Analyzing Web Behavior in Indoor Retail Spaces

Yongli Ren
School of Computer Science and Information Technology, RMIT University, Australia.
Email: yongli.ren@rmit.edu.au

Martin Tomko
Department of Geography, University of Zurich, Zurich, Switzerland.
Email: martin.tomko@geo.uzh.ch

Flora Salim and Kevin Ong and Mark Sanderson
School of Computer Science and Information Technology, RMIT University, Australia.
Email: {flora.salim, kevin.ong, mark.sanderson}@rmit.edu.au

Abstract We analyze 18 million rows of Wi-Fi access logs collected over a one year period from over 120,000 anonymized users at an inner-city shopping mall. The anonymized dataset gathered from an opt-in system provides users' approximate physical location, as well as Web browsing and some search history. Such data provides a unique opportunity to analyze the interaction between people's behavior in physical retail spaces and their Web behavior, serving as a proxy to their information needs. We find: (1) there is a weekly periodicity in users' visits to the mall; (2) people tend to visit similar mall locations and Web content during their repeated visits to the mall; (3) around 60% of registered Wi-Fi users actively browse the Web and around 10% of them use Wi-Fi for accessing Web search engines; (4) people are likely to spend a relatively constant amount of time browsing the Web while their visiting duration may vary; (5) the physical spatial context has a small but significant influence on the Web content that indoor users browse; (6) accompanying users tend to access resources from the same Web domains.
Keywords Indoor Web behavior, indoor spatial context, log analysis

Introduction

While the use of the Web is well understood in many contexts, there is a new context emerging which is little understood: Web access in large indoor spaces, such as shopping malls, airports, universities, and museums. Indoor retail spaces impose various physical, social, and technical constraints, such as location, layout, opening hours, and Wi-Fi connectivity. Owners of these spaces design and manage them under certain economic rationale (Vernor et al., 2009), e.g. the principle of cumulative attraction where similar retail shops tend to be placed near each other. Furthermore, market management research demonstrates that the social context of retail shopping has influences on customers' shopping behaviors (Evans et al., 1996).

In many indoor spaces, free Wi-Fi is increasingly available. Visitors are thus exposed to an engineered environment with a mix of physical, social, and technical factors influencing their needs and desires. Understanding users' physical and Web behavior is fundamental to improving the designs of indoor services – both the physical retail services and the accompanying Web services.

In this paper, Web activities are analyzed based on a large-scale log of Web activity of around 120,000 users, collected over a 1 year period. Additional data about the physical environment are provided by the owner of the mall, including the floor maps of the stores, their shop categories, and the location of the Wi-Fi access points.

Corresponding Author: Yongli Ren, School of Computer Science and Information Technology, Melbourne, VIC 3001, RMIT University, Australia. Email: yongli.ren@rmit.edu.au

The diverse aspects of the physical and Web behavior of indoor users and their relationships are explored through the following research questions:

- What are the temporal characteristics of indoor Web use?
- What are the spatial (physical) characteristics of indoor Web use?
- How do the physical and social contexts influence the accessed Web content?

The main contribution of this paper is a comprehensive report of user indoor behavior. This includes an analysis of the correlation between users' physical visiting patterns and their Web behaviors; the establishment of the significant influence of the physical spatial context on the content that indoor users consume on the Web; and finally, the analysis of the correspondence between indoor users' social context and their Web behaviors. To the best of our knowledge, this is the first such research conducted on a dataset of a significant size in large indoor spaces.

Related Work

Information behavior is a term to describe the ways in which people interact with information (Bates, 2010). When envisaging information services for (indoor) use, one should consider the purpose for which mobile devices may be used. Here we review users' information behaviors on the Web, where users either search for information in a goal oriented manner or browse the content to satisfy their information needs. Web usage mining as a way to infer individualized content has been perceived superior to manually created profiles or individual user content rating-based recommendations due to the reduced subjectivity of the method, relying on actual activity patterns (Mobasher et al., 2000). The connection between indoor physical behavior (captured using mobile devices) and Web behavior has so far been insufficiently investigated – in particular on large-scale real-world datasets.

Two early studies of desktop based Web search used logs from Excite (Jansen, 2000; Spink et al., 2001) and AltaVista (Silverstein et al., 1999). They examined key characteristics of Web search queries, such as the number and distribution of terms. There are some other recent studies focusing on analysis of Web logs. For example, West et al. (2013) studied the spatiotemporal characteristics of population-wide dietary preferences. Specifically, they applied the number of recipes that users searched as a proxy for their food consumption, and they found there were two periodic components in users' dietary preferences, one yearly and the other weekly, and regional differences were also discovered.

Mobile Web use is significantly different from desktop (Kamvar & Baluja, 2006), e.g. how, when and where users search and browse the Web. Cui and Roto (2008) presented a study on how people use the Web on mobile devices, focusing on contextual factors and Web activities. They found people tend to use mobile Web while stationary and in short sessions, and proposed a Web activity taxonomy: information seeking, communication, transaction, and personal space extension. Church et al. (2007) focused on the differences between mobile browsing and mobile searching, showing mobile browsing was more common than mobile searching, although the later was increasingly popular. Church & Smyth (2009) analyzed the intent behind mobile information needs through a diary study. They found contexts influence the types of information, the goal, and the topics that users are interested in.

Contextual influence on Web use

Other studies investigated the contextual influence on mobile Web use. Teevan et al. (2011) performed a diary study on a larger scale, finding that mobile local searches were strongly influenced by context (e.g,

geographic features). Church and Oliver (2011) noticed how users increasingly use mobile Internet in more stationary and familiar settings and explored the popularity of mobile usage in different contexts. Recently, Absar et al. (2014) studied how social contexts influence and are influenced by mobile information behavior, and Patel (2015) investigated the contextual influence of wearable mobile devices in Gym.

Almost all of the previous work only modeled spatio-temporal contexts coarsely, e.g. "at home/work/Gym", "traveling abroad". In this study, we investigate two kinds of contexts, physical context in terms of shop categories, and social context in terms of user accompanying status.

Indoor behavior tracking

Indoor movement is structured by hallways and rooms (Jensen, et al., 2010), segregating spaces hierarchically by functional, organizational and social constraints (Richter et al., 2011). The structure of indoor space has been extensively analyzed by researchers of indoor navigation systems (Ruetschi, 2007; Richter et al., 2011), and is related to the constraints the space imposes on movement. Biczok et. al. analyzed users' indoor spatial mobility through MazeMap, a live indoor/outdoor positioning and navigation system (Biczok et al., 2014). They found strong logical ties between different locations in users' spatial mobility. The LiveLabs project (Misra & Balan, 2013) is an example of an in-device positioning approach for indoor user behavior tracking using a smartphone app to track users indoors. In a related study, a controlled investigation of thirty participants in a shopping mall was conducted to infer the buying intent of shoppers (Sen et al., 2014).

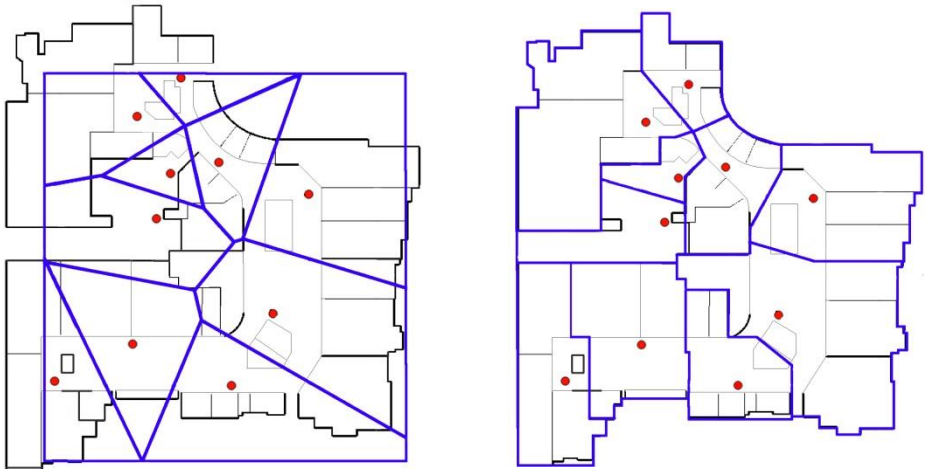
Since the organizational requirements of indoor positioning are poorly understood (Kjærgaard et al., 2014), most work focuses on limited populations of individuals over limited periods of time in instrumented settings. In contrast (as in our study), most indoor environments are set up with Wi-Fi networks to primarily provide Internet access to visitors and are optimized for coverage rather than positioning accuracy. The utility of large scale indoor tracking datasets collected as a by-product of their primary purpose over a long time for user behavior analysis is thus unknown and the applicability of insights from experiments conducted in carefully instrumented environments is uncertain.

