This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2017.2766059, IEEE Transactions on Knowledge and Data Engineering

A Location-Query-Browse Graph for Contextual Recommendation

Yongli Ren, Martin Tomko, Flora Salim, Jeffrey Chan, Charles L.A. Clarke, Mark Sanderson

Abstract—Traditionally, recommender systems modelled the physical and cyber contextual influence on people's moving, querying, and browsing behaviours in isolation. Yet, searching, querying and moving behaviours are intricately linked, especially indoors. Here, we introduce a tripartite location-query-browse graph (LQB) for nuanced contextual recommendations. The LQB graph consists of three kinds of nodes: locations, queries and Web domains. Directed connections only between heterogeneous nodes represent the contextual influences, while connections of homogeneous nodes are inferred from the contextual influences of the other nodes. This tripartite LQB graph is more reliable than any monopartite or bipartite graph in contextual location, query and Web content recommendations. We validate this LQB graph in an indoor retail scenario with extensive dataset of three logs collected from over 120,000 anonymized, opt-in users over a 1-year period in a large inner-city mall in Sydney, Australia. We characterize the contextual influences that correspond to the arcs in the LQB graph, and evaluate the usefulness of the LQB graph for location, query, and Web content recommendations. The experimental results show that the LQB graph successfully captures the contextual influence and significantly outperforms the state of the art in these applications.

Index Terms—location-query-browse graph, contextual recommendation, query log analysis, information retrieval

1 Introduction

Studying users' behavioural patterns captured in mobile access logs enables the understanding of users's intent and the provision of personalised information and services. Research to date, however, focused merely on analyzing individual aspects of behaviour in isolation, e.g. Web site browsing or querying for studying cyber behaviour, and Wi-Fi associations for studying physical behaviour. This largely limits the quality of modelling in terms of provided services.

In this paper, we introduce a tripartite location-query-browse graph to address this gap and capture linked influences. We propose the location-query-browsing (LQB) graph as a representation of the interactive knowledge about people's behaviour across the physical and cyber spaces. The LQB graph contains three kinds of nodes, representing locations, queries and Web domains. The LQB graph thus models the physical and cyber contextual influence on people's moving, querying, and browsing behaviours. It contains arcs between heterogeneous nodes only, and is used to infer connections between homogeneous nodes from corresponding contextual influences. We evaluate this LQB graph in an indoor retail scenario with three types of logs: a Wi-Fi access point association log that records users' physical movement, a Web browsing log, and a query log containing users' interaction with search engines. These logs were collected from over 120,000 (anonymized, opt-in) users for over one year in a large inner-city mall in Sydney, Australia. We characterise the corresponding physical and cyber contextual influence captured in these logs and examine the usefulness of the LQB graph in three applications: location recommendation, Web content recommendation, and query recommendation.

The main contributions of the paper are: (1) A formalisation of the LQB graph model, a concise representation of user behavior across the physical and cyber spaces; (2) A comprehensive analysis of the physical and cyber contextual influence on people's moving, querying, and browsing behaviours in an indoor retail space; and (3) The application of the LQB graph model to location, Web content and query recommendation in this retail space.

- *Yongli Ren, Flora Salim, Jeffrey Chan and Mark Sanderson, are with School of Science, RMIT University, Melbourne, Australia. E-mail: yongli.ren@rmit.edu.au; flora.salim@rmit.edu.au; jeffrey.chan@rmit.edu.au; mark.sanderson@rmit.edu.au*
- *Martin Tomko is with Department of Infrastructure Engineering, the University of Melbourne, Australia. E-mail: tomkom@unimelb.edu.au*
- *Charles L.A. Clarke is with the School of Computer Science, University of Waterloo, Waterloo, ON, Canada. Email: claclark@plg.uwaterloo.ca*

2 Related Work

2.1 Query Logs and Browsing Logs

Query logs record rich data about users' behaviour patterns that can be mined for information about immediate interests and preferences. Two early Web search studies on traditional desktop-based queries are the Excite study [1]–[5] and the AltaVista study [6]. They summarised key characteristics of Web search queries, including the number, type, and distribution of terms in queries, queries per session and session distributions, the use of advanced search features, and interaction with retrieved results. Subsequent studies also examined geographic queries [7]–[9], religious information in search engines [10], and sponsored search [11].

Recent Web search analysis shifted focus to mobile query logs. One of the earliest studies examined the queries from Google's two mobile search interfaces at the time [12]. They analysed the key characteristics in mobile queries, e.g., query length and distribution, session length and click through rates. While the average number of terms per query was similar to desktop queries, the average number of queries per session (around 1.6) differed significantly

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2017.2766059, IEEE Transactions on Knowledge and Data Engineering

(2.02 [6], 2.3 [1] and 2.84 [5]), and nearly 70% of sessions included only one query. Adult-related queries were popular in one Google search interface but not other; the authors suspected the reason was different user demographics.

An early European study on mobile search compared mobile browsing and mobile searching [13]. They found that while browsing was still dominating the mobile information access, searching was gaining in popularity, but mobile queries were shorter than desktop queries, with about half (45%) of the query sessions consisting of a single query. We note that these results come from the time when mobile search and mobile search interfaces were still at an early stage. A later study from the same authors [14] highlighted the key characteristics of mobile search, revealing that almost 90% of searches fail to attract any user clicks on the retrieved results, and that adult-related queries still dominate search activity.

Like query logs, browsing logs contain rich information about users' browsing behaviours on the Web. We briefly review only studies of browsing logs related to search activities. Agichtein et. al. [15] incorporated users' browsing behaviour as implicit feedback to improve Web search ranking, and suggested that they augment other query-relevant factors and improve rankings. Liu et. al. [16] investigated the transitions between pages in users' browsing history, and thus computed Web page importance. They suggested that browsing-based models outperform link-based ranking. White and Drucker studied post-query browsing trails and found dramatic differences in variability in users-engaged Web search activities. Later [17], they demonstrated that these post-query trails provide users benefits in terms of coverage, diversity, novelty, and utility over origins (landing pages) and destinations (pages where trails end). White et. al. [18] further suggested that people's general browsing behavior as recorded in a browsing log far outweigh direct search engine interaction as an information-gathering activity. Tasagkias and Blanco [19] also found that textual features of articles browsed by users to be useful for article recommendations. Chiarandini et. al. [20] studied the browsing patterns on social photo sites, and Chiarandini et. al. [21] used browsing patterns for topic discovery and photostream recommendations. Trevisiol et. al. [22] studied image ranking and user browsing behaviour by exploring both internal and external factors (e.g. links within and outside *Flickr*), and quantified the impact of these factors.

2.2 Context Modelling

Studies exploring a particular aspect of mobile Web access–contextual dependence–followed. An early study on contextual influence in mobile search was conducted by [23]. Consecutively, Sohn et. al. [24] conducted a two-week diary study involving twenty participants, that found that around 72% of the participants' mobile information needs were prompted by contextual factors. Hinze et. al. [25] performed another small-scale diary study, in which participants were required to record their location, time, information needs, and how much their needs were related to the current location and time. Contextual factors strongly influenced needs, e.g. location, conversion, and activity; the type of asked questions varied across locations. To infer context, they found the query key words were not sufficient.

Teevan et. al. [26] performed a similar study on a larger scale and found that mobile local searches were highly influenced by contextual parameters, such as geographic features, temporal aspects, and searchers' social context. Chua et. al. [27] examined context factors finding that location, intended activity, and social surroundings triggered information needs while location, time, current activity, and social surrounding influenced information needs. Song et. al. [28] compared the differences between searches on mobile phones and tablets in terms of search location distribution and found that mobile phone users searched the Web at a variety of different locations while tablet users mainly searched from home. Exploiting GPS sensors in mobile phones, Lymberopoulos et. al. [29] studied the influence of location on local search issued by US users using a dataset of two million queries. They analysed mobile click behaviour across different spatial scales, e.g. city, state and country, and introduced location-aware features to improve local search click prediction by encoding information from the ZIP code where the query was issued.

Yom-Tov and Diaz [30] investigated the influence of social and physical detachment on users' information needs and demonstrated how to use these factors to improve retrieval results. Chiarandini et. al. [20] described the influence of the Websites they arrived from on their usage of social image site. Recently, Zhang et. al. [31] found that the installed apps might indicate users preferences in sports, business or other fields, and proposed an application-aware approach for query auto-completion, which shows improved accuracy on mobile devices.

There are also some research focusing on associating locations to queries or Web content. For example, Zhuang et. al. [32] proposed to exploit the geographical probability distributions of user clicks to infer locality information for queries, and found this leads to better results in terms of query classification. Zong et. al. [33] attempted to determine the spatial semantics to Web pages by assigning them place names. Backstrom et. al. [34] studies the spatial variation in search engine queries, and proposed a probabilistic model to determine the query's geographic center and its spatial dispersion by utilizing geolocation techniques to find locations of IP addresses where queries are issued. However, they are different from the focus on this paper: modelling the contextual influence among locations, queries and Web content for contextual recommendations.

2.3 Graph Representations of Querying, Browsing and Physical Associations

Directed graph representations of sequential activity (whether it is sequential querying activity, browsing activity or movement through space) have gained prominence in for their expressive power and mathematical grounding in graph theory. Graph representations of the query activity, browsing activity and physical acitivty can be built in a large number of ways, depending on the formalisation of the graph nodes and arcs (oriented edges).

The query graph is a compact representation of the information about user querying behaviours extracted from the query log. There are several kinds of query-based graphs, depending on how the nodes of the graph and their associations are represented. These include *query-query* graphs,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2017.2766059, IEEE Transactions on Knowledge and Data Engineering

and *query-click* graphs that have two kinds of nodes, *queries* and *urls* (or documents), respectively. Baeza-Yates et. al. identified five types of *query-query* graphs [35], including 1) *word graph*: connects queries having the same word(s); 2) *session graph*: connects queries in the same session; 3) *URL cover graph*: connects queries by which users clicked on the same url; 4) *URL link graph*: connects queries whose clicked urls are linked; 5) *URL terms graph*: connects queries whose clicked urls have common terms. They suggested these graphs can be used in the following applications, e.g. recognition of polysemic words, as well as related and similar queries.

Zhang et. al. used a one dimensional graph to model consecutive queries and applied a damping factor to weight the arcs between them [36]. They defined the similarity of two queries as the multiplication of the values of the arcs that join them. If they are consecutive, the similarity will be the damping factor. They finally combined this graph-based similarity with the content-based similarity to do query recommendations. Boldi et. al. proposed a *query-flow* graph, which is a *query-query* graph by connecting two consecutive queries in a session [37]. They built the query-flow graph by mining the time, textual information of the queries, and suggested that the *query-flow* graph is helpful for finding logical sessions and query recommendations [38], and query similarity measurement [39]. Albakour et. al. enriched the *query-flow* graph model by utilizing clickthrough information to adjust the arc weights, and found this modified graph was more valuable than the standard *query-flow* graph for query recommendations [40].

