yongli.ren@rmit.edu.au, tomkom@unimelb.edu.au, kevin.ong@rmit.edu.au,
mark.sanderson@rmit.edu.au

ABSTRACT

We report a preliminary study of mobile Web behaviour in a large indoor retail space. By analysing a Web log collected over a 1 year period at an inner city shopping mall in Sydney, Australia, we found that 1) around 60% of registered Wi-Fi users actively browse the Internet, and the rest 40% do not, with around 10% of these users using Web search engines. Around 70% of this Web activity in the investigated mall come from frequent visitors; 2) the content that indoor users search for is different from the content they consume while browsing; 3) the popularity of future indoor search queries can be predicted with a simple theoretical model based on past queries treated as a weighted directed graph. The work described in this paper underpins applications such as the prediction of users' information needs, retail recommendation systems, and improving the mobile Web search experience.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process; H.1.2 [**User/Machine Systems**]: [Human factors]

General Terms

Experimentation

Keywords

Indoor mobile Web behaviour, query popularity propagation

1. INTRODUCTION

Large-scale indoor spaces are increasingly equipped with free Internet via Wi-Fi, the use of which can be logged and studied. With such infrastructure in place, it is possible to collect a variety of parameters about user behaviour. These environments are visited on a regular basis by a large number of visitors. For instance, The Mall in Dubai attracted 75 million visitors in 2013 [2]. Tracking the location of users in time and space along with their Web activity allows for the study of their needs, establishing whether these are appropriately supported by the environments, and exploring how the indoor spaces cope with the presence of diverse visitors; an important consideration in, for example, hospitals [11].

An indoor space imposes a range of social, technical, and physical constraints. While Web query logs have been widely studied, both on desktop machines [9, 5] and on mobile devices [12, 7, 4, 8], there are few published studies that analyse Web search in large-scale indoor spaces. By analysing a log of Web activities of more than 120,000 users in over 1 year period, we study how people behave on the Web in indoor spaces and whether information access trends can be predicted. Such predictions have practical applications, e.g. for predicting consumer behaviour.

In this paper, we address the following questions:

- How do people behave on the Web in large indoor retail spaces? (Section 3)
- What is the indoor Web search behaviour? (Section 4)
- Can the popularity of indoor Web search queries be predicted? (Section 5)

2. DATA ACQUISITION

We study an anonymized dataset of internet accesses taken from a free Wi-Fi network operated by a large inner-city shopping mall. The mall is covered by around 70 Wi-Fi Access Points (AP). The dataset includes three kinds of logs: a Wi-Fi Access-point association Log (AL), a Web Browsing Log (BL) and a Web Query Log (QL), collected between September 2012 and October 2013. Table 1 shows summary statistics of the three logs. For this research, all user identifiable information was replaced by a hash key in an non-invertible way.

Only devices connected to the free Wi-Fi provided by the mall are logged. The logs do not track users but rather mobile devices through the device's Wi-Fi MAC address, which was replaced by a hash key. Thus, there is no ground truth about the identities of the users (e.g. shoppers or mall employees). The term *Users* is used to refer to unique devices appearing in AL, a subset of such users are *browsers* who appear in the BL, and *searchers* are those users who appear in the QL. The BL includes traffic to the Web originating from the mobile Web browser, as well as from other apps. Traffic from apps currently cannot be easily filtered out. The QL was extracted from the BL (it is a subset of BL), by isolating searches pointing to search engines, including Google(110148, 92.4% of QL), Yahoo (6915, 5.8%

CIKM'14, November 03 - 07 2014, Shanghai, China

http://dx.doi.org/10.1145/2661829.2661929.

Table 1: Summary statistics of the AL, BL and QL.