Data Acquisition and Processing

We study an anonymized dataset of Internet accesses by registered users of a free opt-in Wi-Fi network operated by a large inner-city shopping mall covered with 67 Wi-Fi Access Points (AP) across 90,000 square meters. Visitors of the mall can register for free Wi-Fi usage and have to accept the terms and conditions of the service provider. They also provide their email addresses that are further used for dissemination of special offers and event announcements by the mall management. This Wi-Fi system is administered by the mall operators and is not set up by the authors of this study. Every year the mall attracts around 20 million visitors, who are mainly local residents and domestic and international tourists. The mall contains over 200 stores that belong to 34 shop categories as defined by the mall operator (e.g., Women/Men's Fashion, General Footwear, Café, Travel).

Floor plans of the mall were overlaid with AP locations and the service areas of the APs were approximated by Voronoi regions (Okabe, Boots, Sugihara, & Chiu, 1999), each centered on a single AP, which encompasses all the points that are closest to that AP. The regions were manually rectified to

correspond better with the frontages of physical stores in the mall (Fig. 1). Shop frontages are the main determinants of context as the Wi-Fi network is meant to cover common spaces in the mall.



(a) Theoretical Voronoi cells (b) Rectified Voronoi cells

Figure 1. An example of APs and the corresponding Voronoi cells. The black lines show the outlines of the stores. The red dots denote the Wi-Fi APs, and the blue lines show the Voronoi cells.

The dataset consists of three kinds of logs: a Wi-Fi AP association Log (AL), a Web Browsing Log (BL) and a Web Query Log (QL), collected between September 2012 and October 2013 (Table 1). Before analysis, all user identifiable information in the logs (e.g., user device's MAC address) was scrambled in an irreversible way. *Users* is the term we use in this paper to refer to devices appearing in AL, a subset of such users are *browsers* who appear in the BL, and *searchers* are those who appear in the QL.

Table 1. Aggregate statistics of the AL, BL and QL

Wi-Fi AP Log (AL)	
Number of users:	120,548
Number of AP association:	907,084
Number of User Visits:	261,369
Web Browsing Log (BL)	
Number of users browsing:	70,196 (58.3% of AL users)
Number of issued URLs:	18,088,018
Number of User Visits:	139,004
Query Log (QL)	
Number of users searching:	11,169 (9.3% of AL users)
Number of queries:	119,196
Number of query sessions:	20,637

Characteristics of the datasets

The Wi-Fi AP Association Log (AL). The AL captures information about user physical behavior characterized by the following parameters: user device's MAC address uniquely identifying the associated device; the users' IP address; the ID of the Wi-Fi AP associated with the user's mobile device

at a given point in time, used as a proxy for the user's location; the time-stamp of users' association/disassociation with the AP; the duration of users' association with the AP.

The Web Browsing Log (BL). The BL includes the users' Web information behavior, characterized by: the time-stamp of the Web request; the users' IP address; the Web page requested, as defined by the URL; the location of the user at the time of the request by AP ID.

Following (Kumar & Tomkins, 2010; Song, Ma, Wang, & Wang, 2013; Church et al., 2007), we define a browsing session as *a series of URL requests by a single user delimited by 30 minutes of inactivity on the Web*. The duration of a session is defined as the time period between the first and the last URL in the session. We assume the time within a session is spent on the Web and the time between sessions is not. For user visits accessing only a single URL (around 2.6% of overall user visits in BL) the duration is not defined and they are not further considered.

The Query Log (QL). The QL was extracted from the BL by identifying URL requests associated with search engines, including Google (92.4%), Yahoo (5.8%), Bing (0.8%), Baidu (0.9%), AOL (0.04%) and ASK (0.04%). The QL was processed as follows: search queries were treated as case insensitive; a query term was defined as any unbroken string of characters in a query delimited by white-space; the concept of sessions was applied consistently with the processing of the BL. Note, we use a narrower definition of *search* than that applied in (Hodkinson, Kiel, & McCollKennedy, 2000) and restrict this term only for search-engine based search.

Limitations of the datasets

The logs contain tracking data of mobile devices associated with the Wi-Fi network, by storing the device's (anonymized) MAC address[1]. Only those devices associated with the free Wi-Fi network provided by the mall are logged. This means that a user with a registered device may not be present in a log if they did not associate with the studied Wi-Fi network, since the user may be connected to another free Wi-Fi in the mall (e.g., a fast food chain's network) or is on their own cellular data. As current smartphones typically disassociate from Wi-Fi within a few seconds after the sleep mode turns on, disassociations are frequent and the tracking of users in the mall is not continuous. However, many apps send out URL pings frequently, thus keeping smartphones connected. This has possible impact on the analysis of user visiting duration and social context analysis. Fig. 2 shows the overall sampling tree of indoor users, and we focus on those, who are with phones, Wi-Fi enabled, joint the opt-in free network. In addition, we have no access to demographic information about the users and the reasons they visit the mall (e.g. shoppers or mall employees). We assume a MAC address remains representative of a single user across the study. Our AL data capture the timestamp of each device association with a given AP, but movement inside the region served by an AP is not captured. We define a user visit as the combination of all AL records from the same device on a single day.

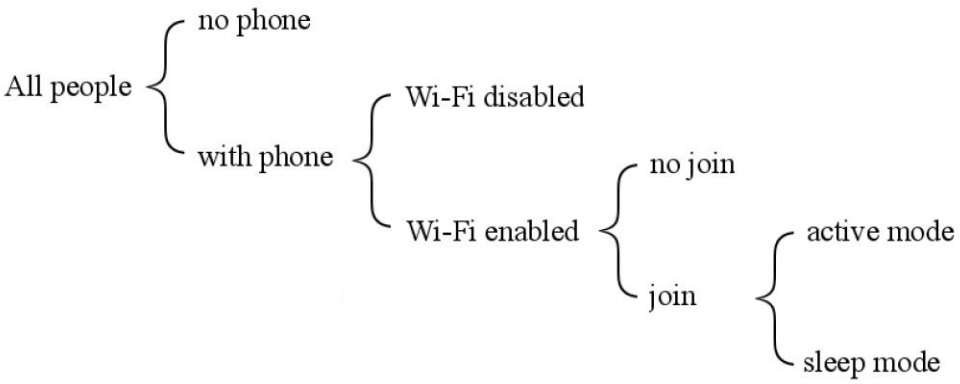


Figure 2: Sampling tree

The BL contains the unencrypted Web requests from both Web browsers and apps. We treat them equally as we analyze the requested URLs in terms of Web page categories and Web domains in this study.

The QL contains, only Web accesses over an http connection. This is of note, as in 2011 Google, started to roll out default encryption (via https) of all its queries. We examined the QL to try to determine if the move to encrypted services was an issue for our data. We split the QL in half chronologically and examined if the balance of logged searches between Google and another popular search engine (Yahoo) changed over the two halves. (Note, Yahoo was picked as it did not introduce encryption until after our logs were captured). The results show no notable difference in the two halves. Additionally the proportions recorded were found to be close to the market share reported by *netmarketshare.com*. From this result, we concluded that Google's query encryption did not influence our data.

Definitions and Terminology

Physical behavior

We study the spatio-temporal characteristics of the physical behavior of mall visitors. Their physical behavior largely equates to way-finding activity and may have goal oriented (roaming) and directed search aspects (Wiener, Büchner, & Hölscher, 2009). We restrict our focus on the manifested locomotion of the visitor but in future work hope to be able to detect the nature of the locomotion captured in the data. We denote $A = \{a_1, \dots, a_m\}$ as the set of all available Wi-Fi APs, where m is the number of APs.