Another popular query-related graph is *query-document* graph, built by using query clickthrough data. A *query-document* graph is a bipartite graph with two types of nodes: queries and documents. A link is introduced if there is a query that is submitted and a corresponding document is clicked by a user. Based on a *query-document (urls)* graph representation, Beeferman and Berger [41] proposed an agglomerative clustering approach to discover similar queries and similar URLs corresponding to similar needs. Craswell and Szummer [42] proposed a Markov random walk model on the *query-document* graph to rank documents for a given query in a image search engine, and demonstrated that the proposed model is effective on ranking those un-clicked documents. Mei et. al. [43] proposed a query recommendation approach based on query rankings with hitting times. The authors define hitting times as the first time that a random walk is at a node in the *query-document (url)* graph. They found hitting time is effective in producing semantically consistent queries. Zhang et. al. [44] represented query logs as an *entity-auxiliary* bipartite graph with additional relevant information, e.g. contextual words and clicked URLs, and suggested a ensemble framework based on label propagation to learn the types of both entities and its auxiliary signals. Recently, Qi, Wu and Mamoulis [45] proposed a location-aware *query(keyword)-document* graph, which can capture the spatial distance between the resulting documents and the user location, as well as the semantic relevance between keyword queries. They found that the document proximity is important and can lead to better query recommendations.

Browse graphs are built based on users' Web browsing history, where the nodes are Web pages and the edges are the transitions between them in a users' browsing log. A users' web browsing behaviour was studied resulting in the *BrowseRank* model [16], [46], which built a graph of Web pages with edges representing the transitions between pages. This model was found more reliable than link-based graphs for inferring page importance, e.g. PageRank [47] and TrustRank [48]. Liu et. al. [49] studied the structure, evolution and application of the *browse graph* by comparing with link-based graphs. Trevisiol et. al. [22] compared different ranking techniques on a *browse graph* in the field of image ranking by using a datset from *Flickr*. *Browse graphs* have been used to tackle the cold-start recommender problems in the news domain, and achieved high accuracy with sparse data [50]. Recently, Trevisiol et. al. [51] investigated the local ranking problem on the *browse graph* for news item ranking, and showed the distance between rankings are predictable based on the structural information of the graph.

There are other relevant research studied with data mining techniques, ranging from traditional recommendation techniques, Location-based services (LBS) to Point of Interests (POI) Web search. For example, Sun et. al. [52] summarised the traditional recommendation techniques that investigating the spatio-temporal joint influence with probabilistic generative models and network embedding models. Xie et. al. [53] tackled location-based recommendations by modelling the relationships among POI, region, time and word in graphs with embedding learning techniques, and found this is effective in cold-start POIs. Zhao et. al. [54] proposed a geo-temporal sequential embedding rank model to capture the contextual check-in information in sequences, and found this works well for POI recommendations. However, based on our experiments, the data across the physical and cyber spaces are too sparse for applying these existing methods, because very few people have long or complete trajectories, hence a different approach is needed, looking at aggregate likelihoods, as in this study.

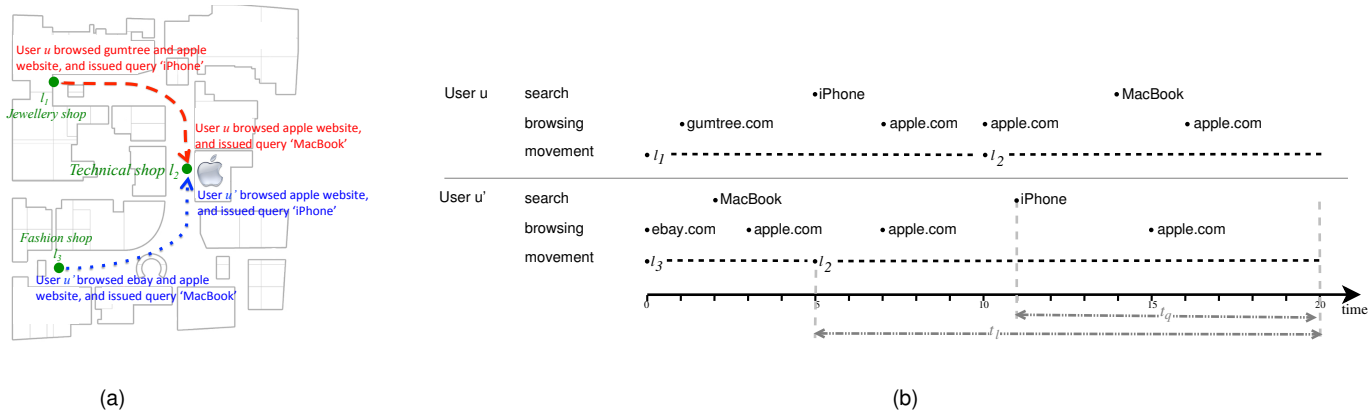
Overall, there is no work in the state of the art that examines a users' query, browsing, and movement log together with a focus on the physical and cyber contextual influence among them.

3 Terminology and Definitions

3.1 User Behaviour Logs

The conjunction between people's physical and cyber behaviour is recorded in the query, browsing and physical movement logs.

As shown in Fig. 1a, we deploy a running example to illustrate the definitions, the terms and the construction process of the subgraphs of the LQB graph. Specifically, the example includes two users, and their Web browsing/searching activities while they are moving in a shopping mall environment. They start from different locations in the mall, and browse and search something online, then move towards the Apple retail store. Fig. 1b shows the alignment of the search log, browsing log and movement log in time for two users. Table 1 illustrates the content of these logs corresponding to the example. To simplify the example we show only one session per user.



(a) (b)

Fig. 1. (a): Example of two users in a shopping mall environment. The green dots are locations with their names and types; the red dashed line and the blue dotted line are the movement trajectories for user u and u', respectively. (b): The illustration of the example logs by aligning the search log, browsing log and movement log in time order for each individual user. The dot • indicates the time stamp for the query, browsed Web domain and location, respectively. The black dotted line for movement logs indicates the duration of visiting a location. t_q and t_l are the time spent on query "iPhone" and in location l_2 for user u'.

TABLE 1
The corresponding logs of the illustration example in Fig. 1a. $\mathbf{P}_{l_1} = [0, 1, 0]$ is a binary representation of vector $[Technology, Jewellery, Fashion]$, where 1 means l_1 belongs to the corresponding location (shop) type (jewellery).

logs	records
search log	$\langle$iPhone, u, 5:00$\rangle$
$\langle q_i, u_i, t_i\rangle$	$\langle$MacBook, u, 14:00$\rangle$
	$\langle$MacBook, u', 2:00$\rangle$
	$\langle$iPhone, u', 11:00$\rangle$
browsing log	$\langle$gumtree.com, u, 1:00$\rangle$
$\langle b_i, u_i, t_i\rangle$	$\langle$apple.com, u, 7:00$\rangle$
	$\langle$apple.com, u, 10:00$\rangle$
	$\langle$apple.com, u, 16:00$\rangle$
	$\langle$ebay.com, u', 0:00$\rangle$
	$\langle$apple.com, u', 3:00$\rangle$
	$\langle$apple.com, u', 7:00$\rangle$
	$\langle$apple.com, u', 15:00$\rangle$
movement log	$\langle l_1, u,$ 00:00, 10, $\mathbf{P}_{l_1} = [0, 1, 0]\rangle$
$\langle l_i, u_i, t_i, d_i, \mathbf{P}_{l_i}\rangle$	$\langle l_2, u,$ 10:00, 10, $\mathbf{P}_{l_2} = [1, 0, 0]\rangle$
	$\langle l_3, u',$ 00:00, 5, $\mathbf{P}_{l_3} = [0, 0, 1]\rangle$
	$\langle l_2, u',$ 05:00, 15, $\mathbf{P}_{l_2} = [1, 0, 0]\rangle$

Query Log A typical query log contains information about users interactions with search engines, including the queries submitted, the time stamp, the returned documents/URLs as the results of the query, and the document/URLs clicked by the users. Here, we do not use any information from the search engine results page (SERP), thus we define a query log as a set of records: $\langle q_i, u_i, t_i\rangle$, where q_i is the submitted query, u_i denotes the user, and t_i denotes the time stamp when the query is submitted. A *search session* S_i is a series of query requests by a single user u_i within a specific time period, which is represented as:

$$S_i = \langle\langle q_{i_1}, u_i, t_{i_1}\rangle, \ldots, \langle q_{i_k}, u_i, t_{i_k}\rangle\rangle,$$

where $t_{i_1} \leq \cdots \leq t_{i_k}$, and $S_i \in \mathcal{S}$ that is the set of all search sessions. Together with the Web browsing log, we note that a *search session* contains all the issued queries, URL clicks and other Web pages navigated from the SERPs.

Web Browsing Log Similarly, a Web browsing log records information about users online access, and can be defined as a set of the following records: $\langle b_i, u_i, t_i\rangle$, where b_i denotes the browsed Web domain. Similarly, we define a *browsing session* B_i as a series of URL requests by a single user u_i within a specific time period, which is represented as:

$$B_i = \langle\langle b_{i_1}, u_i, t_{i_1}\rangle, \ldots, \langle b_{i_k}, u_i, t_{i_k}\rangle\rangle,$$

where $t_{i_1} \leq \cdots \leq t_{i_k}$, and $B_i \in \mathcal{B}$ that is the set of all browsing sessions.

Physical Movement Log A physical movement log contains users moving histories– symbolic trajectories, consisted of a visited location, the time stamp, the stay duration, and the type of the location. We define the physical movement log as a set of records: $\langle l_i, u_i, t_i, d_i, \mathbf{P}_{l_i}\rangle$, where l_i denotes the *id* of the visited location, d_i is the duration of the visit at l_i, and $\mathbf{P}_{l_i}$ is a vector, denoting the physical context of l_i in terms of location types when the log is being recorded. The symbolic location can be expressed at different scales, e.g. room or shop, coverage area, floor, building, or city. We assume, however, that the locations in a movement log are of homogeneous scale. For instance, in our example (Fig. 1a), the location is the service area of a WiFi access point, covering multiple types of shops, e.g. Technology shops, Jewellery shops, and Fashion shops. Thus, $\mathbf{P}_{l_2}$ for l_2 will be a vector of three: $[1, 0, 0]$, where 1 denotes there is a technology shop (Apple retail store) at l_2, but no jewellery or fashion shops there.

Similarly, we define a *movement session* $M_i \in \mathcal{M}$ as a series of movements by a single user u_i within a specific time period, which is represented as:

$$M_i = \langle\langle l_{i_1}, u_i, t_{i_1}, d_{i_1}, \mathbf{P}_{l_{i_1}}\rangle, \ldots, \langle l_{i_k}, u_i, t_{i_k}, d_{i_k}, \mathbf{P}_{l_{i_k}}\rangle\rangle,$$

where $t_{i_1} < \cdots < t_{i_k}$, and $M_i \in \mathcal{M}$ that is the set of all movement sessions.