Wi-Fi Access point Log (AL)			
Number of users:	120,548		
Number of AP association:	907,084		
Number of User Visits:	261,369		
(hour)	Mean	(Max, Min)	SD
Mean duration per visit:	2.77	(21.67, 0.08)	1.52
Duration per visit:	#	%	
(0, <1h)	43,680	16.71%	
(1h, <2h)	14,119	5.4%	
(2h, <3h)	10,589	4.1%	
(3h, <4h)	172,149	65.86%	
(4h, max)	20,832	7.97%	
Web Browsing Log (BL)			
	#	%(AL)	
Number of browsers:	70,196	58.23%	
Number of Web accesses:	18,088,018		
Number of User Visits:	139,004		
	Mean	(Max, Min)	SD
Web accesses per visit:	135	(5393, 1)	209
	#	%	
(0, <50)	58,629	42.18%	
(50, <100)	26,068	18.75%	
(100, <150)	15,956	11.48%	
(150, max)	38,351	27.59%	
Query Log (QL)			
	#	%(AL)	%(BL)
Number of searchers :	11,169	9.27%	15.91%
Number of queries:	119,196		
Number of query sessions:	20,637		
	Mean	(Max, Min)	SD
Mean queries per session:	5.78	(622, 1)	11.00
Queries per session:			
1	6,333	28.3%	
2	4,787	21.4%	
3+	11,255	50.3%	

of QL), Bing (954, 0.8% of QL), Baidu (1086, 0.9% of QL), AOL (43, 0.04% of QL) and ASK (50, 0.04% of QL).

The AL captures information about user physical behaviour through the following parameters: (1) users' location in the mall defined by the location of the Wi-Fi access point associated with the user's mobile device; (2) timestamp and duration of users' association with the access point. The BL includes the users' information behaviour, characterised by: (1) the timestamp of the Web request; (2) the Uniform Resource Locator (URL) of the requested Web page.

The QL was processed as follows: (1) search queries were treated as case insensitive; (2) a query term was defined as any unbroken string of characters in a query delimited by the whitespace symbol (`U+0020`); (3) Following [7, 4], sessions were defined as "a series of queries by a single user made within a small range of time". Following a recent study in Mobile Web search [12], we also use 30 minutes as the threshold for maximum session duration; (4) session length was defined as the number of queries in a session; (5) query length was defined as the number of terms in a query.

3. PHYSICAL BEHAVIOUR AND WEB INFORMATION BEHAVIOUR

In this study, we analyse the physical behaviour and Web information behaviour for indoor users by dividing them into four groups (number of unique users in brackets): (1) *Open (112,322)* denotes the users who appear during the mall's business hours (when shops are open); (2) *Closed (8,226)* denotes the users who appear when the mall is closed (there are some restaurants which are still open to the public after the retail part of the mall is closed); (3) *Once (80,900)* denotes the users who appear only once in the collected period when the mall is open; (4) *Many (31,422)* denotes the users who appear more than once in the collected period when the mall is open. We observe that the majority of users come from the *Open* group, and there are more *Once* users than *Many* users. The *Closed* group was not split due to its small size. The majority of this group visited only once. Table 2 shows the comparison of the physical, browsing, and query activities among these user groups.

From AL, we can observe that (Table 1 and Table 2): (1) around 66% of users stay in the mall between three and four hours per visit, with 26% staying less than two hours and 8% exceeding three hours; (2) while *Once* users outnumber *Many* users, *Many* users tend to contribute more to the AL logs than *Once* users, due to their repeated visits to the mall (56.99% of *Many* visited twice, 16.13% of them visited 3 times, and 26.88% visited 4 times or more); (3) *Many* stay in average longer than *Once* users (2.9 hour vs 2.4 hour), while *Closed* users stay longest (3.2 hour). Overall, mall users are most likely to stay in the mall for around 3 hours on average across different user groups.

From the BL, we observe that: (1) around 60% of Wi-Fi users in the AL actively browsed the Web, while the rest visited the mall but did not use the Web; (2) around 60% of these users accessed fewer than 100 URLs; (3) *Many* is the most Web-active user group (around 144 URLs per visit, significantly more than values for groups *Once* and *Closed*); (4) the number of BL logs from *Closed* is fairly low, with users only averaging 13 URLs per visit. Table 3 shows a breakdown of the content these users browse for on the Web. The top-10 popular Web categories of indoor browsing[1]. While some of the top categories are related with retail activities (e.g. *Business and Economy* and *Web Advertisements*), many are related to social networking, infotainment and Personal information management (PIM). These results indicate that frequent visitors tend to be the most active on the Web, where they use diverse, and not only retail related content.