Definition 1. *The user's physical behavior during a single visit v is captured by their trajectory, which is expressed as a vector $\mathbf{P}_v$ of the durations p_{vk} that the user spent associated with an AP a_k during the visit: $\mathbf{P}_v = [p_{v1}, \dots, p_{vk}, \dots, p_{vm}]$. If a user was associated with an AP multiple times in a visit, the total duration of time spent at this AP is stored, while for unvisited APs, the duration is zero.*

Web behavior

We define the indoor users' Web behavior from two aspects, visits and indoor locations (captured through AP association). We restrict our focus on the subset of information needs that are satisfied through Web interaction, and are unable to consider other social or physical information sources.

We denote $C_w = \{c_w^1, \dots, c_w^n\}$ as the set of all Web page categories, where n is the number of categories. In this paper, we applied the Web categories defined by the Webroot Content Classification Service (WCCS), BrightCloud (http://bcws.brightcloud.com)[2]. We define two kinds of user Web behavior. First, the behavior during a visit v, denoted as $\mathbf{W}_v$:

Definition 2. $\mathbf{W}_v$ *is defined as a vector of the number* W_{vk} *of URLs that are issued during* v *and belong to* $c_w^k \in C_w$: $\mathbf{W}_v = [W_{v1}, \dots, W_{vk}, \dots, W_{vn}]$.

Second, the behavior at a given AP a_i (the overall average Web behavior at an AP), denoted as $\mathbf{B}_i$:

Definition 3. $\mathbf{B}_i$ *is defined as a vector of the average number* b_{ik} *of URLs that are issued through AP* a_i *and belong to* $c_w^k \in C_w$: $\mathbf{B}_i = [b_{i1}, \dots, b_{ik}, \dots, b_{in}]$.

Physical contexts

We define physical contexts in terms of shop categories (a list of categories for each shop was provided by the mall owner), and denote $C_s = \{c_s^1, \dots, c_s^h\}$ as the set of all shop categories, where h is the number of categories. Then, we denote the spatial indoor context for each AP as $\mathbf{E}_i$:

Definition 4. $\mathbf{E}_i$ *is defined as a vector of the number* e_{ik} *of shops that are located in the Voronoi regions of AP* a_i *and belong to* $c_s^k \in C_s$, *giving* $\mathbf{E}_i = [e_{i1}, \dots, e_{ik}, \dots, e_{ih}]$.

Vector $\mathbf{E}_i$ is computed for each AP through a spatial overlay operation between the Voronoi region and the outline of shop footprints from the mall floor layout.

Social contexts

When users are visiting the mall, they may be accompanied by others. To investigate how any social relationship relates to information behavior, we define *social context* by focusing on users with highly correlated physical behaviors. We define a pair of users as *accompanying* if they: 1) both appear in the AL associated with the same AP ±1 min; 2) there is a >90% overlap in the time recorded in the AL over one visit; 3) at least three different APs are recorded in the AL for both users; 4) the average distance between the users during their visits should be no more than one AP, which means they access the Wi-Fi network via, at most, an adjacent AP.

Definition 5. *The topological distance between two user visits* v_i *and* v_j *is defined as the average step-distance between APs in the Wi-Fi signal topology, with which they are associated during their overlapped visiting time:*

$$d(v_i, v_j) = \frac{\sum_t d(a_{ik}^t, a_{jl}^t)}{\sigma},$$

where σ *is the overlapping time between* v_i *and* v_j *in seconds,* $d(a_{ik}^t, a_{jl}^t)$ *denotes the topology distance at time* t, *when these two users are visiting AP* a_k *to* a_l, *respectively.*

We focus on users recorded in the AL during opening hours of the mall. We are measuring a topological (step) distance in a graph representation of adjacencies of the service areas of APs as metric distance between the actual positions of users cannot be calculated from the log.

Basic Behaviors of Indoor Visitors to Retail Environments

Here we describe an overview of the indoor physical and Web behavior of visitors.

Basic indoor physical and social behavior

We find that the use of the Wi-Fi network corresponds to the opening hours of the mall. Starting from 09:00, the fraction of associations with the network for each hour in a day begins to increase quickly, peaking (12.69%) at 14:00, then begins to decrease until the end of the day. Examining the logs, we can also determine when users last accessed the network (a disassociation). We find there are more users associating than disassociating with the Wi-Fi network before 15:00, with disassociations peaking at 17:00, around the time when the mall is about to close. The difference between the associations and disassociations enables us to estimate the number of Wi-Fi users in the mall. Moreover, we observe that *Thursday* is the most popular day of the week for visiting the mall (17.09%). Thursdays are the typical shopping day in Australia, given the extended opening hours. There are, on average, 14.70% of visits on weekdays vs. 13.25% on weekends. Since this mall is a city center mall, and the observation results may be different at suburban malls.

Using the association and disassociation times as a proxy of visit duration, we can compute that the average time a user stays (online) is 2.77 hours, with a minimum duration of 0.08 hours and the maximum of 12.00 hours. We manually remove some outliers here by considering the opening hours of the mall, and the amount of the identified outliers is tiny. Around 66% of user visits lasted between three and four hours; 17% lasted less than one hour and around 10% lasted between one and three hours.

People have habits that lead to highly repetitive and ultimately predictable patterns (De Domenico et al., 2013). Here, we explore whether such regularities are present also in the repetitive patterns of returns to a retail environment, hinting at the satisfaction of repetitive needs. About 67% of users only used the Wi-Fi network once in the monitored period. Of the rest, Fig. 3 shows the distribution of the kinds of user visits classified as a function of the difference in days between two consecutive visits of the same user, and we observe the distribution of the return visits does not follow an uniform decreasing pattern, but a strong impact of a seven-day periodicity is captured in the data. We present this analysis in the Discussion section.

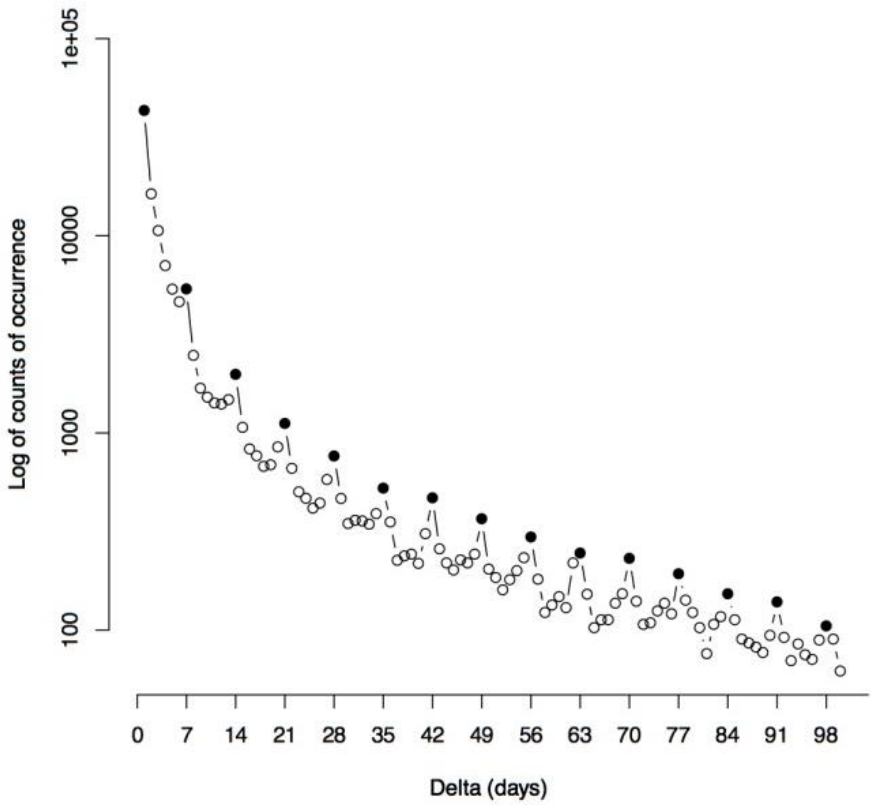


Figure 3. Counts of consecutive visits of all visitors binned by the Δ in days.