Let $C = \{c_1, \ldots, c_h\}$ denote the h available location types, and the physical context for each location l is represented by the vector $\mathbf{P}_l$, which records the likelihood of belonging to each location type. It is formally defined as:

Definition 1. For each location l, let p_{lk} denote the likelihood of belonging to location type $c_k \in C$. Then $\mathbf{P}_l$ is defined

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2017.2766059, IEEE Transactions on Knowledge and Data Engineering

TABLE 2
Symbols

Symbol	Description
u	the user
l	the location
q	the query
b	the browsed Web domain
$\|\cdot\|$	size of a set
$\mathcal{S} = \{S_i\}$	set of search sessions ($i = 1, \ldots, \|\mathcal{S}\|$)
$\mathcal{B} = \{B_i\}$	set of browsing sessions ($i = 1, \ldots, \|\mathcal{B}\|$)
$\mathcal{M} = \{M_i\}$	set of movement sessions ($i = 1, \ldots, \|\mathcal{M}\|$)
q_i, b_i, l_i, u_i	q, b, l, u, t in the corresponding search, browsing or movement session.
t_i	start time stamp of the i-th session in $\mathcal{S}, \mathcal{B}$, or $\mathcal{M}$
t_q, t_b, t_l	time stamp of issuing q, browsing b, visiting l
C	the set of location types
c_k	the k-th type in C
h	the number of location types
$\mathbf{P}_l, \mathbf{P}_b, \mathbf{P}_q$	the physical context of l, b, and q
p_{lk}	likelihood of location l belonging to category $c_k \in C$
L_b, L_q	set of locations where b or q is issued.
V, A, W	sets of nodes, arcs and weights in a graph
G_{lqb}	the LQB tripartite graph: $G_{lqb} = (V_{lqb}, A_{lqb}, W_{lqb})$
G_{ql}	the query-location bipartite subgraph: $G_{ql} = \{V_{ql}, A_{ql}, W_{ql}\}$
G_{bl}	the browse-location bipartite subgraph: $G_{bl} = \{V_{bl}, A_{bl}, W_{bl}\}$
G_{qb}	the query-browse bipartite subgraph: $G_{qb} = \{V_{qb}, A_{qb,qb}\}$
$w(\cdot,\cdot)$	the weight on the arcs connecting two nodes
$\hat{G}$	the projection of a bipartite subgraph
f_q, f_b	the frequency of q and b
X	the transition matrix
$\mathbf{I}$	the unit matrix
$\mathbf{e}_v$	the vector that only have the v-th component equal to 1 and others equal to 0
α	the damping factor
$r(\cdot)$	the random walk values on vertices
$l(\cdot)$	the ranks obtained from corresponding random walk values
β_1, β_2, θ	the scaling factors

as a vector of size h, with entry k storing the likelihood p_{lk} of belonging to c_k:

$$\mathbf{P}_l = [p_{a1}, \ldots, p_{lk}, \ldots, p_{ah}]. \quad (1)$$

3.2 Definitions and symbols

Table 2 lists the main symbols used throughout this paper. Let $G_{lqb} = (V_{lqb}, A_{lqb}, W_{lqb})$ denote the LQB graph, where:

- $V_{lqb} = V_l \cup V_q \cup V_b$ is the union of three sets of different kinds of nodes: the set of distinct physical locations V_l, the set of distinct queries V_q, and the set of distinct browsed Web domains V_b.
- A_{lqb} denotes the set of arcs (oriented edges) among these nodes. There are only arcs between heterogeneous nodes, representing the contextual influence, as discussed in the following section.
- $W_{lqb} : A_{lqb} \to (0, 1]$ denotes the weights on the arcs.

Even if a query has been issued multiple times by a user or by multiple users, it is denoted as a single node in the LQB graph. This also applies to location and Web domain nodes.

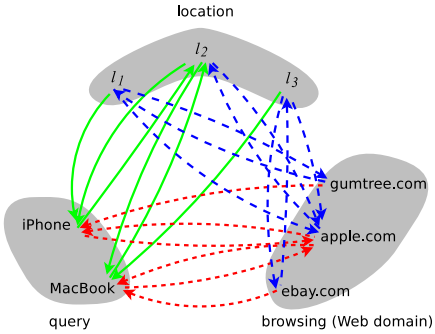


Fig. 2. The LQB Graph of the illustration example in Fig. 1a.

The physical contexts of browsing Web domains $\mathbf{P}_b$ and the physical context of queries $\mathbf{P}_q$ are defined as follows:

Definition 2. $\mathbf{P}_b$ is defined as the average of $\mathbf{P}_l, l \in L_b$, where L_b denotes the set of locations where b is browsed:

$$\mathbf{P}_b = \frac{\sum_{l \in L_b} \mathbf{P}_l}{|L_b|}. \quad (2)$$

Definition 3. $\mathbf{P}_q$ is defined as the average of $\mathbf{P}_l, a \in L_q$, where L_q denotes the set of locations where q is issued:

$$\mathbf{P}_q = \frac{\sum_{l \in L_q} \mathbf{P}_l}{|L_q|}. \quad (3)$$

4 Tripartite LQB Graph

We propose the LQB graph, a tripartite graphical representation of how people behave in the conjunction of physical and cyber spaces by focusing on the contextual influence. Fig. 2 shows the corresponding LQB graph for the illustration example in Fig. 1a. This graph includes three kinds of nodes: location, query, and browsed Web domain. There are only arcs between heterogeneous nodes, representing contextual influences.

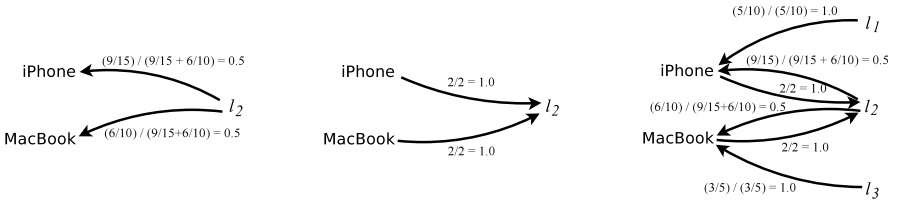
The LQB graph is constructed based on a set of search sessions $\mathcal{S}$, a set of browsing sessions $\mathcal{B}$, and a set of movement sessions $\mathcal{M}$, extracted from users' behaviour logs as defined in Sec. 3. Note, V_{lqb} includes the distinct sets of queries, Web domains and locations, while the corresponding sessions include all instances of issuing/browsing/visiting these distinct queries, Web domains and locations. We describe the tripartite LQB graph through the three partial bipartite graphs: (1) query-location G_{ql}; (2) browse-location G_{bl}; and (3) query-browse G_{qb}, capturing the contextual influence among corresponding nodes.

4.1 Query-Location Bipartite Subgraph G_{ql}

Query activities occur in a certain physical context, and we leverage this information into our graph formulation. This is achieved by aligning the search sessions $\mathcal{S}$ and the movement sessions $\mathcal{M}$ in time order for each user. Then, we define a bipartite subgraph $G_{ql} = \{V_{ql}, A_{ql}, W_{ql}\}$, where $V_{ql} = V_q \cup V_l$, $A_{ql} \subset V_q \times V_l$ denotes the set of arcs connecting queries and locations.

Given a query $q \in V_q$ and a location $l \in V_l$, the arc from l to q is introduced if there is at least one user u who issued

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2017.2766059, IEEE Transactions on Knowledge and Data Engineering



(a) from location to query (b) from query to location (c) the query-location subgraph

Fig. 3. The construction of G_{ql} for the example. (a) the weighted arcs from location node l_2 to query nodes "iPhone" and "MacBook", where the weights are calculated by Eq. 5;(b) the weighted arcs from query nodes "iPhone" and "MacBook" to location node l_2, where the weights are calculated by Eq. 7; (c) the subgraph G_{ql} of the example.

q when s/he is at l. Specifically, the arc from l to q is subject to the following conditions:

- there exists at least one search session $S_i \in \mathcal{S}$ and one movement session $M_j \in \mathcal{M}$ from the same user, which means $u_i = u_j$, where $u_i \in S_i$ and $u_j \in M_j$;
- S_i includes the issuing of q: $q = q_i \in S_i$;
- M_j includes the visiting of l: $l = l_j \in M_j$;
- the user u issued q while s/he is at l in M_j, which means: $t_j^m \leq t_i^s < (t_j^m + d_j^m)$, where t_j^m is the time stamp while u starts visiting l_j in M_j, d_j^m is the duration spent at l_j, and t_i^s is the time stamp while u issued q_i in S_i.

Thus, the arcs from V_l to V_q are defined as:

$$\begin{aligned} A(l,q) =& \{(l,q) | \exists S_i \in \mathcal{S}, \exists M_j \in \mathcal{M}, \\ & \text{so that } u_i = u_j \bigwedge q = q_i \in S_i \bigwedge l = l_j \in M_j \\ & \bigwedge t_j^m \leq t_i^s < (t_j^m + d_j^m)\} \end{aligned} \tag{4}$$

The weight $w(l,q)$ on arc $A(l,q)$ is defined as the normalised ratio of the time spent on query q over the time spent at l where q was issued:

$$w(l,q) = \frac{\eta(l,q)}{\sum_{q' \in V_q} \eta(l,q')}, \text{where } \eta(l,q) = \frac{\sum_{sm_i \in SM} \frac{t_q^i}{t_l^i}}{|SM|}, \tag{5}$$

SM denotes the pairs of search sessions and movement sessions as specified in Eq. 4, t_q^i denotes the time spent on q in the corresponding search session in sm_i when the user is at l, t_l^i denotes the time that the user spent at l in the corresponding movement session in sm_i. Specifically, t_q^i is calculated as the time gap in seconds between q and next query, or the end of the search session if q is the last query in the search session, or the end of the visit of the current location l. For example, the time spent on query "iPhone" and the time spent in location l_2 by user u' are shown as t_q and t_l in Fig. 1b. Fig. 3a shows the weighted arcs from location node l_2 to query node, "iPhone" and "MacBook", based on the example log shown in Table 1.

On the other side, the arcs from V_q to V_l are defined based on l's physical context. Specifically, when q is issued by a user u, a link is defined from q to l if

- there exists at least one search session $S_i \in \mathcal{S}$ and one movement session $M_j \in \mathcal{M}$ from the same user, which means $u_i = u_j$, where $u_i \in S_i$ and $u_j \in M_j$;
- S_i includes the issuing of q: $q = q_i \in S_i$;
- M_j includes the visiting of l: $l = l_j \in M_j$;

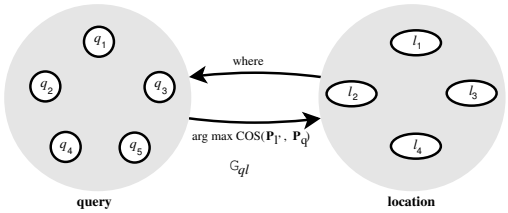


Fig. 4. Illustration of G_{ql}

- $l_j \in M_j$ is visited by u after s/he issued $q_i \in S_i$, which means $t_i^s \leq (t_j^m + d_j^m)$, where t_i^s denotes the time stamp when q_i is issued in S_i, t_j^m the time stamp when u starts visiting l_j in M_j, and d_j^m is the duration spent at l_j;
- l_j has the most similar physical context to q's, which means $l_j = \arg\max_{l' \in M_j} COS(\mathbf{P}_{l'}, \mathbf{P}_q)$, where $COS(\cdot,\cdot)$ denotes the cosine similarity.