Consequently, we explore what indoor users search for on the Web. Is this content similar to their general search? From the QL, we find that: (1) around 10% of overall users in the AL (16% of browsers in BL) are searchers; (2) there is no big difference for indoor searching among different user groups in terms of the number of queries per session. (3) around 50% of the searchers issued three or more queries per session, which is different from general Web searchers [13]: most general Web searchers issue only one query per session. Potential explanations for these differences include: (a) people search differently in indoor retail spaces; (b) mobile search behaviours have changed since the publication of earlier studies; or (c) the behaviour of the modern interfaces has altered the patterns detectable in the logs (e.g., query suggestion by Google); (4) the top 10 indoor search categories (query-click) (Table 3, right column) are dominated by retail-related activities, e.g. *Travel*, *Shopping*,

[1]The Web page categories were generated through the public Webroot Content Classification Service *bcws.brightcloud.com*. Although DMOZ has higher accuracy, its coverage is to limited for our study. E.g. *www.gumtree.com.au* is not categorized in DMOZ but correctly categorized as *shopping* by *BrightCloud*.

Table 2: Physical and Web activities by user group

Log	statistics	Open		Closed
		Once	Many	
AL	No. of AP association	257,444	530,028	119,612
	% of AP association	28.4%	58.4%	13.2%
	mean duration (hour)	2.4	2.9	3.2
BL	No. of Web accesses	4,202,073	12,738,140	1,147,805
	% of Web accesses	23.2%	70.4%	6.3%
	mean Web access	47.4	143.7	13.0
QL	No. of Queries	29,802	83,997	5,397
	% of Queries	25%	70.5%	4.5%
	mean queries per session	5.50	5.87	5.85

Table 3: Top 10 Categories of Browsing and Query-Click

Browsing	Query-Click
Social Networking (20%)	Travel (12%)
Content Delivery Networks (13%)	Entertainment and Arts (9%)
Computer and Internet Info (12%)	Society (8%)
Search Engines (11%)	News and Media (8%)
Business and Economy (10%)	Shopping (8%)
Personal Storage (5%)	Reference and Research (7%)
Web based email (3%)	Social Networking (6%)
Web Advertisements (3%)	Business and Economy (6%)
News and Media (3%)	Personal Sites and Blogs (4%)
Internet Portals (2%)	Computer and Internet Info (4%)

Reference and Research and *Business and Economy*; (5) the search patterns of indoor users are different from their browsing activity, e.g. while *Social Networking* is the most popular browsing category (consistently with mobile Internet usage [3]); *Travel* is the most popular query-click category. Specifically, *Travel* makes 1.4% of browsing but 12% of searching and *Social Networking* takes 20% in browsing but only 6% in searching.

These differences imply that indoor browsing and searching should be treated differently to improve users' Web experience, because users are likely to satisfy different information needs via browsing and searching, respectively. For example, top browsing activities tend to be retail irrelevant, and may be accessed to satisfy common information needs not directly linked to the retail environment; while top search activities tend to be linked to retail, which may indicate tighter dependence on the spatial context of the retail environment. The role of specialised apps is also likely contributing to the difference in browsing vs. query traffic – users are more likely to use pre-installed specialised apps (e.g., Facebook) for social interaction rather than using the mobile version of the sites, and in any case the related Websites represent known target that need not be searched for. Moreover, while adult-related search was popular in general mobile search [12, 7, 4], it is not popular in both indoor browsing and searching. We suspect one main reason is because the data were collected in a public indoor retail space.

Overall, we conclude that indoor users of the *Many* category tend to be much more active in the Web browsing. What indoor users browse for is different from what they search for. The Web categories of the searching activity are more related with the retail environment. Moreover, while the user groups differ in the number of URLs per visit in their Web browsing patterns (for a deeper analysis of BL, see [10].), the difference is small in the average number of search queries per session. We focus on the analysis of QL in the following Section 4.

Table 4: The session length l, $mean(l)$ and $|q|$ of indoor Web queries, general mobile Web queries and general Web queries. "-" means that the value was not listed in the corresponding paper.

l	Indoor	General Mobile			General Web Search	
		Bing	Google	Euro	Dogpile06	Dogpile05
1	**28.3%**	-	68.0%	45.0%	52.8%	53.9%
2	21.4%	-	19.0%	17.0%	17.2%	16.6%
3+	50.3%	-	13.0%	38.0%	30.0%	29.4%
$mean(l)$	5.78	1.48	1.6	5.78	-	2.85
$\|q\|$	2.79	3.05	2.3	2.06	2.83	2.79

