Graph-Based Recommendation Integrating Rating History and Domain Knowledge: Application to On-Site Guidance of Museum Visitors

Einat Minkov
Information Systems Department, University of Haifa, Haifa 31905, Israel. E-mail: einatm@is.haifa.ac.il

Keren Kahanov
Computer Science Department, University of Haifa, Haifa 31905, Israel. E-mail: kahanovk@gmail.com

Tsvi Kuflik
Information Systems Department, University of Haifa, Haifa 31905, Israel. E-mail: tsvikak@is.haifa.ac.il

Visitors to museums and other cultural heritage sites encounter a wealth of exhibits in a variety of subject areas, but can explore only a small number of them. Moreover, there typically exists rich complementary information that can be delivered to the visitor about exhibits of interest, but only a fraction of this information can be consumed during the limited time of the visit. Recommender systems may help visitors to cope with this information overload. Ideally, the recommender system of choice should model user preferences, as well as background knowledge about the museum's environment, considering aspects of physical and thematic relevancy. We propose a personalized graph-based recommender framework, representing rating history and background multi-facet information jointly as a relational graph. A random walk measure is applied to rank available complementary *multimedia presentations* by their relevancy to a visitor's profile, integrating the various dimensions. We report the results of experiments conducted using authentic data collected at the Hecht museum.[1] An evaluation of multiple graph variants, compared with several popular and state-of-the-art recommendation methods, indicates on advantages of the graph-based approach.

Introduction

Visitors to museums and other cultural heritage (CH) sites can be overwhelmed by the richness and diversity of the information items that these sites offer (Davey, 2005). Visitors therefore need assistance in getting the best experience from their visit. Obviously, visitors differ in their preferences, and expectations. CH recommender systems aim at generating personalized recommendations that fit the visitors' individual preferences and needs. Such personalized services can be implemented using dedicated mobile applications (Ardissono, Kuflik, & Petrelli, 2012).

At present, mobile devices are typically available at CH sites, offering complementary information about exhibits of interest, albeit not in a personalized manner. Importantly, personalized information can be both delivered and collected as part of the interaction with the mobile device, having feedback on viewed items collected explicitly, or in a non-intrusive manner. For example, it is possible to track the user's behavior by analyzing signals transmitted by her mobile device (Dim & Kuflik, 2015; Kuflik, Kay, & Kummerfeld, 2012).

Nevertheless, making personalized recommendations at the museum is a challenging problem. Crucially, the collected feedback information is sparse: every user gets to view and provide feedback for a small number of items, and in the majority of cases, the visitor is introduced to the museum for the first time (Biran, Poria, & Oren, 2011). The recommender system thus operates in continuous *cold start* conditions. Further, the recommended items are not standalone artifacts—they are directly associated with some exhibit in the museum, which in turn is located in a specific room. In order for recommendations to be effective, the system must consider this location context. Additional useful *background knowledge* may map the museum's environment and items into a semantic space, for example, associating exhibits with

Received February 18, 2016; revised September 11, 2016; accepted November 3, 2016

 • Published online 0 Month 2017 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/jasist.23837

[1]http://mushecht.haifa.ac.il/

specific themes. Semantic modeling is especially important considering the sparsity of historical ratings. In order to integrate visitor's feedback with physical and semantic contexts, the recommender system must effectively consolidate such heterogeneous information.

We outline and evaluate a graph-based recommendation approach that handles these challenges gracefully. Multi-source information is represented using a heterogeneous graph scheme, in which typed nodes denote entities, and directed and typed edges denote inter-entity relations. Concretely, the graph nodes denote *users* and *multimedia presentation*[2], as well as physical *positions* and semantic *themes*. The graph edges denote structured relations, for example, *located-in* (between *presentation* items and the *positions* in which they are offered) or *viewed* relations (between *users* and the *presentations* that they rated). Edges further denote elicited relations, such as similarity between *presentations* induced based on their content.