The arcs from V_q to V_l are defined as:

$$\begin{aligned} A(q,l) =& \{(q,l) | \exists S_i \in \mathcal{S}, \exists M_j \in \mathcal{M}, \\ & \text{so that } u_i = u_j \bigwedge q = q_i \in S_i \bigwedge l = l_j \in M_j \\ & \bigwedge t_i^s < (t_j^m + d_j^m) \bigwedge l_j = \arg\max_{l' \in M_j} COS(\mathbf{P}_{l'}, \mathbf{P}_q)\}. \end{aligned} \tag{6}$$

Similarly, the weight $w(q,l)$ on arc $A(q,l)$ is defined as the frequency of q connected to l normalised by the overall occurrence of q:

$$w(q,l) = \frac{f_{ql}}{f_q}, \tag{7}$$

where f_{ql} denotes the frequency of q connected to l, and f_q denotes the number of occurrence of q. Fig. 3b shows the weighted arcs from query node, "iPhone" and "MacBook", to location node l_2, where the value 2 in $2/2 = 1.0$ on the arcs denotes the two instances of issuing "iPhone" and "MacBook" from u and u' on each arc accordingly. Here, it is assumed that l_2's context is most similar to that of the query, "iPhone" and "MacBook". Fig. 3c shows the final subgraph G_{ql} of the example log. Fig. 4 shows an illustration of G_{ql}.

4.2 Browse-Location Bipartite Subgraph G_{bl}

Like queries, browsing activities also occur in a certain physical context, and this is achieved by aligning the browsing sessions $\mathcal{B}$ and the movement sessions $\mathcal{M}$ in time order for each user. Consequently, we define a bipartite subgraph $G_{bl} = \{V_{bl}, A_{bl}, W_{bl}\}$, where $V_{bl} = V_b \cup V_l$, $A_{bl} \subset V_b \times V_l$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2017.2766059, IEEE Transactions on Knowledge and Data Engineering

denotes the set of arcs connecting browsing Web domains and locations.

Given a browse Web domain $b \in V_b$ and a location $l \in V_l$, the arc from l to b is introduced if there is at least one user who browsed b when s/he is at l. As this is similar to the arc from a location l to a query q as detailed in Section 4.1, we do not list these conditions here to avoid repetition. Thus, the arcs from V_l to V_b is formally defined as:

$$\begin{aligned} A(l,b) =& \{(l,b) | \exists B_i \in \mathcal{B}, M_j \in \mathcal{M}, \\ & \text{so that } u_i = u_j \bigwedge b = b_i \in B_i \bigwedge l = l_j \in M_j \\ & \bigwedge t_j^m \leq t_i^b < (t_j^m + d_j^m)\} \end{aligned} \tag{8}$$

The corresponding weight $w(l,b)$ on arc $A(l,b)$ is defined as the normalised ratio of the time spent on browsing Web domain b over the time spent at l where b was accessed:

$$w(l,b) = \frac{\eta(l,b)}{\sum_{b' \in V_b} \eta(l,b')}, \text{where } \eta(l,b) = \frac{\sum_{bm_i \in BM} \frac{t_b^i}{t_l^i}}{|BM|}, \tag{9}$$

BM denotes the pairs of the browsing sessions and movement sessions as specified in Eq. 8, t_b^i denotes the time the user spent at b in the corresponding browsing session in bm_i, and t_l denotes the time that the user spent at l in the corresponding movement session in bm_i. Similarly, t_b^i can be calculated as the total time spent at b when s/he is at l in browsing session s_b, or the end of the browsing session if b is the last Web domain browsed, or the end of the visit of the current location.

Similarly, the arcs from V_b to V_l are defined based on their contexts. When b is browsed by a user, an arc is defined from b to l if 1) l has the most similar physical context to b's; 2) l is visited by the user after s/he browsed b at least once. Thus, we obtain the arcs from V_b to V_l as:

$$\begin{aligned} A(b,l) =& \{(b,l) | \exists B_i \in \mathcal{B}, \exists M_j \in \mathcal{M}, \\ & \text{so that } u_i = u_j \bigwedge b = b_i \in B_i \bigwedge l = l_j \in M_j \\ & \bigwedge t_i^b < (t_j^m + d_j^m) \bigwedge l_j = \arg\max_{l' \in M_j} COS(\mathbf{P}_{l'}, \mathbf{P}_b)\} \end{aligned} \tag{10}$$

The weight $w(b,l)$ on arcs $A(b,l)$ is defined as the normalised frequency of b connected to l:

$$w(b,l) = \frac{f_{bl}}{f_b}, \tag{11}$$

where f_{bl} denotes the frequency of b connected to l, and f_b denotes the number of occurrence of b. Fig. 5a illustrates the G_{bl} of the example log.

4.3 Query-Browse Bipartite Subgraph G_{qb}

We leverage the influence between queries and Web domains in our contextual graph model, and similarly this is achieved by aligning the browsing sessions $\mathcal{B}$ and the search sessions $\mathcal{S}$ in time order for each user. In the cyber context, issuing queries involves two kind of nodes, the query node and the Web domain node, which gives the bipartite subgraph G_{qb} across heterogeneous nodes: $G_{qb} = \{V_{qb}, A_{qb}, W_{qb}\}$, where $V_{qb} = V_q \cup V_b$, $A_{qb} \subset V_q \times V_b$ denotes the set of arcs connecting queries and Web domains.

We refer to the Web domain accessed by users just before a query is issued as transited Web access. Thus, similar to the conditions for $A(l,q)$ in Section 4.1, the arcs from V_b to V_q is defined based on where q was transited:

$$\begin{aligned} A(b,q) =& \{(b,q) | \exists S_i \in \mathcal{S}, \exists B_j \in \mathcal{B}, \\ & \text{so that } u_i = u_j \bigwedge q = q_i \in S_i \bigwedge b = b_j \in B^j \\ & \bigwedge t_j^b \leq t_i^s < t_{j+1}^b\}, \end{aligned} \tag{12}$$

where t_j^b is the time stamp while u browses b_j in B_j, and t_i^s is the time stamp while u issued q_i in S_i. Namely, the arc from b to q is introduced if q was transited from b in any user browsing session. The arcs from V_q to V_b are defined based on their physical contexts. Specifically, the directed connectivity from q to b is introduced if 1) b has the most similar physical context to q's; 2) b is browsed by the user after s/he issued q:

$$\begin{aligned} A(q,b) =& \{(q,b) | \exists S_i \in \mathcal{S}, \exists B_j \in \mathcal{B}, \\ & \text{so that } u_i = u_j \bigwedge q = q_i \in S_i \bigwedge b = b_j \in B^j \\ & \bigwedge t_i^s \leq t_j^b \bigwedge b_j = \arg\max_{b' \in B_j} COS(\mathbf{P}_{b'}, \mathbf{P}_q)\} \end{aligned} \tag{13}$$

Fig. 5b illustrates the G_{qb} of the example log.

The corresponding weights are defined as:

$$w(b,q) = \frac{f_{bq}}{f_b}, \text{ and } w(q,b) = \frac{f_{qb}}{f_q}, \tag{14}$$

where f_{bq} denotes the number of q transited from b, f_b denotes the number of all queries transited from b, f_{qb} denotes the number of q connected to b, and f_q denotes the number of occurrence of q. Namely, $w(b,q)$ is the fraction of q transited from b over all queries transited from b, while $w(q,b)$ is the fraction of b connected from q over the occurrence of q.

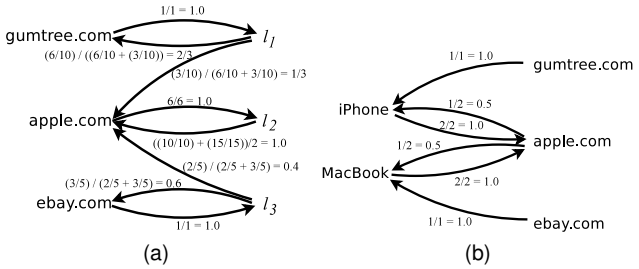


Fig. 5. (a) The construction of G_{bl} and the weights are obtained with Eq. 9 and 11. (b) The construction of G_{qb} and the weights are obtained with Eq. 14.

4.4 Model

Given the LQB graph is a representation of users' moving, browsing and querying behaviours across the physical and cyber spaces, it is appropriate to consider the recommendation problem on this graph as a random walk, as shown in [37]. Specifically, for efficiency and simplicity purposes, we first project the heterogeneous bipartite subgraphs to homogeneous graphs, then deploy a random walk model. Moreover, the LQB graph can be applied to produce three kinds of applications: location recommendation, query recommendation and Web content recommendation, which corresponds to the three kinds of nodes. Here, we describe the modelling of the LQB graph in the application of query

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2017.2766059, IEEE Transactions on Knowledge and Data Engineering

recommendation, and the application on Web content and location recommendation can be obtained in a straightforward manner.

There are three heterogeneous bipartite sugbraphs G_{ql}, G_{qb} and G_{bl}, and each of them can be projected in two ways. For example, G_{ql} can be projected as a location graph $\hat{G}_l^{ql}$ and a query graph $\hat{G}_q^{ql}$. Here, we show how to project G_{ql} to a query graph $\hat{G}_q^{ql}$. Specifically, when projecting G_{ql} to a query graph $\hat{G}_q^{ql}$, there are some approaches that transform a heterogeneous graph to homogeneous graph [55]. Here, we consider three approaches:

- Binary: this is the simplest approach, which defines the weight $w(q,q')_q^{ql}$ as either 1 or 0, depending whether there is at least one path connecting q to q':

$$w(q,q')_q^{ql} = \begin{cases} 1 & \text{if } \exists l \in V_l, w(q,l) > 0, w(l,q') > 0 \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

- Distributional: by considering the distributional information associated with the relationship between queries and locations, the weight $w(q,q')_q^{ql}$ can be defined as the sum of the weights on all possible paths from q to q' via any $l \in V_l$:

$$w(q,q')_q^{ql} = \sum_{l \in V_l} \left(w(q,l) w(l,q') \right). \tag{16}$$

- Macro-Aggregation: this projection approach is proposed in [55] to remove the potential bias from very active users. Specifically, it first treats each user's logs independently and creates the LQB graph, then aggregates the weights across all users:

$$w(q,q')_q^{ql} = \sum_u \left(w_u(q,q')_q^{ql} \right), \tag{17}$$

where $w_u(q,q')_q^{ql}$ is the distributional weight from the LQB graph for user u.

Fig. 6 illustrates the projected query graphs from G_{ql} of the example log with different projection approaches.

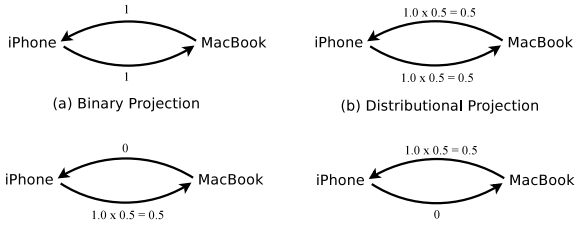


Fig. 6. Illustration of the projection approaches of $\hat{G}_q^{ql}$ from G_{ql}.