The trajectory length of users is on average 3.47 (expressed in number of APs associated with a visitor per day), with a range of 1-64 (mode=1, median=2. Note there are 67 APs). We observe around 28% of user visits accessed Wi-Fi at a single AP, and the majority (over 93%) of user visits associated with fewer than 10 APs overall.

The place of first association (identified by the AP ID) is not necessarily the same as the point of entry into the mall. We hypothesize visitors associate either when satisfying direct information needs (either mall related, e.g. price comparison; or generic, e.g. mail checking), or when filling time – eating, resting or waiting for acquaintances. The type of places where the users associate the Wi-Fi network for the first time is therefore informative.

Based on the floor maps, we manually classify the proximal areas of APs thus: Food-court (11 APs, around 16%), Retail (46 APs, around 69%), and Navigational (10 APs, around 15%: non-retail areas, e.g. near lifts, and toilets). For those APs with a context of both retail and navigational, the AP is classified based on which context area covers over 50% of the Wi-Fi signal coverage. Table 2 shows the distribution of first associations per context. While the majority (63%) of first associations are from a retail context, the number of first Wi-Fi associations per AP is higher in the food-court. Table 3 shows the distribution of visiting time per context. A similar trend to first associations is observed: 7% of users' visiting time is spent in navigational areas, 23% spent in food court, and the rest 70% spent in retail context. Again, the largest average duration per AP is measured in the food-court. In addition, from the average of visiting time per user visit, we observe indoor users tend to spend more time in *retail* context than other physical contexts.

Table 2. Context of first association

Context	% of starting association
Food-court	31% (2.84% per AP)
Retail	63% (1.37% per AP)
Navigational	6% (0.60% per AP)
Total	100% (1.49% per AP)

Table 3. Context in relation to visiting time, as a proportion of all association time spent at a given category of AP (and per AP within category), as well as average time.

Context	% of assoc. time	Avg. time per visit [h]
Food-court	23% (2.06% per AP)	1.39
Retail	70% (1.52% per AP)	2.29
Navigational	7% (0.68% per AP)	1.00
Total	100% (1.49% per AP)	2.77

Examining social contexts, we identified 2,705 accompanied user visits, coming from 2,358 individual users, with the size of groups ranging 2-14. The majority (78%) of such visits are composed of 2 users, 15% are composed of 3 users, 4% are composed of 4 users and only 3% are composed of 5 or more users. In the following section, we analyze how accompanying users actively access Web content. Identified users who only appear in the AL and not the BL are excluded from this analysis, leaving 2,174 accompanied user visits from 1,886 individual users.

Basic indoor Web behavior

The average Web access duration is around 40 minutes, 82% of users accessed the Web for less than an hour. Note the contrast with the distribution of physical visiting time (AL), which showed 66% of users stayed in the mall between 3-4 hours.

Fig. 4 shows the average BL Web duration, the average AL duration (in range of 0 to 4 hours in hourly bins) and the ratio between these visit durations. While the physical durations of visits (AL) in the mall vary widely, BL durations are much more constrained in extent. On average, a user accessed the Web for less than 1 hour during a single visit, resulting in a decreasing ratio between the BL duration and the AL duration in a visit. This indicates indoor users are likely to spend a relatively constant amount of time browsing the Web (less than 1 hour), although this period may be fragmented into a number of Web browsing sessions (the average number of sessions per visit is 1.32).

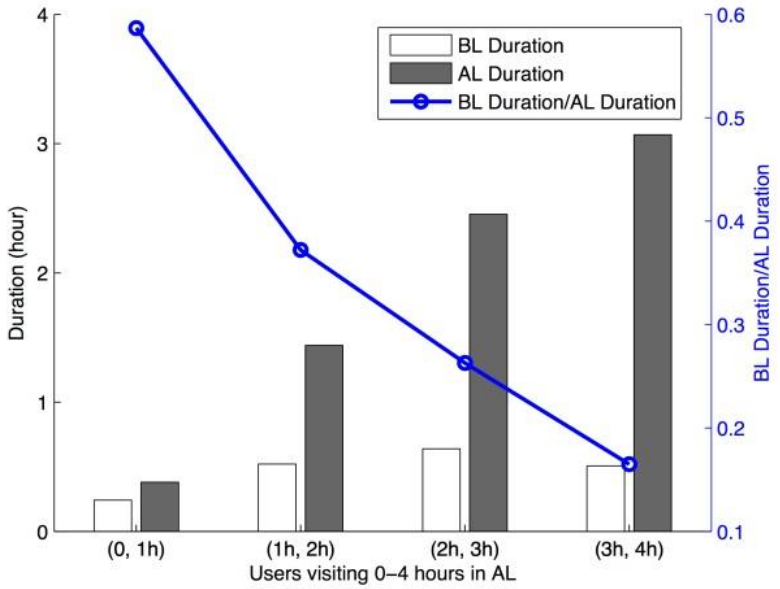


Figure 4. The average BL/AL duration of users visiting 0-4 hours in AL

We now analyze what users browse and search for in the mall. A categorization of Website content (captured by URLs, which was briefly discussed in (Ren et al., 2014)) was performed using BrightCloud. Specifically, *Social Networking* is the most popular browsing category (20%), consistent with overall mobile Web usage (Church & Oliver, 2011). *Content Delivery Networks* (aiming to improve the performance of Web services, e.g., akamaihd.net) and *Computer and Internet info* (e.g., amazonaws.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%.

We further investigate what users search for in the mall by analyzing Google search results that were followed by the users (query-click). Specifically, browsing categories are derived from all URLs in BL while Searching categories are from the click through from Google's Search Engine Results Page (SERP) in BL. *Travel* is the most popular category for Searching but only accounts for 1.4% in Browsing; *Social Networking* takes 20% in Browsing but only 6% in Searching.

Overall, around 80% of indoor Web browsing and searching URLs come from the top 20% of Web categories – showing a typical long-tail distribution characteristic.

Analyzing Indoor Behavior

Commonly accessed web content

We use the concept of entropy to quantify the commonality of a Website category in the Web behavior of users by measuring the *access entropy* across users.

For a URL category c_w, access entropy $H(c_w)$ is defined as:

$$H(c_w) = -\sum_{v \in S(c_w)} p(v|c_w) \log p(v|c_w), \quad (1)$$

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. 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 subset of users. Fig. 5 shows the distribution of $H(c_w)$. *Computer and Internet Info*, *Social Networking* and *Search Engines* are common URL categories with entropies of 10.75, 10.72 and 10.50, respectively. We observe there are some categories of Websites that are more commonly visited than others, and given the x-axis ($H(c_w)$) of Fig. 5 is on a log (bits of entropy) scale, we conclude there is a small number of categories that dominate what user access on the Web.

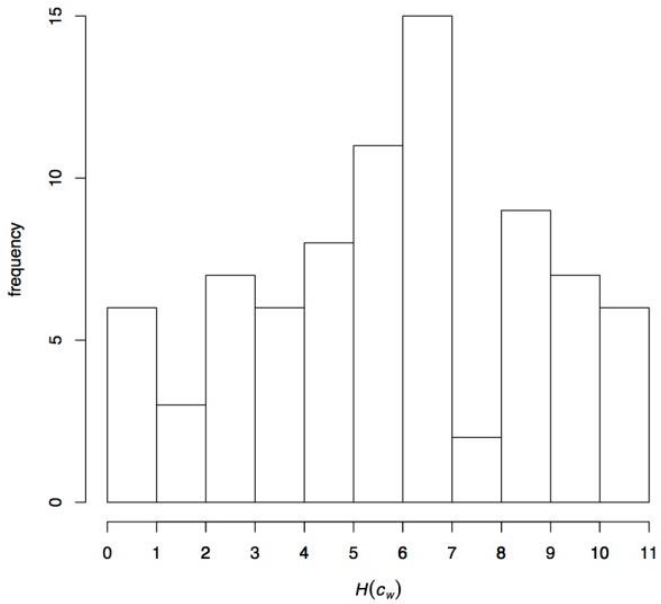


Figure 5. Binned distribution of access entropy $H(c_w)$

Some categories are commonly issued by a large proportion of users during many user visits, but they are not high in absolute numbers in the overall URL traffic. For example, the category *Shareware* & *Freeware* covers URL requests to Web pages containing screensavers, icons, wallpapers, utilities and ringtones. These are commonly accessed (high $H(c_w)$). However, the absolute number of requests to such Web URLs is low.

