The Influence of Indoor Spatial Context on User Information Behaviours

Yongli Ren[1], Martin Tomko[2], Kevin Ong[1], Yuntian Brian Bai[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,
yuntianbrian.bai@rmit.edu.au, mark.sanderson@rmit.edu.au

ABSTRACT

Through analysing a large data set of Web logs collected at a shopping mall, this study shows that the indoor spatial context significantly influences the information contents users search for and access on the Web. Specifically, this study shows that (1) at different locations of a large-scale indoor retail space, users tend to access different kinds of Web pages; (2) at indoor locations with similar context, users tend to request similar Web pages. These findings support a range of research questions in the context of information behaviour research, from a fresh understanding of mobile Web usage in indoor spaces to new applications of mobile surfing that matches users' dynamic indoor spatial context.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Measurement

Keywords

Indoor spatial context, indoor information behaviours

1. INTRODUCTION

Visiting large-scale indoor spaces, such as shopping malls, airports, and museums, has become a pervasive part of modern life. For example, Algethami describes a mall in Dubai, which attracted 75 million visitors in 2013 [1]. The Palace Museum in Beijing attracts approximately 12 million visitors each year [9]. All such buildings are designed to serve particular purposes: shopping malls, for example, are more than just a collection of retail stores [8].

One aspect of such spaces that does not appear to have been studied on a large scale, is to what extent does the context of indoor location affect the information users need? We hypothesize that users information needs can be identified based on their Web activity and consequently, in this paper, we investigate the following research question:

Does the spatial context of a structured indoor space implicitly influence a user's information behaviours on the Web?

ECIR'14 Information Access in Smart Cities Workshop (i-ASC 2014).
April 13, 2014, Amsterdam, the Netherlands.

By analysing an anonymised data set containing 18 million Web accesses from 12 thousand users collected over a 1 year period, it is found that the users' Web information behaviour significantly changes with their indoor spatial context.

Specifically, users at different locations tend to access different Web content, while users at locations with similar spatial context tend to access similar content. To the best of our knowledge, this is the first research concerning the relationship between the context of physical indoor spaces and users' Web surfing behaviours conducted on a dataset of a significant size.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the collected data and Section 4 presents our methodology. We analyse the dataset and describe the identified association between of indoor spatial context on user information behaviours in Section 5. Section 6 concludes the paper.

2. RELATED WORK

Previous research focussed on either indoor spaces or mobile Web searching/browsing information behaviour, but rarely in connection, investigating the influence of indoor spatial context on user Web behaviour. For example, Biczok et. al. analysed the users' indoor spatial mobility through MazeMap, a live indoor/outdoor positioning and navigation system [2]. They found strong logical ties between different locations in users' spatial mobility. Church and Smyth focused on the differences between mobile browsing and mobile searching, showing that browsing was more common than searching, though mobile searching was increasingly popular [5]. Their other work [4] analysed the intent behind mobile information needs through a diary study. Church and Oliver noticed the shift that users are using mobile internet in more stationary and familiar settings, and explored the popularity of mobile usage in different contexts [3]. Similar research focused on the popularity of mobile searching in different contexts [7].

However, all this previous work only analysed context in the general forms, e. g. "at home/work", "travelling abroad", "with friends/family", "in transit/commuting". In our work, we focussed on the influence of specific contexts on user information behaviours, rather than the popularity of mobile usage in different general contexts.

3. DATA ACQUISITION

In this paper, we study a dataset of Web accesses gathered from a publicly available Wi-Fi network at a large inner-city shopping mall. The mall has over 200 stores and is covered

Table 1: The statistics of the query log data

Feature	Value
Number of users:	120,548
Number of access point association:	907,084
Number of Web accesses:	18,088,018
Number of days covered:	406

Table 2: Sample shop categories

Category	Category
Women's Fashion	Men's Fashion
Fine Jewellery	Music/Videos/DVDs
Furniture/Floor Coverings	Hair & Beauty
Fruit & Vegetable	Groceries

by between 50-100 Wi-Fi access points. The stores belong to 34 shop categories as defined by the mall operator. The data was collected between September 2012 and October 2013.