The graph-based recommendation process involves inference of node relevancy with respect to a *query*, defined as a distribution over the graph nodes. We apply the Personalized Page Rank (PPR) algorithm (Haveliwala, 2002) to rank *presentations* by their relatedness to a user profile represented as the set of *presentations* already viewed and liked by the user. PPR applies a random walk procedure, assessing inter-node relatedness from a global perspective, thus alleviating the sparsity problem that challenges recommender systems.

The article reports the results of a case study using authentic data obtained at the Hecht Museum, located at the University of Haifa. Following the deployment of a mobile visitors guide system at the museum, data have been collected in the form of visit logs for research purposes (Kuflik, Wecker, Lanir, & Stock, 2014). Given user feedback on viewed *multimedia presentations*, yet unviewed *presentations* are ranked according to the user's interests. We report a set of comparative experiments, showing superiority of graph-based recommendation over popular and state-of-the-art recommendation approaches.

There are several main contributions of this work:

- We show that the graph-based framework delivers accurate recommendations in the challenging CH domain. Unlike alternative methods, this approach models historical ratings jointly with diverse background knowledge, including contextual physical proximity and semantic aspects.
- Relatively few works employed graph based similarity in general, and the PPR measure in particular, for recommendation purposes. We report the results of a comprehensive set of comparative experiments, demonstrating the advantages of the graph-based approach over alternative methods in contextual recommendation setting.
- We evaluate and discuss issues related to graph design, as well as the impact of tuning parametric edge weights on recommendation performance.

[2]Henceforth, we use the terms *multimedia presentations* and *presentations* interchangeably

Background and Related Work

Let us formally define the recommendation problem. Recommender systems estimate the relevancy of yet unseen *items* for individual *users*. We denote the set of users by U, and the finite set of items by I. Let I_u represent the subset of items that have been viewed and rated by an individual user $u \in U$. The rating assigned by user u to item $i \in I_u$ is denoted by r_{ui}. This information corresponds to a sparsely populated matrix, with known users and items as the matrix dimensions, and historical ratings as values. The set of available feedback scores for a given user, $\{r_{ui}, i \in I_u\}$, represents her preferences. Given this rating history, it is desired to predict ratings for the remaining items that the user has yet to experience, $\{I - I_u\}$.

Classical recommendation approaches are roughly categorized into content-based (CB) and collaborative filtering (CF) (Adomavicius & Tuzhilin, 2005, 2011). These approaches face the "cold start" problem, handling new items (both approaches) and new users (CF), for which little rating history exists, as well as general rating sparsity. Unfortunately, classical approaches do not support easy integration of context information into the recommendation model (Konstas, Stathopoulos, & Jose, 2009). In order to model background information about the user, a CF method based on Tensor Factorization has been proposed (Karatzoglou, Amatriain, Baltrunas, & Oliver, 2010), which models the ratings data as a user-item-context N-dimensional tensor. Such modeling comes with the cost of increased sparsity–as it considers a multidimensional space instead of the (already sparse) two dimensional one. Others suggested to partition the user-item rating matrix into groups of ratings with similar contexts (Liu & Aberer, 2013). None of these extensions readily accommodates complex, relational background information, such as the information modeled in this work.

Graph-Based Recommendation

Multiple studies have previously explored graph-based methods for recommendation. Early works used homogenous graphs, in which nodes denoted *items* and edges represented inter-item similarity, for example, (Gori & Pucci, 2007). Other works modeled bipartite graphs, in which edges connected *user* nodes to nodes denoting *items* that they rated (Baluja et al., 2008; Fouss, Pirotte, Renders, & Saeren, 2007).