Examining repetitive patterns

We investigate the stability of users' indoor behavior during consecutive visits. To measure the strength of the correlation of the physical behavior $\mathbf{P}_{v_i}$ of a user during a current visit v_i with $\mathbf{P}_{v_j}$ during a consecutive visit v_j (further called *repeat model*), we compute the Pearson Correlation Coefficient (PCC) over consecutive visits as:

$$PCC_{phy} = r\left(\mathbf{P}_{v_i}, \mathbf{P}_{v_j}\right) = \frac{\sum_{a_k \in A}(p_{ik} - \bar{p}_i)(p_{jk} - \bar{p}_j)}{\sqrt{\sum_{a_k \in A}(p_{ik} - \bar{p}_i)^2 \sum_{a_k \in A}(p_{jk} - \bar{p}_j)^2}}, \quad (2)$$

where $\bar{p}_i$ and $\bar{p}_j$ are the average duration a user spends at each AP for visit v_i and v_j, respectively. A high positive PCC value indicates a strong correlation in physical behavior during consecutive visits. We apply

two baselines to compare with the repeat model: a *random*-paired baseline and an *average* baseline: the former replaces each repeated $\mathbf{P}_{v_j}$ with a randomly selected visit, and the latter replaces with the average physical behavior $\bar{\mathbf{P}}_v$ over all user visits[3]:

$$\bar{\mathbf{P}}_v = \frac{\sum \mathbf{P}_v}{|V|},$$

where V is the set of user visits and $|V|$ is the number of the user visits.

Table 4 shows the PCC values for *repeat*, *random*, and *average* models. We observe the *repeat* model achieves the largest PCC value, which is over two times larger than that of *average* and over twenty times larger than that of *random*. We have analyzed the variance between the means of the *repeat*, *random*, and *average* models through ANOVA and conclude that the differences are statistically significant with a p-value of < 0.0001. This indicates users' physical indoor behavior is repetitive and does not change substantially between two consecutive visits. It also demonstrates visitors return to the same parts of the mall and spend similar amounts of time in them.

Table 4. PCC values of trajectories

	Repeat	Random	Average
All	0.2534 (±0.3922)	0.0123 (±0.1506)	0.1108 (±0.1585)

Similarly, we apply PCC to measure the correlation in Web behavior between two consecutive visits v_i and v_j,:

$$PCC_{web} = r\left(\mathbf{W}_{v_i}, \mathbf{W}_{v_j}\right).$$

Again, we define another two baselines: the *random*-paired baseline, which replaces $\mathbf{W}_{v_j}$ with a randomly selected visit, and the *average* baseline, which replaces $\mathbf{W}_{v_j}$ with the average Web behavior $\bar{\mathbf{W}}_v$:

$$\bar{\mathbf{W}}_v = \frac{\sum \mathbf{W}_v}{|V|},$$

where V is the set of user visits and $|V|$ is the number of the user visits.

The first row of Table 5 shows the PCC results when all Web categories are considered, including those with a high $H(C_w)$. We observe consecutive visits achieve the highest score $r = 0.5902$, which means they are highly similar; *average* follows with $r = 0.5068$ while *random* only reaches $r = 0.2647$. To show the positive correlation between Web accesses in consecutive visits more clearly, we gradually remove common Web categories by setting a threshold for $H(C_w)$[4], and then re-calculate the above experiments (Table 5). The gap between *repeat vs. random* and *repeat vs. average* increases when the common Web categories are gradually removed. A two-tailed, paired t-test was applied to evaluate whether the differences are statistically significant (Table 6). It indicates the PCC values for *repeated* visits are statistically larger than both that for *random* and *average*.

Table 5. PCC values of browsing log (over Brightcloud category) for consecutive visits, random paired visits and between (each visit, average visit profile)

$H(c_w)$	Repeat	Random	Average
$H(c_w) \leq max(H(c_w))$	**0.5902**	0.2647	0.5068
$H(c_w) \leq 10$	**0.4581**	0.0922	0.3010
$H(c_w) \leq 9$	**0.4311**	0.0694	0.2632
$H(c_w) \leq 8$	**0.5261**	0.0287	0.1875
$H(c_w) \leq 7$	**0.4940**	0.0236	0.1505
$H(c_w) \leq 6$	**0.6526**	0.0483	0.2422
$H(c_w) \leq 5$	**0.7986**	0.1096	0.2093

Table 6. Paired *t*-test results for PCC values of browsing log comparison

Methods	Paired-t statistics	
	t	p-value
Repeat vs. Random	8.2	< 0.0001
Repeat vs. Average	4.545	0.0007

We have also examined the browsing differences between different visiting periodicities. We find, as time between revisits increases, there is decay in the likelihood of users repeating what they looked at online compared to last time. The PCC values degrade from around 0.63 for a periodicity of one day to about below 0.55 for a periodicity of 6 days, with a small increase in around 7 days.

Spatial context & information behavior

There are differences in the categories of shops served by different Wi-Fi APs (the association being done using the Voronoi regions). We hypothesize the proximity of different shop categories (the indoor context) leads to a different Web behavior of the mall visitors. At the level of Wi-Fi APs, the influence of spatial context on users' Web behavior can be viewed as the correlation between $\mathbf{B}_i$ and $\mathbf{B}_j$ for every two APs. We again apply PCC to test this association:

$$r(\mathbf{B}_i, \mathbf{B}_j) = \frac{\sum_{c_w^k \in C_w}(b_{ik} - \bar{b}_i)(b_{jk} - \bar{b}_j)}{\sum_{c_w^k \in C_w}(b_{ik} - \bar{b}_i)^2 \sum_{c_w^k \in C_w}(b_{jk} - \bar{b}_j)^2}, \quad (3)$$

where 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.

To test the above hypothesis, we apply a clustering algorithm to group similar APs based on shop categories. From *definition* 4, an AP a_i is represented by a vector $\mathbf{E}_i$ of shop categories. If the users' information behavior is influenced by their indoor context, the users' Web behavior *within* a cluster should be *similar* and the users' Web behavior *between* clusters should be *different*. 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 (Davies & Bouldin, 1979).

To test the association, we apply PCC to measure the similarity between the Web behavior at two APs. The *intra-cluster* similarity (*within*) and the *inter-cluster* similarity (*between*) are defined as follows:

$$within = \frac{1}{k}\sum_{x=1}^{k}\left(\frac{2}{|t_x|(|t_x|-1)}\sum_{\mathbf{B}_i \in t_x}\sum_{\mathbf{B}_j \in t_x, i \neq j} r(\mathbf{B}_i, \mathbf{B}_j)\right), \quad (5)$$

$$between = \frac{1}{k}\sum_{x=1}^{k}\left(\frac{1}{|t_x|(|\mathfrak{B}|-|t_x|)}\sum_{\mathbf{B}_i \in t_x}\sum_{\mathbf{B}_j \notin t_x} r(\mathbf{B}_i, \mathbf{B}_j)\right), \quad (6)$$

where $\mathfrak{B}$ denotes the set of user Web behavior, and $|\mathfrak{B}|$ denotes the size of $\mathfrak{B}$, and k is the number of clusters, t_x denotes the x-th cluster, and $|t_x|$ denotes the size of t_x. We emphasize that the groups of APs are clustered based on their physical context information $\mathcal{E}$, but the r value is defined based on user's Web behavior $\mathfrak{B}$. Hence, the user's Web behavior 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. The *average* r for all $\mathbf{B}_i$ pairs is also applied as another baseline and it is defined as:

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

We also examine the influence of the coarser indoor contexts (food-court, retail, and navigational) on users' Web behavior. Specifically, we treat this as *places*-based clustering results and calculate the corresponding *within* and *between*.