To ensure user privacy, identifying information is not stored in the data set we use. Such identifying information gathered by the operator are hashed in a non-invertible way. Table 1 shows the statistics of the collected data.

The data includes user spatial behaviour and the user Web information behaviour. Specifically, the users' spatial behaviour is characterised by the following parameters (1) users' location in the mall defined through by the location of the Wi-Fi access point with which the user's mobile device is associated; (2) timestamp and duration of users' association with the access point; (3) a computed convex area served by the access point (computed as a Voronoi cell) and related to the physical stores within this area. The users' information behaviour is characterised by: (1) timestamp of the Web request. (2) what Web page is requested, as defined by the uniform resource locator (URL); (3) the location of the users at the time of the request.

4. METHODOLOGY

We explore the associations between users' physical spatial context and their information behaviours in a large-scale indoor space. We investigate such correlation by integrating the spatial context of access points in terms of shop categories and the user information behaviours of Web accesses through Wi-Fi access points. The shop categories were made available to us from the mall owners, and the Web page categories were generated through a public Webroot Content Classification Service (*Brightcloud*[1]). Some sample shop categories are shown in Table 2.

We then define the spatial indoor context for each access point as a vector of shop categories, and the users' information behaviours are defined as a vector of *Brightcloud* categories, as follows:

DEFINITION 1. *The indoor context of access point a_i is defined as a vector of shop categories $\mathcal{C}_s$,*

$$\mathbf{E}_i = [e_{i1}, \cdots, e_{ik}, \cdots, e_{im}],$$

where $\mathcal{C}_s = \{c_s^1, \cdots, c_s^m\}$, e_{ik} is the number of shops, which are located in the Voronoi cells of a_i and belong to $c_s^k \in \mathcal{C}_s$. This vector can be computed for each access point through a spatial overlay operation between the Voronoi cells and the outline of shop footprints from the mall floor layout, although in this case it has been executed manually for quality control.

[1]http://brightcloud.com/resourcecenter/categories.php

DEFINITION 2. *The user information Behaviour at access point a_i is defined as a vector of Web page categories $\mathcal{C}_w$,*

$$\mathbf{B}_i = [b_{i1}, \cdots, b_{ik}, \cdots, b_{in}],$$

where $\mathcal{C}_w = \{c_w^1, \cdots, c_w^n\}$, b_{ik} is the average number of URLs, which users issued through a_i and which belong to $c_w^k \in \mathcal{C}_w$.

At the level of Wi-Fi access points, the influence of spatial context on users' information behaviours can be viewed as the correlation between $\mathbf{B}_i$ and $\mathbf{B}_j$ for every two access points. We use the Pearson Correlation Coefficient (PCC) to test this association, defined as follows:

$$r(\mathbf{B}_i, \mathbf{B}_j) = \frac{\sum_{c_w^k \in \mathcal{C}_w}(b_{ik} - \bar{b}_i)(b_{jk} - \bar{b}_j)}{\sqrt{\sum_{c_w^k \in \mathcal{C}_w}(b_{ik} - \bar{b}_i)^2 \sum_{c_w^k \in \mathcal{C}_w}(b_{jk} - \bar{b}_j)^2}}, \quad (1)$$

where $\mathcal{C}_w$ is the set of URL categories, $\bar{b}_i$ and $\bar{b}_j$ are the average numbers of issued URLs at a_i and a_j, respectively. Results are shown in Fig. 2 and Table 4, and detailed discussion are shown in the following section.