Then, a random walk with restart to a single node is deployed the projected homogeneous subgraphs. Although similar ideas were investigated in [37], they focused on modelling the consequential order of queries. But, we focus on modelling multiple contextual influence among location, query and browsing contexts. Specifically, given a graph G, a random surfer starts from a single node v in the graph, then continues to surf following one of the leaving edges from current node with a probability α, and comes back to the original node v with a probability $(1-\alpha)$. Formally, this can be defined as a Markov chain: $X = \alpha W + (1-\alpha)\mathbf{I}\mathbf{e}_v^T$, where X is the transition matrix, α is the damping factor, W is the weight matrix of the subgraph, $\mathbf{I}$ is the unit matrix, and $\mathbf{e}_v$ denotes a vector that only have the v-th component equal to 1 and others equal to 0. After the stationary distribution is achieved, each node v' (other than the original node) in the graph is allocated with a random-walk score $r(v')_G$, representing the relevance to v. Moreover, the random walk could start with some historical information. For example, if the previous node v' is known for current node v, the corresponding random walk model becomes:

$$X = \alpha W + (1-\alpha)\mathbf{I}\mathbf{e}_{v',v}^T \tag{18}$$

where $\mathbf{e}_{v',v}$ denotes a vector that only have the v'-th and v-th component equal to 1 and others equal to 0. Note, for other ways of allocating values to $\mathbf{e}_{v',v}$, please refer to [37].

In the proposed LQB graph, for the query node, there are two projections, $\hat{G}_q^{ql}$ and $\hat{G}_q^{qb}$, from G_{ql} and G_{qb}, respectively. After running random walks on each of them, for each query node $q' \in V_q$, we obtain two random-walk scores, $r(q')_{\hat{G}_q^{ql}}$ and $r(q')_{\hat{G}_q^{qb}}$. To merge these two ranking results, we deploy a simple rank-based merging function:

$$r(q')_{G_{lqb}} = \beta_1 \frac{1}{l(q')_{\hat{G}_q^{ql}} + 1} + \beta_2 \frac{1}{l(q')_{\hat{G}_q^{qb}} + 1}, \tag{19}$$

where $l(q')_{\hat{G}_q^{ql}}$ and $l(q')_{\hat{G}_q^{qb}}$ are the ranks obtained from $r(q')_{\hat{G}_q^{ql}}$ and $r(q')_{\hat{G}_q^{qb}}$, respectively; β_1 and β_2 are scaling factors that represent the importance of corresponding random-walk scores, which can be obtained by cross-validation. Note, there are other merging functions, e.g. the value-based merging approach in [16]. However, we argue the above rank-based function is more appropriate, because the random-walk scores from $r(q')_{\hat{G}_q^{ql}}$ and $r(q')_{\hat{G}_q^{qb}}$ are not comparable. The final ranking is generated by sorting all queries according to $r(q')_{G_{lqb}}$ in a decreasing order. Consequently, following [18], [37], we suggest the top ranked queries to the user, and call this as query recommendations.

Location and Web content recommendations are generated in a similar way by projecting the LQB graph to homogeneous location graphs and domain graphs, respectively.

5 Validating the LQB Graph in Indoor Retail Environment

In this section, we report on experiments validating the LQB graph on a real-world indoor retail environment. We first characterise the contextual influences on people's behaviours in this scenario by corresponding them to the arcs in the LQG graph and then evaluate the usefulness of the graph in three applications: location recommendation, Web content recommendation and query recommendation.

5.1 Data Acquisition

Data were collected from over 120,000 anonymized users between September 2012 and October 2013 via a free, opt-in Wi-Fi network operated by an inner city shopping mall in Sydney, Australia. The mall is around 90,000 square meters covered by 67 Wi-Fi access points (APs). Three kinds

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2017.2766059, IEEE Transactions on Knowledge and Data Engineering

of logs were collected, including a 1-million rows of Wi-Fi AP association log (AL) capturing physical movement with APs corresponding to locations l and the served shop categories to the location types C, a 18-million rows of Web browsing log (BL), and a 100-thousand rows of query log (QL). There are over 200 stores in the mall, belonging to 34 shop categories defined by the mall operator, e.g. Fashion, Footwear, Travel, Jewellery, Sport, Toys and Hobbies, and Cinemas. Note, the Wi-Fi covers common areas of the mall (so not inside the shops). The logs capture all associations of registered users with the WiFi network, without distinguishing what purpose the visit served.

5.1.1 Wi-Fi AP Association Log

The Wi-Fi AP Log (AL) captures information about user physical behavior characterized by the following parameters (1) user device's MAC address uniquely identifying the associated device (information was hashed to anonymize it); (2) the users' IP address; (3) the ID of the Wi-Fi AP (not MAC address) associated with the user's mobile device at a given point in time, used as a proxy for the user's location; (4) the time-stamp of users' association/disassociation with the access point.

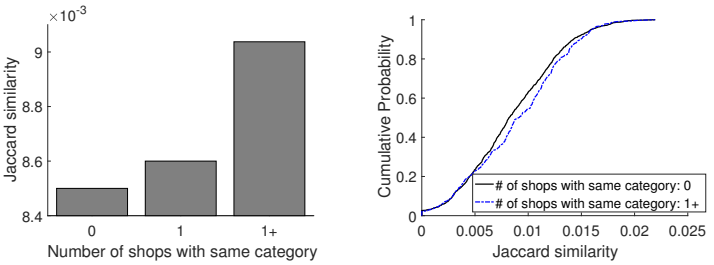
To obtain the type of locations l (APs), floor plans of the mall were overlaid with AP positions and the service areas of the APs were approximated by Voronoi regions [56], each centered on a single AP, that encompass all the points that are closest to that AP. The regions were manually rectified to correspond better with the frontages of physical stores in the mall (see [57] for details). Shop frontages are the main determinants of context as the Wi-Fi network is meant to cover common spaces in the mall. More details about the overlaying of floor maps with APs can be found in [58]. Thus, the physical context of an AP (location l in LQB graph) is defined both by the shops covered within its signal coverage (defined in Def. 1). Moreover, as this log data does not include the purposes of users' visits, we treat user movements equally. In addition, as the locations are WiFi Access Points, which has been determined already in the logs, we use all those available locations, and can not distinguish whether the querying or browsing behaviour happen inside or outside the shop locations

5.1.2 Web Browsing Log

The Browsing Log (BL) includes the users' Web browsing behavior, characterized by: (1) the time-stamp of the Web request; (2) the users' IP address; (3) the Web page requested, as defined by the URL. This contains all out-going URL requests from the device, including app traffic.

We enriched the BL with an attribute identifying the location of the user at the time of the request, by joining the BL with AL records through a composite key of time-stamp and IP address. The first appearance of a users' device in the AL, as well as any consecutive appearance after disconnection always precede appearance in the BL. It is also possible for the user to only connect to the Wi-Fi network and not access Web pages, thereby only appearing in the AL.

We further enriched the BL with an attribute categorising the URL through the Brightcloud service (http://www.brightcloud.com/). We also remove URLs



(a) Jaccard similarity (b) ECDFs of the Jaccard similarity

Fig. 7. Location context influence on queries. (a) Jaccard similarity of query sets vs the number of shops sharing the same category. (b) ECDFs of the Jaccard similarity of query sets.

with auxiliary or advertising content from (1) *Content Delivery Networks* URLs, incl. ads, media, files, images, and video providers; and (2) *Web Advertisements* URLs, incl. ads, media, content, and banners. We also removed *Dead Sites*, which did not respond to http requests. We finally obtained around 1.6 million URL requests. Following [13], [28], [37], [59], we set a timeout threshold to a *browsing sessions*, thus implementing a browsing session as *a series of URL requests by a single user delimited by* 30 *minutes of inactivity on the Web*.

5.1.3 Query Log

The Query Log (QL) was extracted from the general BL (QL $\subset$ BL), by following the steps in [13] to identify certain URL requests from search engines. The final QL includes 104,063 queries from 54 search engines, belonging to three groups: *General* (91.7% from 9 search engines): incl. Google, Naver, Yahoo, Daum, Bing, Baidu, AOL, ASK, searchmobileonline; *Special* (4.2 % from 15 search engines): e.g. Domain, Google Maps, Domain, SEEK, Google Images, Wiki; *E-Commerce* (4.1% from 30 search engines): incl. Gumtree (an Australian online classifieds Ads and community Website), Taobao, Ebay, JB Hi-Fi, Asos, Amazon, Tripadvisor, Booking, etc. Note, as the LQB graph does not use the SERP and the clicked URLs, we did not identify and process them.

The QL was processed as follows: (1) search queries were treated as case insensitive; (2) a query term was defined as an unbroken string of characters in a query separated by whitespace, other special characters (e.g. #,% and /) were treated as normal characters; (3) The QL was segmented by consistently applying the similar processing of the BL. Similar to browsing sessions, we define a search session as *a series of query requests by a single user delimited by 30 minutes of inactivity on the Web*.

5.2 Characterising the Contextual Influence

5.2.1 Querying Behaviour and Location Context

To characterise the influence of physical location on people's querying behaviour, we examine the overlap of the sets of queries in different contexts in terms of Jaccard similarities.

We apply Jaccard similarity to measure the overlap between the queries issued at different APs. Because an AP normally covers several shops, we group the AP by the cardinality of the set intersection between shop categories at pairs of APs: '0' means that the shops covered by two APs have no common categories; '1' means there is one shop from each of two paired APs having the same categories;

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2017.2766059, IEEE Transactions on Knowledge and Data Engineering

'1+' means there is more than one shops from both APs having the same categories. Note, '1+' does not contain many options as, on average, an AP covers around 3 shops. Fig. 7a shows the Jaccard similarity, increasing with the number of same-category shops. Fig. 7b shows the Empirical Distribution Function (ECDF) of Jaccard similarity when there are no same-category shops and more than one same-category shops. The Kolmogorov-Smirnov test result (D = 0.0942, p-value = 0.0270) shows there is a statistically significant difference between these two distributions. These results indicate that people within similar physical locations issue a larger fraction of similar queries than those in dissimilar locations.

5.2.2 Browsing Behaviour and Location Context

The physical location context also influences people's browsing behavior, which we investigate in terms of Web domains by deploying the same analysis method for querying behaviour detailed above. Similar trends have been observed: APs covering shops of the same-category have significantly higher Jaccard similarity than APs without same-category shops, as tested by the Kolmogorov-Smirnov test (D = 0.2075, p-value < 0.0001). The reason might be that queries determine the users' browsing behaviours. This indicates that users in similar location contexts are more likely to browse the same Web domains than users in distinct location contexts. Note, this fine-granularity analysis at the level of Web domains are different from the coarse analysis in [57] at the level of Web content categories (e.g. emails, news), which is too coarse to support contextual recommendations.

5.2.3 Querying Behaviour and Cyber Context

We study the influence of people's current cyber context on their querying behaviour captured through Web domains, by investigating the Jaccard similarity between the query sets transited from different Web domains, including two different kinds of Web domains, *e-commence* (*gumtree* and *ebay*) and *social networks* (*facebook*, *twitter*, and *tumblr*). We observe higher similarity for domains of the same kind, while lower similarity is observed across domain. For example, *gumtree* has around 50% of queries overlapped with *ebay*, while it only has around 20% of queries overlapped with *facebook*. This indicates that people transitioned from *facebook* are issuing a small fraction of queries that are issued by people from *gumtree*. What people query is thus dependent on the cyber context of the browsing Web domains.