Table 7 shows the experiment results and Table 8 the results of the analysis (a two-tailed, paired t-test). We observe: the *within* of k-means is significantly larger than the *between* of k-means, the *within* of *random* and *places*-based methods, and the *average*; the *within* of *places* is significantly larger than its *between* value; the *within* of *random* is not significantly different from its *between* value and the *average*. More importantly, the trends are all in the right direction: all context-based groups lead to higher *within* than its *between*. This indicates there is an influence from indoor spatial context on users' Web behavior.

Table 7. Correlation of user Web behavior in groups of APs with similar spatial context

	$H(c_w)$	PCC r value based on $\mathfrak{B}$						
		k-means		places		random		average
		within	between	within	between	within	between	
Groups of APs based on $\mathcal{E}$	$H(c_w) \leq \max(H(c_w))$	**0.9659**	0.9623	0.9617	0.9613	0.9609	0.9617	0.9619
	$H(c_w) \leq 10$	**0.8601**	0.8509	0.8401	0.8302	0.8493	0.8501	0.8498
	$H(c_w) \leq 9$	**0.7721**	0.7599	0.7540	0.7287	0.7564	0.7573	0.7573
	$H(c_w) \leq 8$	**0.6817**	0.6572	0.6804	0.6556	0.6493	0.6473	0.6483
	$H(c_w) \leq 7$	**0.6410**	0.5966	0.5950	0.5645	0.5767	0.5750	0.5770
	$H(c_w) \leq 6$	**0.5045**	0.4778	0.5001	0.4842	0.4755	0.4751	0.4763
	$H(c_w) \leq 5$	**0.4107**	0.3942	0.4004	0.3837	0.3821	0.3848	0.3863

Table 8. Paired *t*-test results

Methods	Paired-t statistics	
	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 $within$(places)	2.5497	0.0435
$within$(k-means) VS $average$	3.4126	0.0143
$within$(places) VS $between$(places)	4.5326	0.0040
$within$(random) VS $between$(random)	0.2526	0.8090
$within$(random) VS $average$	1.6007	0.1606

Finally, we examine what Web content indoor users accessed in different context. Around 70% of Web sites about *Swimsuits & Intimate Apparel*, *Fashion and Beauty*, *Alcohol and Tobacco*, *Financial Services*, and *Shopping*, are accessed in the retail context; around 50% of Web pages about *Kids*, *Home and Garden*, *Real Estate*, *Individual Stock Advice and Tools*, and *Sports*, are requested in the food-court section; *Dating*, *Search Engines*, *Social Networking*, *Web based email*, and *Fashion and Beauty* are popular services accessed by users in the navigational context.

Social context and Web access

To investigate what the accompanying users access on the Web, we measure the overlap of the accessed Web content captured through Web domains. For two accompanied user visits, v_i and v_j, we define

$$O_{social} = \frac{|D_{v_i} \cap D_{v_j}|}{|D_{v_i} \cup D_{v_j}|}, \tag{7}$$

where D_{v_i} is the set of Web domains that a user visit v_i accessed on the Web.

To show the influence of the social context we compare O_{social} with two baselines:

- $O_{physical}$: when computing the domain overlap as shown in Eq. 7, replace v_j with another random user visit, which is associated with exactly the same Wi-Fi APs associated by v_i. This baseline distinguishes the influence of the accompanying social context from that of the physical context.
- O_{random} : replace v_j with another random user visit when calculating the domain overlap defined in Eq. 7.

Table 9 shows the average values of O_{social}, $O_{physical}$ and O_{random} over various groups of accompanying users whose average distance is ≤ 1. We observe the domain commonality in accompanying users' visits is higher than that modeled by the baselines $O_{physical}$ and O_{random}.

Table 9. Overlap in accessed Web domains amongst members of a group

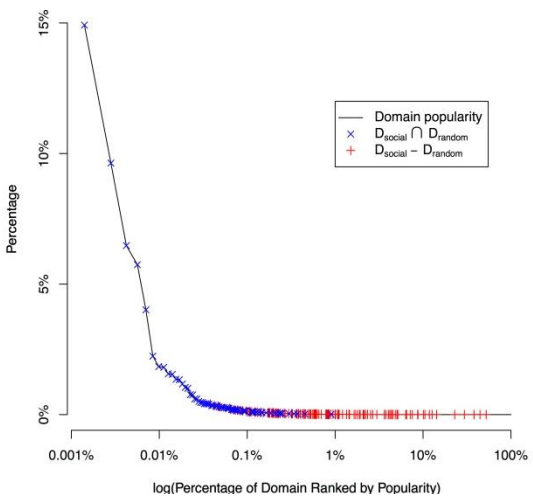
$d(v_i, v_j)$	O_{social}	$O_{physical}$	O_{random}
$d(v_i, v_j) = 0.0$	0.1868	0.1119	0.1031
$d(v_i, v_j) \leq 0.1$	0.1833	0.1173	0.1057
$d(v_i, v_j) \leq 0.2$	0.1780	0.1130	0.1054
$d(v_i, v_j) \leq 0.3$	0.1772	0.1139	0.1067
$d(v_i, v_j) \leq 0.4$	0.1717	0.1147	0.1060
$d(v_i, v_j) \leq 0.5$	0.1670	0.1173	0.1072
$d(v_i, v_j) \leq 0.6$	0.1635	0.1137	0.1084
$d(v_i, v_j) \leq 0.7$	0.1620	0.1175	0.1061
$d(v_i, v_j) \leq 0.8$	0.1628	0.1160	0.1061
$d(v_i, v_j) \leq 0.9$	0.1613	0.1162	0.1043
$d(v_i, v_j) \leq 1.0$	0.1614	0.1157	0.1049

Table 10 shows the paired-t test results among accompanied visits, *physical*-paired visits and *random*-paired visits, and we observe O_{social} is significantly larger than $O_{physical}$ and O_{random}; $O_{physical}$ is significantly larger than O_{random}, which confirms the influence of physical context. These indicate the accompanying social context significantly correlates with the Web content consumed during people's visits to the mall. Namely, visitors belonging to the same social group access similar content on the Web. Furthermore, we show this influence is not just an artifact of the joint physical context (proximity to the same shops).

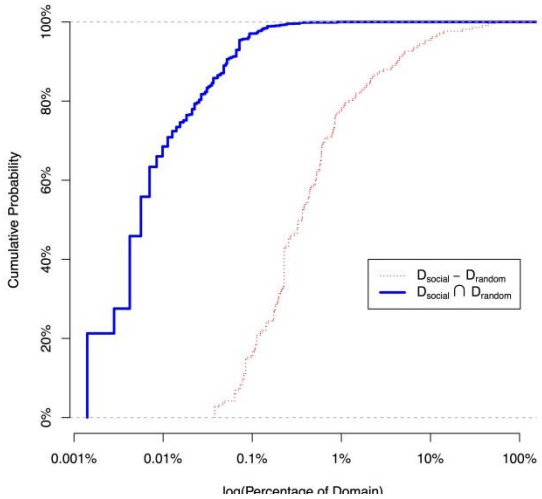
Table 10. Paired *t*-test results for overlap in domains

Methods	Paired-t statistics	
	t	p-value
O_{social} VS $O_{physical}$	19.3371	< 0.0001
O_{social} VS O_{random}	22.8111	< 0.0001
$O_{physical}$ VS O_{random}	13.1395	< 0.0001

Accompanying users are more likely to access the same Web content (domains). Although the extent of this overlap is not large, it is statistically significant. But is this content similar to the overall commonly accessed Web content of indoor users? We first examine the distribution of Web domains in the collected BL (Fig. 6a). The distribution of Web domains is highly skewed and has a long tail. Over 80% of the overall Web accesses go to less than 1% of overall Web domains in the collected data. This is expected following the discussion about basic indoor Web behavior. We investigate what are the commonly accessed Web domains from the accompanying users. Here, we define D_{social} as the union set of domains that are commonly accessed by every accompanied user visits, corresponding to O_{social}; and D_{random} as the union set of domains that are commonly accessed by an accompanied user visit and another randomly selected user visit, corresponding to O_{random}.