5. INDOOR SPATIAL CONTEXT & USER INFORMATION BEHAVIOURS

5.1 Basic Indoor Information Behaviours

We start by investigating common indoor information behaviour patterns by analysing the distribution of the URLs over URL categories. It is observed that around one fifth of URLs are associated with *Social Networking* (e.g., Facebook.com). *Content Delivery Networks* (e.g., akamaihd.net) and *Computer and Internet info* (e.g., apple.com) take roughly the same proportion, around 13%. *Search Engines* are the fourth most popular category at 11%, and followed by *Business and Economy* with 10.6%. However, the users' indoor information behaviours is different from general mobile surfing, as reported by Church and Smyth [4]. Specifically, they reported that there are only 3.2% data for *Email and Social Networking* in general mobile information needs, but this category is much more represented (at 23.1%) in our dataset study. It is possible that, either the indoor context leads to a different information behaviour, or that the information behaviour of mobile users has shifted since the publication of the study of Church and Smyth [4].

To show common information behaviours, we identify the commonality of URL categories by measuring its access entropy. For a URL category c_w, its access entropy $H(c_w)$ is defined as:

$$H(c_w) = -\sum_{v \in S(c_w)} p(v|c_w) \log p(v|c_w), \quad (2)$$

where $S(c_w)$ is the set of visits when users accessed URLs in category c_w, $p(v|c_w)$ is the percentage of accesses to c_w during a visit v out of all visits, and a visit is defined as a single device per day in the mall. A high access entropy $H(c_w)$ means that c_w is a common category among all users; a low entropy means a category is accessed by a sub-set of users.

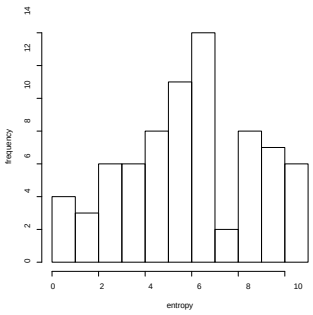

Figure 1: Distribution of $H(c_w)$

Table 3: Top-5 common URL categories and example URLs

Rank	category	example URL
1	Computer and Internet Info	*apple.com*
2	Social Networking	*facebook.com*
3	Search Engines	*google.com*
4	Business and Economy	*kakao.com*
5	Personal Storage	*icloud.com*

For example, *Computer and Internet Info*, *Social Networking* and *Search Engines* are very common URL categories with an entropy 10.75, 10.72 and 10.50, respectively. Table 3 lists the top-5 common categories based on their access entropy value and some corresponding example URLs. Fig. 1 shows the distribution of $H(c_w)$. It is observed that (1) there are some categories of websites that are more commonly visited than others, and (2) around 50% of the categories have an entropy smaller than 6.00. We will investigate this phenomenon of user information behaviour on these common and uncommon categories of websites in our future work.

5.2 Influence from Different Locations

There are differences in the types of shops served by different Wi-Fi access points. The collection of shop categories served by a single access point is what described our indoor context. We have hypothesized that the proximity of different types of shops will lead to a different Web information behaviour of the mall visitors. To test this hypothesis, we analyse the average PCC value r for every pair of access points, as defined in Eq. 1. The overall average of r reflects the general similarity of Web activity throughout the space. A small r indicates that different locations in the mall lead to different user information behaviours.

When using all URL categories, the average value of r is 0.9619, which seems to indicate that there is little difference between the information behaviours at different access points. In fact, the correlation is caused by the large proportion of common Web requests pointing to a small subset of URLs, of well defined categories. The top 5 common URL categories takes over 57.8% of the overall URL records and thus dominate the dataset. This significantly skewed Web behaviour introduces a bias in the r value.