More recently, researchers have started to explore heterogeneous graph representations that consist of multiple node and edge types for context-aware recommendation (Bagci & Karagoz, 2016; Bu et al., 2010; Noulas, Scellato, Lathia, & Mascolo, 2012; Pham, Li, Cong, & Zhang, 2015; Shang, Kulkarni, Cuff, & Hui, 2012; Tiroshi et al., 2014; Wang, Terrovitis, & Mamoulis, 2013). In addition to user-item ratings, the constructed graphs typically include social relations between *users* and link users to social *tags* extracted from domain-specific social networks. Here, we show that the graph-based approach can effectively model diverse

DOI: 10.1002/asi

relational background knowledge, also representing aspects such as physical and thematic relatedness.

Several previous works used the PPR measure to generate a ranked list of items, having the user's preferences represented as a distribution of interest over the graph nodes. Some of these works compared PPR with alternative recommendation methods. Specifically, Konstas et al. (2009) considered recommending music tracks using a graph that included social annotations and friendship relations, and showed improvements using PPR over a user-based kNN method. Yao, He, Huang, Cao, and Zhang (2015) proposed a multilayer graph representation for contextual recommendation, and also showed superiority of PPR over a user-based kNN approach. Noulas et al. (2012) targeted location recommendation, modeling historical user-location visits and user friendships in the graph. The distribution of visits across locations was reported to be very skewed, having the PPR measure be the only method to beat a strong popularity-based baseline.

Nevertheless, we find that to date, a comprehensive comparison against classical recommendation approaches, which have been developed over the course of several decades, is still at need. A main contribution of this study is empirical comparison of PPR graph-based recommendation with multiple classical and state-of-the-art recommendation methods. Consequently, we provide stronger evidence about the advantages and applicability of PPR and similar methods for the recommendation task.

Interestingly, none of the aforementioned works examined graph adaptation by means of edge weight tuning (e.g., Minkov & Cohen, 2010). We show that tuning parametric task-specific edge weights substantially affects the global similarity measure, leading to performance gains.

Recommendation Systems for Museum Visitors

Several recommendation techniques and systems have been designed to enhance the museum visit experience. The overlay model (Stock et al., 2007) "overlays" the user preferences over a domain ontology, and propagates node ratings over the ontology to other exhibits of interest. Grieser, Baldwin, and Bird (2007) aimed at predicting which exhibit a visitor would visit next based on her visit history using a Naive Bayes learning model, taking into consideration exhibit proximity, textual description of the exhibit, and exhibit popularity. The approach proposed here is more comprehensive, as it models collaborative user history jointly with content-based and physical proximity aspects. We believe that the graph-based approach is advantageous to statistical learning in conditions of data sparsity.

The "Geckommender" system (Bohnert, Zukerman, & Laures, 2012) uses a nearest-neighbor CB approach to predict exhibit ratings and generate theme/tour recommendations. Geckommender was only evaluated with respect to different display modes of the predictions. Another work (Bohnert & Zukerman, 2014) targets the prediction of exhibit viewing times using the Spatial Process Model, a collaborative model based on the theory of spatial processes. They model the correlation between observed viewing times in terms of exhibit conceptual distance, encoded in a covariance matrix. Several types of distances are evaluated, including viewing-time similarity, semantic similarity and walking distance. The results, evaluated using 157 visitor histories at the Melbourne Museum, indicate that physical distances yield the best predictions of viewing times. The graph-based approach described here is complementary to this model– one may generate multifacet correlation scores using the global graph-based measure.

Finally, Bartolini et al. (2014) target the recommendation of diverse multimedia materials across cultural heritage sites. They use a graph to represent item similarity, based on semantic annotations and "visiting patterns," indicating the frequency in which two items were consumed consecutively by the same visitor. While they organize the recommended items into paths, the physical aspect as well as historical ratings are not integrated in the graph. Their evaluation focuses on assessing visitor satisfaction in field conditions, whereas we focus on comparing multi-facet graph-based recommendation with classical approaches.

Graph-Based Recommendation at the Museum

In this section, we first formulize the graph representation schema, and describe how it is applied to the case study of item recommendation to museum visitors. We then outline the PPR algorithm, and provide intuitions on why using PPR is beneficial in our contextual recommendation settings.