4. INDOOR WEB SEARCH BEHAVIOUR

In this section, we investigate Web search session length l and query length $|q|$, which are two of the fundamental characteristics in Web search [1]. Table 4 shows the comparison of l, $mean(l)$ and $|q|$ among indoor queries, general mobile Web queries and general Web queries. We compare to past work on: Bing [12], Google Mobile [7], Euro Mobile [4], Dogpile '06 & '05 [5, 9]. These studies are selected because they study general mobile Web search or they define sessions in a similar way to our study. We observe that users tend to issue more queries than general mobile searchers and general Web searchers, but type queries in a similar way. The difference in query numbers again points to our hypothesis above, pointing to the recent introduction of query suggestion mechanisms in search engine interfaces. Overall, these statistics differ significantly for indoor queries from general mobile Web queries and general Web queries. Specifically, in the distribution of l amongst indoor searchers, this user group has the lowest percentage of single query sessions, with the majority of indoor searchers submitting more than one query per session. We further focus on the regular patterns of the session length l for indoor searchers in Section 5.

5. INDOOR WEB SEARCH PATTERNS

We investigate the predictability of popular queries in indoor retail spaces, which has practical implications for the prediction of consumer behaviour.

5.1 Search Patterns

We first determine if the number of queries per session follows a two-parameter Inverse Gaussian (IG) distribution:

$$p(l) = IG(l; \mu, \lambda) = \left[\frac{\lambda}{2\pi l^3}\right]^{\frac{1}{2}} \exp\frac{-\lambda(l-\mu)^2}{2\mu^2 l}, \qquad (1)$$

where $\mu = E(l) = \frac{1}{m}\sum_{i=1}^{m} l^i$ denotes the mean (m is the total number of sessions in the log and l^i is the length of i-th session), $\lambda = \frac{\mu^3}{var(l)}$ is the shape parameter, and $var(l)$ denotes the variance of l .

There are two theoretical underpinnings in IG distribution, which makes itself a good model of l: 1) the indoor searching has a longer tail distribution (around half of the searchers submit more than two query per session). Thus, the consequent μ and λ in IG can form a large and long tail; 2) the relatively larger head of the l distribution can be fitted by the asymmetry feature of the IG distribution.

As Aijferuke et. al. [1] pointed out, when the sample size is large, the data can not be satisfactorily fitted using goodness-of-fit techniques, including chi-square χ^2 test and Kolmogorov-

Table 5: PCC r Values (Google and Yahoo are investigated here, because they are the two most popular search engines on mobile devices in 2012 and 2013 *netmarketshare.com* and they are also the top two search engines in our QL)

methods	All	Closed	Open	Google	Yahoo
$C(q_i)$	**0.9280**	**0.7560**	**0.9269**	**0.9211**	**0.9364**

Smirnov test. Following [6], we apply the coefficient of variation R^2 as an indication of the closeness of the fit.

To make a precise measurement of the quality of the theoretical fit, we perform a quantile-quantile analysis and observe that the R^2 value is 0.9597 ($P < 0.001$), which means that the theoretical fit describes almost all of the observed l's variance.

5.2 Predicting Query Behaviour

Here, we apply Eq. 1 to model and predict the popularity of queries in terms of query counts. We treat the collection of queries as a weighted directed graph $G = (Q, E)$, where the nodes $Q = \{q_1, \cdots, q_n\}$ denote the set of queries and weighted, directed arcs $E = \{e_{ij}, \cdots, e_{hk}\}$ connect consecutive queries. If query q_j is issued by a user immediately after query q_i, e_{ij} is defined as $e_{ij} = (q_i, q_j)$. The weight w_{ij} on e_{ij} is defined as the fraction of users who search q_i and then continue to q_j: $w_{ij} = \frac{o_{ij}}{\deg^-(q_i)}$, where o_{ij} is the number of sequential co-occurrences from q_i to q_j, and $\deg^-(q_i)$ is the in-degree of query q_i.

The prediction of the query popularity can then be modelled as the propagation process of query searching on the given graph G. Let N^j_{l-1} denote the number of users who reach q_j after searching $l-1$ queries, then the number of users who reach q_i after searching l queries is defined as: $N^i_l = \delta_l \sum_{j=1}^{n} w_{ji} N^j_{l-1}$, where $\delta_l = \frac{\int_l^{l+1} IG(\mu,\lambda)}{\int_{l-1}^{l} IG(\mu,\lambda)}$ and $IG(\mu, \lambda)$ is defined in Eq. 1. δ_l is interpreted as the fraction of users who search more than l queries over users who search more than $l-1$ queries. This propagation process stops when the majority of modelled users stops searching.