Thus, we conduct another experiment to isolate the influence of these frequent Websites. We remove top common URL categories identified by Eq. 2 based on $p(v|c_w)$, and the r value is calculated by Eq. 1 based on $\mathbf{B}_i$. Thus, the calculation of r is independent from the identification

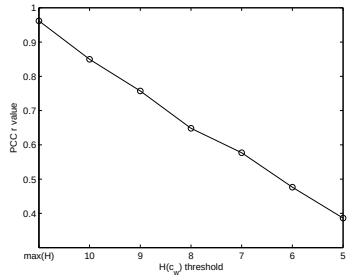

Figure 2: PCC r value without common c_w

of URL commonality. Namely, there is no logical influence between the calculation of r and the URL elimination based on $p(v|c_w)$. To show the influence of indoor location on user information behaviours, we calculate the r value by progressively eliminating common URL categories. Specifically, we select c_w based on its access entropy, $H(c_w)$ with a threshold, and we vary the threshold from $max(H(c_w))$ to 5 with a unit step[2]. Fig. 2 shows the r value over various thresholds. It is observed that when common URLs are removed from the calculation of r, differences in information behaviours at different access points appear. The more common URL categories we remove, the more substantial a difference we observe indicating that there is an influence from the local context of access points on user information behaviours.

5.3 Influence of Indoor Context

To show the influence of indoor context, we apply a clustering algorithm to group similar access points into clusters based on their indoor context. From *definition* 1, the surrounding indoor context information for an access point a_i is represented by a vector $\mathbf{E}_i$ of shop categories. We apply the k-means clustering algorithm to cluster $\mathcal{E}$ by treating each $\mathbf{E}_i \in \mathcal{E}$ as an instance. We set $k = 6$ because it achieves a relatively low value of the Davies-Bouldin index [6].

The k-means algorithm groups similar spatial contexts into clusters. If the users' information behaviour is influenced by their indoor context, the users' information behaviours *within* a cluster should be *similar* and the users' information behaviours *between* clusters should be *different*. To verify this association, we apply PCC, from Eq. 1, to measure the similarity between the information behaviours at two access points. The *intra-cluster* similarity (*within*) and the *inter-cluster* similarity (*between*) are defined as follows:

$$within = \frac{1}{k}\sum_{x=1}^{k}\left(\frac{2}{|t_x|(|t_x|-1)}\sum_{\mathbf{B}_i\in t_x}\sum_{\mathbf{B}_j\in t_x, i\neq j} r(\mathbf{B}_i,\mathbf{B}_j)\right), \tag{3}$$

where k is the number of clusters, t_x denotes the x-th cluster, and $|t_x|$ denotes the size of t_x.

$$between = \frac{1}{k}\sum_{x=1}^{k}\left(\frac{1}{|t_x|(|\mathcal{B}|-|t_x|)}\sum_{\mathbf{B}_i\in t_x}\sum_{\mathbf{B}_j\notin t_x} r(\mathbf{B}_i,\mathbf{B}_j)\right), \tag{4}$$

[2]When $H(c_w) \leqslant 4$, some $\mathbf{B}_i$ become empty, which renders the calculation of PCC r undefined. So, we analysed in the cases when $H(c_w) > 4$.

Table 4: Correlation of user information behaviours in groups of access points with similar spatial context

	$H(c_w)$	PCC r value based on $\mathcal{B}$				
		k-means		random		average
		within	between	within	between	
Groups of Access Point based on $\mathcal{E}$	$H(c_w) \leqslant max(H(c_w))$	**0.9659**	0.9623	0.9609	0.9617	0.9619
	$H(c_w) \leqslant 10$	**0.8601**	0.8509	0.8493	0.8501	0.8498
	$H(c_w) \leqslant 9$	**0.7721**	0.7599	0.7564	0.7573	0.7573
	$H(c_w) \leqslant 8$	**0.6817**	0.6572	0.6493	0.6473	0.6483
	$H(c_w) \leqslant 7$	**0.6410**	0.5966	0.5767	0.5750	0.5770
	$H(c_w) \leqslant 6$	**0.5045**	0.4778	0.4755	0.4751	0.4763
	$H(c_w) \leqslant 5$	**0.4107**	0.3942	0.3821	0.3848	0.3863

where $\mathcal{B}$ denotes the set of user information behaviours, and $|\mathcal{B}|$ denotes the size of $\mathcal{B}$. We emphasize that the groups of access points are clustered based on their physical context information $\mathcal{E}$, but the r value is defined based on user's information behaviours $\mathcal{B}$. Hence, the user's information behaviour is isolated from the clustering process.