There are significant differences in the categories of Web content clicked by people transitioned from different Web domains. Table 3 shows the top popular Web categories that are queried and clicked by people transitioned from *gumtree*, *ebay*, *facebook*, *twitter* and *tumblr*. Here, we used only queries and clicks issued to the *google* search engine, because it is both the most popular and most general purpose search engine in the QL (e.g. a search within a special e-commerce website has a high chance of leading to another page within that site, which means the type of the click-through is biased to be *Shopping*). One difference is the order of the types of queried Web content. Specifically, for *gumtree* and *ebay*, the most popular type of query-click content is *Shopping*. This is expected, as both of them are *e-commerce* Web sites, and people might compare prices between different e-commerce sites or check customer reviews on products. For *facebook* and *twitter* the most popular kind of query-click content is *Business & Economy*. Another important difference is that transitions from *Social Network* domains, including *facebook*, *twitter* and *tumblr*, shows strong interests in *News & Media* and *Entertainment & Arts*, which is not observed in transitions from *gumtree* and *ebay*. This indicates an influence of people's cyber context (Web domains) on their querying behaviour.

Overall, we observe mutual and bi-directional influences between people's physical location context, cyber browsing context, and their querying context. Thus, while people who are in the similar cyber and location contexts tend to issue the similar queries, currently issued queries can also reflect their cyber and location contexts. These influences are captured by the LQB graph.

5.3 Experimental Results

We now test the ability of the LQB graph to provide users with recommendations about future interests in locations, queries, and Web content based solely on user's current location l, or current query q, or current browsing Web domain b.

We apply 5-fold cross validation to evaluate the performance of the proposed LQB graph. Specifically, we chronologically divide the logs into 5 equal sized sets to get reliable experimental results, which means no search/browsing/movement sessions are split into training and test set at the same time. Note there are three kinds of applications: location recommendation, Web content recommendation and query recommendation. For each application, we perform the procedure illustrated on the example of query recommendations on each of the 5 experiment sets: 1) For each search session in the current set, randomly select a query q as current query and leave the remaining following queries as ground-truth. 2) Build the LQB graph on the remaining 4 experiment sets and train the scaling factors β_1 and β_2 with cross-validation; 3) Calculate the accuracy of the query recommendations generated using the LQB graph on the ground-truth queries. The reported results are the average accuracy of all 5 experiment sets, with the damping factor α set to 0.85. Following [18], to remove sampling bias in the experiments (since some users in the logs were much more active than others), we randomly selected at most 10 days of logs from each user, resulting in $120,548$ users, 67 Wi-Fi APs, $56,281$ Web domains, and $54,647$ distinct queries.

5.3.1 Measurement Metrics and Baselines

We apply three standard metrics to evaluate the ranking accuracy of contextual recommendations [18], [60]: (1) Precision in the top k ($p@k$) as the average fraction of the top k true items found in the recommendation list; (2) Recall in the top k ($r@k$) as the average fraction of the true items that are successfully retrieved; and (3) Mean Reciprocal Rank (MRR) is one over the rank of the top ranked relevant item.

To examine the effectiveness of the LQB graph, we compare the performance of the following methods: (1) $random$ model: the recommendation list is generated by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2017.2766059, IEEE Transactions on Knowledge and Data Engineering

TABLE 3
Top popular Web content queried by people from different Web domains

gumtree	*ebay*	*facebook*	*twitter*	*tumblr*
Shopping	*Shopping*	Business & Economy	Business & Economy	Social Network
Business & Economy	Auctions	*News & Media*	*Shopping*	*Entertainment & Arts*
Travel	Business & Economy	*Shopping*	Reference & Research	Personal sites & Blogs
Society	Travel	Travel	Travel	*News & Media*
Reference & Research	Society	Reference & Research	*Entertainment & Arts*	Business & Economy

TABLE 4
Results of location, Web content and query recommendations

Recommendations	Methods	current location l					knowing previous location $l' \rightarrow l$				
		p@5	p@10	r@5	r@10	MRR	p@5	p@10	r@5	r@10	MRR
Location Recommendations	$random$	0.0371	0.0373	0.0744	0.1495	0.0869	0.0371	0.0373	0.0744	0.1495	0.0869
	topology	0.0389	0.0408	0.0787	0.1651	0.1010	0.0389	0.0408	0.0787	0.1651	0.1010
	ap-flow	0.0615	0.0491	0.1263	0.2015	0.1891	0.0907	0.0624	0.1863	0.2563	0.2538
	valueMerge	0.0588	0.0549	0.1212	0.2264	0.1863	0.0867	0.0676	0.1789	0.2788	0.2467
	G_{lqb}^{binary}	0.0660	0.0527	0.1346	0.2151	0.1958	0.0875	0.0696	0.1784	0.2838	0.2601
	G_{lqb}^{macro}	0.0401	0.0377	0.0802	0.1877	0.1268	0.0542	0.0438	0.1302	0.1873	0.1533
	$\hat{G}_l^{ql}$	0.0625	0.0562	0.1283	0.2308	0.1953	0.0884	0.0699	0.1816	0.2873	0.2612
	$\hat{G}_l^{bl}$	0.0663	0.0571	0.1362	0.2345	0.2062	0.0998	0.0696	0.2011	0.2857	0.2718
	G_{lqb}^{distro}	**0.0704**	**0.0577**	**0.1446**	**0.2329**	**0.2063**	**0.0999**	**0.0717**	**0.2050**	**0.2946**	**0.2719**
Web Content Recommendations	$random$	0.3884	0.3351	0.2466	0.3596	0.6548	0.3884	0.3351	0.2466	0.3596	0.6548
	domain-flow	0.5906	0.5681	0.3671	0.5551	0.7965	0.5943	0.5324	0.3619	0.5398	0.8243
	valueMerge	0.5473	0.4291	0.4544	0.5042	0.7065	0.5134	0.4181	0.4244	0.5340	0.6903
	G_{lqb}^{binary}	0.5243	0.4023	0.4597	0.5557	0.7319	0.4292	0.3589	0.4246	0.5407	0.7971
	G_{lqb}^{macro}	0.4203	0.3892	0.2783	0.3889	0.7001	0.4079	0.3397	0.2599	0.3678	0.6077
	$\hat{G}_b^{bl}$	0.5575	0.4758	0.3033	0.4911	0.7445	0.5410	0.5142	0.2972	0.4846	0.8001
	$\hat{G}_b^{qb}$	0.6712	0.6001	0.4766	0.5616	0.8306	0.6278	0.5561	0.4526	0.5704	0.8294
	G_{lqb}^{distro}	**0.6939**	**0.6594**	**0.5054**	**0.6061**	**0.8406**	**0.6487**	**0.5593**	**0.4693**	**0.5796**	**0.8377**
Query Recommendations	$random$	0.0003	0.0002	0.0003	0.0004	0.0008	0.0003	0.0002	0.0003	0.0004	0.0008
	query-flow	0.0337	0.0190	0.0425	0.0477	0.1328	0.0629	0.0323	0.0799	0.0821	0.1901
	valueMerge	0.0216	0.0189	0.0152	0.0266	0.1069	0.0405	0.0284	0.0285	0.0398	0.1357
	G_{lqb}^{binary}	0.0340	0.0166	0.0371	0.0502	0.1303	0.0396	0.0334	0.0545	0.0693	0.1621
	G_{lqb}^{macro}	0.0150	0.0127	0.0090	0.0172	0.0787	0.0178	0.0133	0.0141	0.0237	0.1053
	$\hat{G}_q^{ql}$	0.0269	0.0146	0.0339	0.0369	0.1224	0.0566	0.0292	0.0721	0.0743	0.1919
	$\hat{G}_q^{qb}$	0.0364	0.0200	0.0459	0.0504	0.1388	0.0625	0.0336	0.0795	0.0855	0.1818
	G_{lqb}^{distro}	**0.0394**	**0.0211**	**0.0496**	**0.0530**	**0.1423**	**0.0672**	**0.0347**	**0.0855**	**0.0881**	**0.1998**

random selection; (2) $query\text{-}flow$ like model for query recommendation as defined in [37], and for location recommendation adapted by using the consecutive visits of Wi-Fi APs, giving an $ap\text{-}flow$ model. Similarly, for Web content recommendation, we build a $domain\text{-}flow$ model; (3) *topology* is a baseline for location recommendation. It generates the recommendation list based on the topology of the Wi-Fi network by using the rules of suggesting APs based on their topology distances to the current location l; (4) *valueMerge*: rather than using Eq. 19, following [16], *valueMerge* uses a value-based merging function: $r(q')_{G_{lqb}} = \theta \cdot r(q')_{\hat{G}_q^{ql}} + (1-\theta) \cdot r(q')_{\hat{G}_q^{qb}}$, where θ is a scaling factor, and it is obtained by cross-validation; (5) G_{lqb}^{binary}: is the LQB graph with *binary* projection approach (Eq. 15); (6) G_{lqb}^{macro}: is the LQB graph with *macro-aggregation* projection approach (Eq. 17); (7) $\hat{G}_l^{ql}$, $\hat{G}_l^{bl}$, $\hat{G}_b^{bl}$, $\hat{G}_b^{qb}$, $\hat{G}_q^{ql}$, and $\hat{G}_q^{qb}$: are methods, which make the recommendations based on the projected location/browsing/query graph of corresponding bipartite graph with *distributional* projection approach (Eq. 16). They are deployed to investigate the influence of physical and cyber contexts; (7) G_{lqb}^{distro}: is the LQB graph with *distributional* projection approach (Eq. 16).

5.3.2 Location Recommendation

Here, we recommend in terms Wi-Fi APs, suggesting where a user is likely to visit, based on their current location (AP). Specifically, as we discussed in Section 4.4, we project G_{ql} and G_{bl} into $\hat{G}_l^{ql}$ and $\hat{G}_l^{bl}$, respectively. Then, the random walk model is deployed for location recommendation. Note, because of the continuous movement, we treat location recommendations equivalent to location predictions here, suggesting based on where we predict a user will go next.