The query log was chronologically split $80\% - 20\%$ into continuous *training* and *test* sets, respectively. The *training* set is used to build the graph G, to estimate μ and λ for the Inverse Gaussian distribution, and to initialize N^i_1 for all queries. The popularities of queries in the *test* are predicted by the method described above in terms of query counts, and the prediction of the query counts $C(q_i)$ for query q_i is defined as: $C(q_i) = \sum_{k=1}^{l} N^i_k$. To measure the accuracy of prediction, we apply the Pearson Correlation Coefficient (PCC). As shown in Table 5, the predicted counts fit 4 out of 5 datasets with a r value over 0.90. The fit for the *Closed* subset is relatively low, a possible reason is that the search behaviour of the *Closed* group is less conditioned by the indoor environment (since shops are closed) and their Web search behaviour is therefore more varied and less predictable.

6. CONCLUSION & FUTURE WORK

In this paper, we report the key characteristics of users' physical and Web behaviours in large indoor spaces. This shows an initial understanding of how mall visitors behave on the Web, and provides a chance to improve their Web and shopping experience. We have established that the indoor query behaviour is predictable in terms of query popularity, which has a practical application for the detection of search trends and recommendation services. In future work, we will further characterize and model location-based indoor search behaviour in order to better suggest contextualised results.

Acknowledgement

This research is supported by a Linkage Project grant of the Australian Research Council (LP120200413). We would like to thank Stefano Mizzaro for helpful discussions, and the anonymous reviewers and CIKM shepherd for valuable suggestions.

7. REFERENCES

[1] I. Ajiferuke, D. Wolfram, and F. Famoye. Sample size and informetric model goodness-of-fit outcomes: a search engine log case study. *Journal of Information Science*, 32(3):212–222, June 2006.

[2] S. Algethami. Dubai Mall welcomes more than 200,000 shoppers a day. *Gulfnews*, 2014.

[3] K. Church and N. Oliver. Understanding Mobile Web and Mobile Search Use in Today's Dynamic Mobile Landscape. In *MobileHCI'11*, pages 67–76, 2011.

[4] K. Church, B. Smyth, P. Cotter, and K. Bradley. Mobile information access: A study of emerging search behavior on the mobile Internet. *ACM TWEB*, 1(1), May 2007.

[5] B. J. Jansen, C. C. Ciamacca, and A. Spink. An Analysis of Travel Information Searching on the Web. *Information Technology & Tourism*, 10(2):101–118, June 2008.

[6] S. Joo, D. Wolfram, and S. Song. Nonparametric estimation of search query patterns. In *iConference 2013 Proceedings*, pages 919–924, 2013.

[7] M. Kamvar and S. Baluja. A large scale study of wireless search behavior: Google mobile search. In *CHI*, pages 701–709, 2006.

[8] 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. In *WWW*, pages 801–810, 2009.

[9] A. Kathuria, B. J. Jansen, C. Hafernik, and A. Spink. Classifying the user intent of web queries using k-means clustering. *Internet Research*, 20(5):563–581, 2010.

[10] Y. Ren, M. Tomko, K. Ong, Y. B. Bai, and M. Sanderson. The Influence of Indoor Spatial Context on User Information Behaviours. In *i-ASC Workshop in conjunction with ECIR 2014*, 2014.

[11] A. J. Ruiz Ruiz, H. Blunck, T. S. Prentow, A. Stisen, and M. B. Kjærgaard. Analysis Methods for Extracting Knowledge from Large-Scale WiFi Monitoring to Inform Building Facility Planning. In *PerCom*, pages 130–138. IEEE, 2014.

[12] Y. Song, H. Ma, H. Wang, and K. Wang. Exploring and Exploiting User Search Behavior on Mobile. In *WWW*, pages 1201–1212, 2013.

[13] A. Spink, D. Wolfram, M. B. Jansen, and T. Saracevic. Searching the Web: The Public and Their Queries. *JASIST*, 53(3):226–234, 2001.