We vary $H(c_w)$ from $max(H(c_w))$ to 5 with a unit step. We apply a *random* clustering method as a baseline to show the influence of indoor context[3]. The mean r for all $\mathbf{B}_i$ pairs is also applied as another baseline, and is defined as:

$$average = \frac{2}{|\mathcal{B}|(|\mathcal{B}|-1)} \sum_{\mathbf{B}_i} \sum_{\mathbf{B}_j, i \neq j} r(\mathbf{B}_i, \mathbf{B}_j). \quad (5)$$

Table 4 shows the results of the experiment and Table 5 the results of the analysis, where a two-tailed, paired t-test is applied to evaluate whether the observed influence is significant or not. We observe: (1) the *within* of k-means is significantly larger than the *between* of k-means. (2) the *within* of k-means is significantly larger than the *within* of *random* method. (3) the *within* of k-means is significantly larger than the *average*. (4) the *within* of *random* is not significantly different from its *between* value. (5) the *within* of *random* is not significantly different from the *average*. As shown in the first row of Table 4, even when no common URL categories are removed, the *within* value of k-means 0.9659 is larger than the corresponding *between* value 0.9623, and is also larger than that of *random* 0.9609 and the *average* 0.9619.

Table 5: Paired t-test results

Methods	t	p-value
within(k-means) VS *between*(k-means)	3.7962	0.0090
within(k-means) VS *within*(*random*)	3.5871	0.0115
within(k-means) VS *average*	3.4126	0.0143
within(*random*) VS *between*(*random*)	0.2526	0.8090
within(*random*) VS *average*	1.6007	0.1606

The results show that the observed influence is statistically significant (see paired-t statistics in Table 5). This indicates that there is an influence from indoor spatial context on users' information behaviours.

6. CONCLUSION

Based on a large data set collected through the public Wi-Fi system of a large-scale shopping mall, we present an analysis of the influence of indoor spatial context on users' Web information behaviours in large-scale retail indoor spaces. We have found that the users' indoor information behaviour manifests a significant location-based bias when the baseline, common information behaviour is excluded. Furthermore, this location-based element captured by the indoor spatial context leads to similar information behaviours between indoor locations with similar contexts. In other words, users in similar indoor contexts tend to access similar categories of Web pages, while users in dissimilar indoor contexts tend to request dissimilar Web pages. This study has raised many new research questions: 1) what are the specific differences in user Web behaviours in two kinds of indoor contexts? 2) can the differences in information behaviours help identify and recommend different spatial indoor locations that can satisfy these needs? We leave the analysis of these possible implicit effects for future work.

[3]Both *random* and k-means are run 10 times, then averaged.

7. REFERENCES

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

[2] G. Biczok, S. Martinez, T. Jelle, and J. Krogstie. Navigating MazeMap: indoor human mobility, spatio-logical ties and future potential. *CoRR*, arXiv:1401, 2014.

[3] K. Church, P. Ernest, 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 and B. Smyth. Understanding the intent behind mobile information needs. *IUI*, pages 247–256, 2009.

[5] 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.

[6] D. L. Davies and D. W. Bouldin. A Cluster Separation Measure. *IEEE TPAMI*, 1(2):224–227, 1979.

[7] J. Teevan, A. Karlson, S. Amini, a. J. B. Brush, and J. Krumm. Understanding the importance of location, time, and people in mobile local search behavior. In *MobileHCI '11*, pages 77–80. ACM Press, 2011.

[8] J. D. Vernor, M. F. Amundson, J. A. Johnson, and J. S. Rabianski. *Shopping Center Appraisal and Analysis*. 2009.

[9] C. G. Wayne. A Better Space. *Smithsonian Magazine*, 2011.