(a) the log plot of domain popularity and the distribution of $D_{social} \cap D_{random}$ and $D_{social} - D_{random}$ over it.



(b) the empirical CDFs of $D_{social} \cap D_{random}$ and $D_{social} - D_{random}$

Figure 6. The domain popularity and the relationship between D_{social} and D_{random}

Thus, $D_{social} \cap D_{random}$ reflects the domains that are commonly accessed by an indoor user regardless of whether they are accompanied or not, and $D_{social} - D_{random}$ reflects the domains that are shared among accompanying users but not non-accompanying users. Finally, we obtain $|D_{social}|$ = 208, $|D_{random}|$ = 88, $|D_{social} \cap D_{random}|$ = 70 and $|D_{social} - D_{random}|$ = 138.

The blue-cross points and the red-plus points in Fig. 6a show the distribution of $D_{social} \cap D_{random}$ and $D_{social} - D_{random}$ over the distribution of overall Web domains, respectively. We observe $D_{social} \cap D_{random}$ are composed of Top popular Web domains, namely Top 1%; $D_{social} - D_{random}$ are composed of unpopular Web domains, which mainly come from the 'tail' of the overall domain distribution. Furthermore, we examine the difference between the distributions of $D_{social} \cap D_{random}$ and $D_{social} - D_{random}$, as shown in Fig. 6b in terms of empirical CDFs. We observe their CDFs are different, and the Kolmogorov-Smirnov test has been applied to measure whether the differences are significant. The detailed result is (D = 0.8804, *p*-value < 0.0001), which means the differences are statistically significant. This indicates apart from accessing popular Web domains, the accompanying users tend to access some less popular Web domains which may be specific to their information needs.

Then, we examine what Web categories D_{social} includes. Around 65% of accompanying users used the same *Social Networks*; 25% of them accessed the same *Personal Storage websites*; 20% of them accessed the same *Web based email servers*. Note, two accompanying users may access Web domains from more than one category. These show accompanying users tend to have similar habits and needs in the mall.

Moreover, we find around 15% of accompanied user visits accessed at least one Website having the same country domain, which are over 26 countries. On contrary, only 3% of randomly paired users (corresponding to D_{random}) accessed the Websites over only around 5 countries. Note, for the domain au (Australia), we eliminate the effect of the host location of the investigated mall by limiting the Websites to local services (e.g., vodafone.com.au) other than localized well-known world-wide services (e.g.,

google.com.au). This indicates there is a good probability to see shopping accompanies having the same nationality, assuming people tend to use their native language in daily life.

Discussion

Temporal patterns of users visits

The analysis of the length of visits to the indoor environment shows an uneven distribution with the majority of visitors spending 3-4 hours in the mall, while visits shorter than 1 hour are common. Note, users' visit duration may be underestimated if their phones turn to sleep mode and disassociate from the Wi-Fi network before they physically leave the mall. We plan to perform a user study to investigate the impact of smartphone sleep mode on access logs in future work. The likelihood of a user returning to the mall is higher if the time difference since the last visit is proportionate to a weekly pattern or its multiple. These two patterns may point to different *purposes* of the visits to the retail environment and the related nature of the indoor physical behavior. The trip may be related to the satisfaction of repetitive needs, further emphasized by a preference for a specific day of the week for shopping when conducting goal-oriented regular shopping trips. Additional rounding on the periodicity capturing individual flexibility in the choice of the day of the week for the shopping trip may emphasize this effect. Less regular shoppers may visit on an ad-hoc basis related to an activity satisfying rare needs. These patterns may prove useful for the detection of customer groups. A first venture in this direction is the analysis of the locations that users visit during repeat visits. We show the closer together visits are in the sequence, the more similar their pattern is likely to be. Fig. 7 shows the movement and the accessed Web content in two consecutive visits of a sample user. Combined with a deeper analysis of social shopping contexts and shopper groups, our future research will focus on the predictability and characterization of these groups.

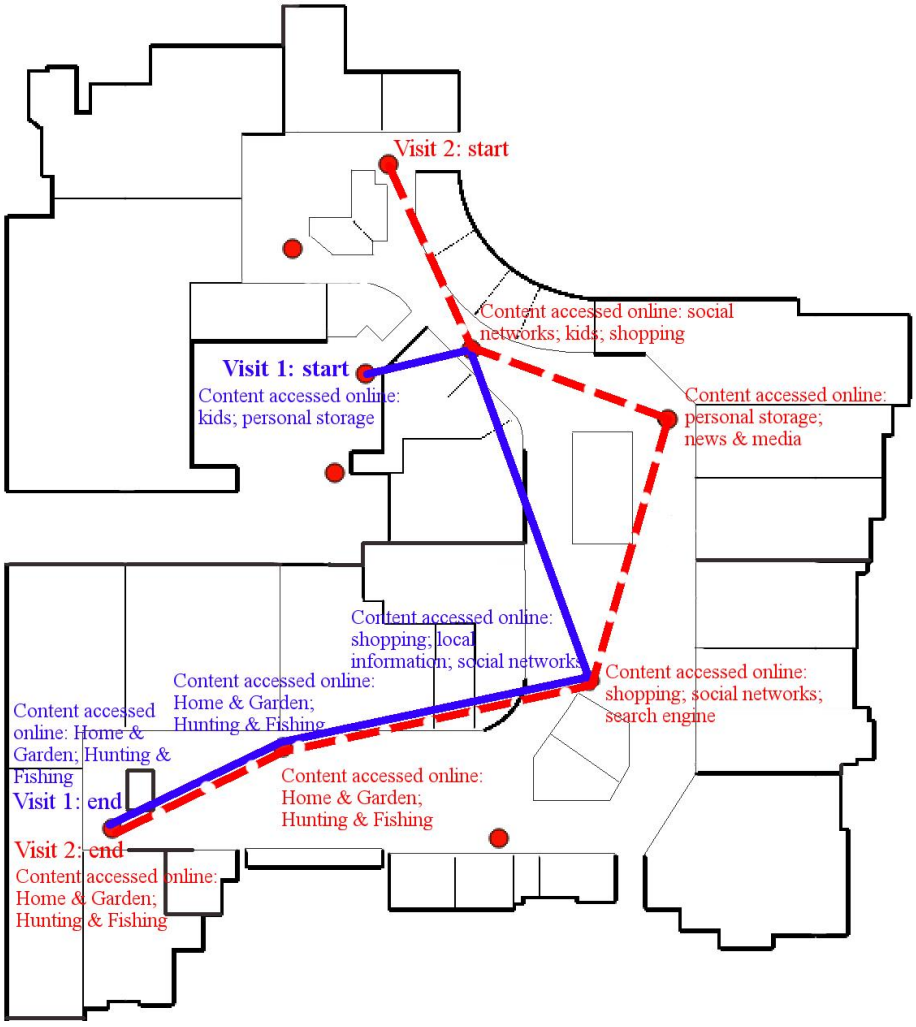


Figure 7: The movement and Web use pattern during two consecutive visits of a sample user. The blue (solid) line/text show the movement (in terms of Wi-Fi APs) and the accessed Web content (in terms of BrightCloud categories) for visit 1, while the red (dashed) line/text show that for visit 2.

Spatial patterns of indoor visits

Recall that the visitors are only logged if actively interacting with their mobile devices. The short length of indoor trajectories detected might indicate that indoor visitors use Wi-Fi in a relatively static manner, for instance while eating at a food court (phones enter sleep mode when not actively used during walking or shopping). Food courts are also locations of the longest average connection duration per AP and proportionally they are also the most likely place of first association. While about 70% of AP accesses occurs in the normal retail context (naturally as most of the mall is dedicated to this function), a high proportion of Web use occurs in the food-court context. The length of association with APs in the food-court context is also high (on average 1.39 hours). We conclude that the food-court context in shopping malls has a dominant role in the visitors' Web behavior and it is therefore critical to improve the quality of services in this part so as to satisfy users' information needs better.