The results of location recommendations based on different methods are shown at the top of Table 4. It is observed that the LQB graph with *distributional* projection G_{lqb}^{distro} achieves the best performance in all evaluation metrics; $\hat{G}_l^{ql}$ and $\hat{G}_l^{bl}$ also outperform the standard *ap-flow* model. For example, given current location l, G_{lqb}^{distro} outperforms *ap-flow* by 14.5% in p@5, 15.6% in r@10, 9.1% in MRR. For the three projection approaches, G_{lqb}^{binary} performs slightly worse than G_{lqb}^{distro}, indicating that the linking relationship is important in the LQB graph; G_{lqb}^{distro} outperforms G_{lqb}^{macro} with a large gap, which is consistent to findings in [55]. The reason would be that users' behaviour data is very sparse across the physical and cyber spaces, and G_{lqb}^{macro} treats each

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2017.2766059, IEEE Transactions on Knowledge and Data Engineering

TABLE 5
Top 10 location recommendation for $l =$ "*ap28 (Jewellery)*" from *ap-flow*, $\hat{G}_l^{ql}$, $\hat{G}_l^{bl}$ and G_{lqb}^{distro}. The information in brackets are the categories of the shops covered by the corresponding APs.

ap-flow	$\hat{G}_l^{ql}$	$\hat{G}_l^{bl}$	G_{lqb}^{distro}
ap16 (Fashion)	ap16 (Fashion)	ap16 (Fashion)	ap16 (Fashion)
ap46 (Cafe, Fashion, Jewellery)	**ap46 (Cafe, Fashion, Jewellery**)	ap07 (Fashion)	**ap46 (Cafe, Fashion, Jewellery)**
ap53 (Fashion, Restaurant)	ap07 (Fashion)	ap35 (Fashion, Footwear)	ap07 (Fashion)
ap49 (Cafe, Fashion)	**ap21 (Fashion, Jewellery)**	ap31 (Footwear, Newsagents)	ap35 (Fashion, Footwear)
ap07 (Fashion)	ap53 (Fashion, Restaurant)	ap40 (Fashion, Footwear)	ap31 (Footwear, Newsagents)
ap21 (Fashion, Jewellery)	ap40 (Fashion, Footwear)	**ap38 (Footwear, Watches, Jewellery)**	ap40 (Fashion, Footwear)
ap44 (Takeaway)	ap43 (Takeaway)	ap60 (Takeaway)	**ap21 (Fashion, Jewellery)**
ap63 (Footwear, Hair & Beauty)	ap31 (Footwear, Newsagents)	**ap23 (Fashion, Jewellery, Bakeries)**	ap60 (Takeaway)
ap52 (Restaurant , Delicatessen)	ap35 (Fashion, Footwear)	**ap22 (Fashion, Jewellery)**	**ap23 (Fashion, Jewellery, Bakeries)**
ap59 (Takeaway)	**ap23 (Fashion, Jewellery, Bakeries)**	**ap46 (Cafe, Fashion, Jewellery)**	**ap38 (Footwear, Watches, Jewellery)**

user equally that introduces more noises. Moreover, the two-tailed, paired t-test with a 95% confidence level shows that G_{lqb}^{distro} significantly outperforms all the compared methods, which demonstrates that the contextual links in the LQB graph are more reliable than the consecutive AP-based links in location recommendation.

Table 5 shows the top 10 location recommendations for AP/location *ap28*, which covers only *Jewellery* shops. Specifically, *ap-flow* starts with focus on relevant jewellery locations, then lost in other popular categories, e.g. *Takeaway*, *Restaurant*, and *Footwear*. $\hat{G}_l^{ql}$ and $\hat{G}_l^{bl}$ perform better than *ap-flow*, with 3 and 4 relevant jewellery locations retrieved, respectively. However, $\hat{G}_l^{bl}$ ranked those *Jewellery* locations lower than those *Fashion* locations, which is possibly caused by the popularity of *Fashion* stores in the mall. Overall, it is observed that G_{lqb}^{distro} achieves the best results by ranking those *Jewellery* APs/locations highly, and obtaining 4 relevant locations.

5.3.3 Web Content Recommendation

Following [18], 1) other than predicting the Web domains, we present users' interests as a list of BrightCloud URL categories, which are assigned to the Web domains in users' browsing history; 2) some categories of Web domains are highly common, which would make the prediction of user interest either too easy or too difficult. We also removed the following categories of Web domains from our Web content recommendation analysis: social networks, search engines and portals. Note, they are removed at the recommendation stage, not at the construction of the LQB graph.

The results of Web content recommendation are shown in the middle of Table 4. We observe that G_{lqb}^{distro} achieves the highest accuracy in all measurement metrics, and it substantially outperforms all baseline models. For example, given current Web domain b, in terms of $r@5$, G_{lqb}^{distro} outperforms $random$ model by 105%, *domain-flow* by 38%, and *valueMerge* by 11%. G_{lqb}^{distro} also outperforms G_{lqb}^{binary} and G_{lqb}^{macro}, which is consistent to the results of location recommendations. We also note that $\hat{G}_b^{qb}$ outperforms $\hat{G}_b^{bl}$, and the reason might be that queries determines users browsing behaviours. This indicates that for Web content recommendation, the cyber contexts are more important than the physical contexts. In addition, we also observe that $\hat{G}_b^{bl}$ does not perform as high as *domain-flow*, but the final integrated results with $\hat{G}_b^{qb}$ is higher than both $\hat{G}_b^{qb}$ and *domain-flow*. This confirms that the physical contexts from $\hat{G}_b^{bl}$ complements the cyber contexts from $\hat{G}_b^{qb}$, which is consistent from the results on location recommendation. The paired t-test results show G_{lqb}^{distro} significantly outperforms all the compared methods for Web content recommendation.

5.3.4 Query Recommendation

The results of query recommendation are shown at the bottom of Table 4. Together with a paired-t-test, it is observed that the LQB graph (G_{lqb}^{distro}) also achieves statistically significantly better performance than all the compared models. Specifically, when only currently query is available, G_{lqb}^{distro} outperforms *query-flow* by 16.91% in terms of $p@5$, 16.95% in terms of $r@5$, and 7.19% in terms of MRR. Among G_{lqb}^{binary}, G_{lqb}^{macro} and G_{lqb}^{distro}, G_{lqb}^{macro} perform badly on query recommendations, and the reason is not many people issue queries in the mall environment [57], which makes the data extremely sparse here. Moreover, similarly like that in Web content recommendation, the cyber contexts appears to be more important than the physical contexts here, as $\hat{G}_q^{qb}$ outperforms $\hat{G}_q^{ql}$. However, the integration of them G_{lqb}^{distro} achieves better results than either of them.

Take the query recommendations for a real query as an example. Table 6 shows the top 10 suggested queries for query "virgin blue" (an Australian budget airline). *query-flow* initially suggests "ebookers", an online flight booking website, which is relevant. Subsequent recommendations soon lead to a lost of focus, suggesting frequent flight destinations (e.g. "New Zealand", "cheap cairns to melbourne flights") and other unrelated queries (e.g. "male halloween costume"). In contrast, $\hat{G}_q^{ql}$ and $\hat{G}_q^{qb}$, lead to query recommendations nuanced by the physical and cyber contexts, respectively. Using $\hat{G}_q^{ql}$ leads to recommendations of queries related to physical locations and physically close stores; while $\hat{G}_q^{qb}$ suggested relevant queries targeting competing airlines (e.g. jal (Japanese Airline), qantas, tiger, singapore airlines), as well as a number of irrelevant queries (e.g. yahoo, la senza triple gel bikini). However, the combination of the three influences in G_{lqb}^{distro} leads to a set of query recommendations that are all topic relevant. Specifically, airline related queries are ranked higher than "peter pan", which is searched for the travel website *peterpans.com*.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2017.2766059, IEEE Transactions on Knowledge and Data Engineering

TABLE 6
Top 10 query recommendation for $q =$ "*virgin blue*" from *query-flow*, $\hat{G}_q^{ql}$, $\hat{G}_q^{qb}$ and G_{lqb}^{distro}.

query-flow	$\hat{G}_q^{ql}$	$\hat{G}_q^{qb}$	G_{lqb}^{distro}
virgin blue	virgin blue	virgin blue	virgin blue
ebookers	sydney- australie	jal	jal
new zealand	cashier	qantas flights	qantas flights
cheap oz flights	hilton sydney	quantus	ebookers
flights australia	mantra on kent	virgin australia	cheap oz flights
worth seeing new zealand	city westfield	virgin	virgin australia
cheap cairns to melbourne flights	wedding dresses sydney	tiger airways	quantus
peter pan	fitness first	yahoo	flights australia
peter pan adventures	peter pan	singapore airlines	tiger airways
male halloween costume	princess polly	la senza triple gel bikini	peter pan

6 Discussion and Conclusions

We have proposed a heterogeneous LQB graph as a representation of the interactive knowledge about people's behaviour across the physical and cyber spaces. We have highlighted the utility of the LQB graph method in an indoor retail scenario, based on the analysis of a large dataset capturing the indoor physical and Web activity of registered WiFi users. Following an analysis on the contextual influence on people's information and physical behaviour, we confirm the strong inter-dependencies between people's querying, browsing and spatial behaviours, as previously suggested [20], [26], [30]. In contrast to previous literature, we explored these interdependencies in a constrained and controlled physical environment (a shopping mall) and across all three contextual influences. To do this, we populated the locations subgraph of the LQB graph with Wi-Fi AP associations, the query subgraph with queries, and the browse graph with URL domains. We have then shown that the tripartite LQB graph successfully models the physical and Web content aspects of context and outperforms the state-of-the-art models in location, query and Web content recommendation. The proposed LQB graph model significantly outperforms even the baselines achieved using the partial graphs as well as the more naive models.

We have also quantified the relative impact of query, browsing and location activity on each other. Simply put, the capture of multifaceted contextual information improves recommendations, but not equally. The physical location recommendation depends more strongly on browsing behaviour, while both browsing and query behaviour influence each other more than the physical location where querying and browsing occurs. This may mean that immediate information needs in a physical location are satisfied by browsing or navigating to a known URL, while more exploratory information behaviours, even if triggered at a certain location (e.g., is this the best price for this item) are satisfied at a different, deffered time and location (e.g., querying for alternatives and specifications in the food court). While previous work also hinted at the importance of social contextual factors [26], [30], this aspect could not be sufficiently tested with the dataset at hand and remains subject to future research.

The proposed model so far contains no elements of personalization, yet it is able to produce significantly better recommendations from a heterogeneous pool of suggested items (locations, queries and browsing content), outperforming existing baselines. This is important for numerous scenarios, such as recommendations in retail environments where a large number only visit once ($> 70\%$ in our dataset). The cold start problem is a significant hurdle for better recommendations for physical retail environment operators, struggling to remain competitive against their online competition. Another advantage is the LQB graph can be constructed incrementally as logs capture more queries, Web domain or as the space is remodelled and location added. This is because the LQB graph is built by sequential processing of the logs. However, it is necessary to update the corresponding graph projections for ranking-based recommendations. In future work, we plan to enrich our model to improve recommendations for frequent visitors where personalization and the capture of the social ties may be possible.

Acknowledgments

This research is supported by ARC LP120200413.

References

[1] M. B. Jansen, A. Spink, J. Bateman, and T. Saracevic, "Real Life Information Retrieval: A Study Of User Queries On The Web," *ACM SIGIR Forum*, vol. 32, no. 1, pp. 5–17, 1998.

[2] B. J. Jansen, "Real life, real users, and real needs: a study and analysis of user queries on the web," *Information Processing and Management*, vol. 36, no. 2, pp. 207–227, Mar. 2000.

[3] A. Spink, D. Wolfram, M. B. Jansen, and T. Saracevic, "Searching the Web: The Public and Their Queries," *J. Am. Soc. Inf. Sci. Technol.*, vol. 53, no. 3, pp. 226–234, 2001.

[4] D. Wolfram, A. Spink, B. J. Jansen, and T. Saracevic, "Vox populi: The public searching of the web," *J. Am. Soc. Inf. Sci. Technol.*, vol. 52, no. 12, pp. 1073–1074, 2001.

[5] A. Spink, B. Jansen, D. Wolfram, and T. Saracevic, "From e-sex to e-commerce: Web search changes," *Computer*, vol. 35, no. 3, pp. 107–109, 2002.

[6] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz, "Analysis of a Very Large Web Search Engine Query Log," in *ACM SIGIR Forum*. New York, NY, USA: ACM, 1999, pp. 6–12.