Web content use and context dependence

We find, for the different groups of visitors grouped by the length of time spent in the mall, the amount of time spent on the Web does not vary much and is less than an hour for about 82% of visitors. This may

indicate, for the majority of indoor visitors, accessing the Web is not the primary activity pursued in the mall and that their information needs satisfied via Web require a relatively constant amount of time, independent of the total amount of time spent in the mall.

There is a pronounced difference between the Web content browsed and searched, with a dominant representation of social networking services *browsed to*, rather than *searched for*. We further show that once common Websites are filtered out (the top 5 common URL categories take over 57.8% of the overall URL records), the Web behavior of the visitors reveals strong contextual dependence. Compared to a baseline generated by random and average models, the *repeat model* (taking into account the Web content accessed in the previous visit) allows for a substantial improvement in content prediction in the consecutive visit. Thus, content access is correlated in time and space, with different Web content accessed in different parts of the mall, as well as different parts of the mall with the same *context* (mixture of shop categories) inciting users to consume similar Web content. Finally, we have demonstrated how visitors belonging to the same social group have a Web behavior biased to a larger proportion of joint Web content consumed within the mall. Note, the 90% threshold, which is used to identify users' social contexts, may underestimate the number of accompanying users if some of the phones are in sleep mode; we use this high time-based overlap as a constraint to ensure a good precision of identifying accompanying users.

We have thus shown that the visitor's Web and physical behaviors are predictable and highly contextualized. The discovered behaviors can be modeled beyond individual visitors, such as in groups of visitors that can be detected purely based on their spatio-temporal characteristics.

Conclusions and Future Work

Based on a large data set collected over a one year period through an opt-in public Wi-Fi network of a large urban shopping mall in Australia, we present an analysis of how people use the Web in the context of indoor retail spaces.

The study established the extent of the predictability of contextualized indoor information behavior, a first step towards visitor modeling. The patterns in suburban shopping malls or in malls in other countries may differ. The study also raised many new research questions, e.g. How to improve users' Web experience in the context of indoor retail spaces? What are the specific differences in indoor users' Web behaviors in two kinds of indoor contexts? Can the differences in Web behavior help to identify the spatial context of user preferences, and can this knowledge be utilized further to provide contextual preference-aware recommendations to satisfy user needs? How do different groups of users behave in indoor retail spaces? We plan to do a supervised user classification to discover how shopping behaviors change between individuals or their groups. We hope we contributed to a better understanding of people's indoor information behavior in retail environments. As over 80% of shoppers check the price online before purchase (Regalado, 2013), and 27% of smartphone users do research while in store, a better understanding of indoor information behavior can help improve services to shoppers.

Acknowledgements

This research is supported by a Linkage Project grant of the Australian Research Council (LP120200413).

Footnote:

1. Scrambled is applied, where a temporary database is stored for the pre- and post- strings, and is deleted after the anonymization is done.
2. There are other WCCS, such as DMOZ, but our testing found that its coverage was too narrow for our study. E.g., the highly popular Australian classifieds Website *www.gumtree.com.au* is not categorized in DMOZ but categorized as *shopping* by *BrightCloud*.
3. All random processes in this research are repeated ten times, and averaged.
4. When $H(c_w) \leq 4$, some W_i become empty, which renders the calculation of PCC_{web} undefined. So, we analyzed in the cases when $H(c_w) > 4$.

References

Absar, R, O'Brien H, & Webster E. (2014). Exploring Social Context in Mobile Information Behavior. In ASIS&T'14.

Bates, M. J. (2010). Information Behavior. In Encyclopedia of Library and Information Sciences, 3, 2381–2391.

Biczok, G., Martinez, S., Jelle, T., & Krogstie, J. (2014). Navigating MazeMap: indoor human mobility, spatio-logical ties and future potential. CoRR.

Church, K., & Oliver, N. (2011). Understanding Mobile Web and Mobile Search Use in Today's Dynamic Mobile Landscape. In Mobilehci'11 (pp. 67–76).

Church, K., & Smyth, B. (2009). Understanding the intent behind mobile information needs. In IUI'09.

Church, K., Smyth, B., Cotter, P., & Bradley, K. (2007, May). Mobile information access: A study of emerging search behavior on the mobile Internet. ACM Trans. Web, 1(1).

Cui, Y., & Roto, V. (2008). How people use the web on mobile devices. In WWW'08 (pp. 905–914).

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. IEEE Trans. Pattern Anal. Mach. Intell., 1(2), 224–227.

De Domenico, M., Lima, A., & Musolesi, M. (2013, December). Interdependence and predictability of human mobility and social interactions. Pervasive Mob. Comput., 9 (6), 798–807.

Evans, K. R., Christiansen, T., & Gill, J. D. (1996). The impact of social influence and role expectations on shopping center patronage intentions. Journal of the Academy of Marketing Science, 24(3), 208–218.

Hodkinson, C., Kiel, G., & McColl-Kennedy, J. R. (2000, May). Consumer web search behaviour: Diagrammatic illustration of wayfinding on the web. Int. J. Hum. Comput. Stud., 52 (5), 805– 830.

Jansen, B. J. (2000, March). Real life, real users, and real needs: a study and analysis of user queries on the web. Information Processing and Management, 36(2), 207–227.

Jensen, C. S., Lu, H., & Yang, B. (2010). Indoor-a new data management frontier. IEEE Data Engineering Bulletin, 33(2), 12–17.

Kamvar, M., & Baluja, S. (2006). A large scale study of wireless search behavior: Google mobile search. In CHI'06 (pp. 701–709).

Kjærgaard, M. B., Krarup, M. V., Stisen, A., Prentow, T. S., Blunck, H., Grønbæk, K., & Jensen, C. S. (2014). Indoor positioning using wi-fi–how well is the problem understood? In IPIN'13.

Misra, A., & Balan, R. K. (2013, December). Livelabs: Initial reflections on building a large-scale mobile behavioral experimentation testbed. SIGMOBILE Mob. Comput. Commun. Rev., 17(4), 47–59.

Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on web usage mining. Communications of the ACM, 43(8), 142–151.

Okabe, A., Boots, B., Sugihara, K., & Chiu, S. N. (1999). Spatial tesselations: Concepts and applications of voronoi diagrams (2nd ed.) John Wiley and Sons.

Patel, M, O'Kan A. (2015). Contextual Influences on the Use and Non-Use of Digital Technology While Exercising at the Gym. In CHI'15.

Regalado, A. (2013). It's all e-commerce now [Electronic Book Section].

Ren, Y., Tomko, M., Ong, K., & Sanderson, M. (2014). How people use the web in large indoor spaces. In CIKM'14 (pp. 1879–1882).

Richter, K.-F., Winter, S., & Santosa, S. (2011). Hierarchical representations of indoor spaces. Environment and Planning-Part B, 38(6), 1052.

Ruetschi, U.-J. (2007). Wayfinding in scene space: Transfers in public transport. Phd. dissertation.

Sen, S., Chakraborty, D., Subbaraju, V., Banerjee, D., Misra, A., Banerjee, N., & Mittal, S. (2014). Accommodating user diversity for in-store shopping behavior recognition. In ISWC'14 (pp. 11–14).

Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999). Analysis of a Very Large Web Search Engine Query Log. In ACM SIGIR forum (pp. 6–12).

Spink, A., Wolfram, D., Jansen, M. B., & Saracevic, T. (2001). Searching the Web: The Public and Their Queries. J. Am. Soc. Inf. Sci. Technol., 53(3), 226–234.

Teevan, J., Karlson, A., Amini, S., Brush, A. J. B., & Krumm, J. (2011). Understanding the importance of location, time, and people in mobile local search behavior. In Mobilehci'11 (pp. 77–80).

Vernor, J. D., Amundson, M. F., Johnson, J. A., & Rabianski, J. S. (2009). Shopping Center Appraisal and Analysis.

West, R., White, R. W., & Horvitz, E. (2013). From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. In WWW'13 (pp. 1399–1410).

Wiener, J. M., Büchner, S. J., & Hölscher, C. (2009). Taxonomy of human wayfinding tasks: A knowledge-based approach, Spatial Cognition & Computation, 9(2), 152-165.