[7] M. Sanderson and J. Kohler, "Analyzing geographic queries," in *Workshop on Geographic Information Retrieval*, Sheffield, UK, 2004.

[8] Q. Gan, J. Attenberg, A. Markowetz, and T. Suel, "Analysis of Geographic Queries in a Search Engine Log," in *Proceedings of the 1st Workshop on Location and the Web*. ACM, 2008, pp. 49–56.

[9] S. Aloteibi and M. Sanderson, "Analyzing geographic query reformulation: An exploratory study," *J. Am. Soc. Inf. Sci. Technol.*, vol. 65, no. 1, pp. 13–24, 2014.

[10] R. Wan-chik, P. Clough, and M. Sanderson, "Investigating religious information searching through analysis of a search engine log," *JASIST*, vol. 64, no. 12, pp. 2492–2506, 2013.

[11] S. Pandey, K. Punera, M. Fontoura, and V. Josifovski, "Estimating Advertisability of Tail Queries for Sponsored Search," in *SIGIR*. New York, New York, USA: ACM Press, 2010, pp. 563–570.

[12] M. Kamvar and S. Baluja, "A large scale study of wireless search behavior: Google mobile search," in *SIGCHI*, 2006, pp. 701–709.

[13] K. Church, B. Smyth, P. Cotter, and K. Bradley, "Mobile information access: A study of emerging search behavior on the mobile Internet," *ACM Trans. Web*, vol. 1, no. 1, May 2007.

[14] K. Church, B. Smyth, K. Bradley, and P. Cotter, "A large scale study of European mobile search behaviour," in *MobileHCI*, ACM. ACM, 2008, pp. 13–22.

[15] E. Agichtein, E. Brill, and S. Dumais, "Improving web search ranking by incorporating user behavior information," *SIGIR '06*, p. 19, 2006.

[16] Y. Liu, B. Gao, T.-Y. Liu, Y. Zhang, Z. Ma, S. He, and H. Li, "Browserank: Letting web users vote for page importance," in *SIGIR'08*. New York, NY, USA: ACM, 2008, pp. 451–458.

[17] R. W. White and J. Huang, "Assessing the scenic route: Measuring the value of search trails in web logs," in *SIGIR*, 2010, pp. 587–594.

[18] R. W. White, P. Bailey, and L. Chen, "Predicting user interests from contextual information," in *SIGIR '09*. ACM, 2009, pp. 363–370.

[19] M. Tsagkias and R. Blanco, "Language intent models for inferring user browsing behavior," in *SIGIR*. ACM, 2012, pp. 335–344.

[20] L. Chiarandini, M. Trevisiol, and A. Jaimes, "Discovering social photo navigation patterns," in *ICME*, 2012, pp. 31–36.

[21] L. Chiarandini, P. Grabowicz, M. Trevisiol, and A. Jaimes, "Leveraging Browsing Patterns for Topic Discovery and Photostream Recommendation." in *ICWSM*, 2013, pp. 71–80.

[22] M. Trevisiol, L. Chiarandini, L. M. Aiello, and A. Jaimes, "Image ranking based on user browsing behavior," in *SIGIR*, 2012, pp. 445–454.

[23] M. Kamvar and S. Baluja, "Deciphering trends in mobile search," *IEEE Computer*, vol. 40, no. 8, pp. 58–62, 2007.

[24] T. Sohn, K. A. Li, W. G. Griswold, and J. D. Hollan, "A Diary Study of Mobile Information Needs," in *SIGCHI*, 2008, pp. 433–442.

[25] A. M. Hinze, C. Chang, and D. M. Nichols, "Contextual Queries express Mobile Information Needs Categories and Subject Descriptors," in *MobileHCI*. ACM, 2010, pp. 327–336.

[26] 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*, 2011, pp. 77–80.

[27] A. Y. K. Chua, R. S. Balkunje, and D. H.-L. Goh, "Fulfilling Mobile Information Needs : A Study on the Use of Mobile Phones," in *IMCOM*. ACM, 2011, pp. 92:1–92:7.

[28] Y. Song, H. Ma, H. Wang, and K. Wang, "Exploring and Exploiting User Search Behavior on Mobile and Tablet Devices to Improve Search Relevance," in *WWW*. ACM, 2013, pp. 1201–1212.

[29] D. Lymberopoulos, P. Zhao, C. König, K. Berberich, and J. Liu, "Location-aware Click Prediction in Mobile Local Search," in *CIKM*. ACM, 2011, pp. 413–422.

[30] E. Yom-Tov and F. Diaz, "Out of Sight , Not Out of Mind: On the Effect of Social and Physical Detachment on Information Need," in *SIGIR*, 2011, pp. 385–394.

[31] A. Zhang, A. Goyal, R. Baeza-Yates, Y. Chang, J. Han, C. A. Gunter, and H. Deng, "Towards mobile query auto-completion: An efficient mobile application-aware approach," in *WWW*, 2016, pp. 579–590.

[32] Z. Zhuang, C. Brunk, and C. L. Giles, "Modeling and visualizing geo-sensitive queries based on user clicks," in *LOCWEB '08*, 2008, pp. 73–76.

[33] D. H. l. Goh, E. p. Lim, A. Sun, D. Wu, and W. Zong, "On assigning place names to geography related web pages," in *JCDL '05*, June 2005, pp. 354–362.

[34] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak, "Spatial variation in search engine queries," in *WWW*, 2008, pp. 357–366.

[35] R. Baeza-Yates, "Graphs from search engine queries," in *SOFSEM*. Springer-Verlag, 2007, pp. 1–8.

[36] Z. Zhang and O. Nasraoui, "Mining search engine query logs for query recommendation," in *WWW*, 2006, pp. 1039–1040.

[37] P. Boldi, F. Bonchi, and C. Castillo, "The query-flow graph: model and applications," in *CIKM*, 2008, pp. 609–617.

[38] P. Boldi, F. Bonchi, C. Castillo, D. Donato, and S. Vigna, "Query suggestions using query-flow graphs," in *WSCD '09*. New York, NY, USA: ACM, 2009, pp. 56–63.

[39] I. Bordino, C. Castillo, D. Donato, and A. Gionis, "Query similarity by projecting the query-flow graph," in *SIGIR '10*. New York, NY, USA: ACM, 2010, pp. 515–522.

[40] M.-D. Albakour, U. Kruschwitz, I. Adeyanju, D. Song, M. Fasli, and A. De Roeck, "Enriching query flow graphs with click information," in *AIRS'11*. Springer-Verlag, 2011, pp. 193–204.

[41] D. Beeferman and A. Berger, "Agglomerative clustering of a search engine query log," in *SIGKDD*, 2000, pp. 407–416.

[42] N. Craswell and M. Szummer, "Random walks on the click graph," in *SIGIR '07*. ACM, 2007, pp. 239–246.

[43] Q. Mei, D. Zhou, and K. Church, "Query suggestion using hitting time," in *CIKM*, 2008, pp. 469–478.

[44] J. Zhang, L. Jie, A. Rahman, S. Xie, Y. Chang, and P. S. Yu, "Learning entity types from query logs via graph-based modeling," in *CIKM'15*. ACM, 2015, pp. 603–612.

[45] S. Qi, D. Wu, and N. Mamoulis, "Location aware keyword query suggestion based on document proximity," *IEEE TKDE*, vol. 28, no. 1, pp. 82–97, Jan 2016.

[46] Y. Liu, T.-Y. Liu, B. Gao, Z. Ma, and H. Li, "A framework to compute page importance based on user behaviors," *Inf. Retr.*, vol. 13, no. 1, pp. 22–45, Feb. 2010.

[47] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," in *WWW*, 1998, pp. 161–172.

[48] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen, "Combating web spam with trustrank," in *VLDB '04*, 2004, pp. 576–587.

[49] Y. Liu, Y. Jin, M. Zhang, S. Ma, and L. Ru, "User Browsing Graph: Structure, Evolution and Application," in *WSDM'09*, 2009.

[50] M. Trevisiol, L. M. Aiello, R. Schifanella, and A. Jaimes, "Cold-start news recommendation with domain-dependent browse graph," in *RecSys'14*. ACM, 2014, pp. 81–88.

[51] M. Trevisiol, L. M. Aiello, P. Boldi, and R. Blanco, "Local ranking problem on the browsegraph," in *SIGIR*, 2015, pp. 173–182.

[52] Y. Sun, H. Yin, and X. Ren, "Recommendation in context-rich environment: An information network analysis approach," in *WWW '17 Companion*, 2017, pp. 941–945.

[53] M. Xie, H. Yin, H. Wang, F. Xu, W. Chen, and S. Wang, "Learning graph-based poi embedding for location-based recommendation," in *CIKM*. ACM, 2016, pp. 15–24.

[54] S. Zhao, T. Zhao, I. King, and M. R. Lyu, "Geo-teaser: Geo-temporal sequential embedding rank for point-of-interest recommendation," in *WWW '17 Companion*, 2017, pp. 153–162.

[55] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme, "Evaluating similarity measures for emergent semantics of social tagging," in *WWW*. ACM, 2009, pp. 641–650.

[56] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu, *Spatial Tesselations: Concepts and Applications of Voronoi Diagrams*, 2nd ed., ser. Wiley Series in Probability and Statistics. John Wiley and Sons, 1999.

[57] Y. Ren, M. Tomko, F. D. Salim, K. Ong, and M. Sanderson, "Analyzing web behavior in indoor retail spaces," *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 1, pp. 62–76, Jan. 2017.

[58] Y. B. Bai, S. Wu, Y. Ren, K. Ong, G. Retscher, A. Kealy, M. Tomko, H. Wu, and K. Zhang, "A New Approach for Indoor Customer Tracking Based on a Single Wi-Fi Connection," in *IPIN*, 2014.

[59] R. Kumar and A. Tomkins, "A Characterization of Online Browsing Behavior," in *WWW*, 2010, pp. 561–570.

[60] A. Dean-Hall, C. L. A. Clarke, J. Kamps, J. Kiseleva, and E. M. Voorhees, "Overview of the TREC 2015 contextual suggestion track," in *TREC*, 2015.

Yongli Ren is a Lecturer at the School of Science, at RMIT University in Melbourne, Australia. He has a PhD degree in Information Technology from Deakin University, Australia.

Martin Tomko is a Lecturer at the Department of Infrastructure Engineering of the University of Melbourne, Australia. He holds a PhD in Geomatics from the University of Melbourne, Australia.

Flora Salim is a Senior Lecturer at the School of Science, RMIT University. She obtained her PhD in Computer Science from Monash University.

Jeffrey Chan is a Lecturer at the School of Science, at RMIT University in Melbourne, Australia. He holds a PhD in computer science from the University of Melbourne, Australia.

Charles L. A. Clarke is a Professor in the Cheriton School of Computer Science at the University of Waterloo in Canada. He has published in question answering, XML, filesystem search, user interfaces, statistical NLP, and the evaluation of information retrieval systems. He received his Ph.D. from Waterloo in 1996.

Mark Sanderson is a Professor at the School of Science, at RMIT University in Melbourne, Australia. His research focuses on information retrieval, bibliometrics and user analytics. He is associate editor of ACM Transactions on the Web and IEEE Transactions on Knowledge and Data Engineering.