The Museum as a Graph

A graph $G = < V, E >$ consists of a set of nodes V, and a set of labeled and directed edges E. We denote nodes by lower-case letters such as x, y, or z. Every node x has a type, denoted $\tau(x)$. The set of possible types is pre-determined and fixed. An edge from x to y is typed with relation ℓ, denoted as $x \xrightarrow{\ell} y$. Typically, for every edge in the graph, there exists an edge going in the other direction, denoting an inverse relation. This implies that the graph is cyclic and highly connected.

We describe the museum's environment using the following classes of entities, represented using distinct node types:

- *Positions*. Physical *points of interest* (POIs), in which multimedia information is available about exhibits nearby. The POIs are spread in the museum environment over multiple rooms and floors.
- *Presentations*. Multimedia presentations are offered for viewing on the visitor's mobile device. Presentations are associated with concrete exhibits. Once the user is tracked at some POI, she is offered to view presentations about exhibits associated with that position.
- *Themes*. The multimedia presentations in Hecht Museum have been associated with specific themes, for example, Religions, or Art symbols and Maritime (Katz et al., 2006). We represent each of these themes as a node in the graph.

DOI: 10.1002/asi

TABLE 1. Relation types in the museum graph.

Source node type		Edge type		Target node type
presentation	(309)	located-in	(1:1)	*position*
		has-theme	(1:1)	*theme*
		similar-to	(1:*)	*presentation*
		viewed^{-1}	(1:*)	*visitor*
position	(49)	nearby	(1:*)	*position*
		located-in^{-1}	(1:*)	*presentation*
theme	(9)	has-theme^{-1}	(1:*)	*presentation*
		similar-theme	(1:*)	*theme*
visitor	(287)	viewed	(1:*)	*presentation*

Note: The table includes the total number of nodes of each type in our experimental data set, as well as relation cardinality (1:1 denotes a functional relation, and 1:* implies that the source node may be connected to any number of nodes over that relation.) we bound relation cardinality as part of graph design, as described in Data.

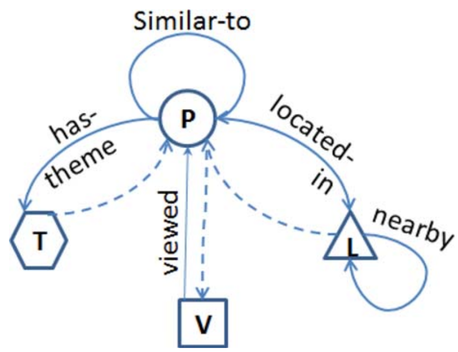


FIG. 1. A schematic view of the relation types in the museum graph. The different node shapes denote entity types, including: *Presentations* (P), *positions* (L), *themes* (T), and *visitors* (V). Dashed lines denote inverse edges. [Color figure can be viewed at wileyonlinelibrary.com]

- *Visitors*. These nodes represent individual visitors. Although the other node types describe static aspects of the museum, historical information about visitors is dynamic, being accumulated over time.

The graph edge types are detailed in Table 1, and illustrated schematically in Figure 1. We directly link each *presentation* with the *position* with which it is associated over an edge of type *located-in*. *Has-theme* edges link each *presentation* with the respective *theme* node. Both of these relation types are functional, having each presentation map to a single position and theme. In order to maintain high connectivity in the graph, edges are added in the opposite direction between the respective node pairs, denoting the inverse relations *located-in*$^{-1}$ and *has-theme*$^{-1}$.

We further model aspects of inter-entity similarity in the museum's environment. *Positions* that reside in high physical proximity are linked over *nearby* edges. In addition, *presentations* that exhibit high content similarity are interlinked over the *similar-to* relation, and *similar-theme* edges connect semantically related *themes*. Details about the computation of inter-item similarity and derivation of the respective edges are given in Impact of Graph Tuning.

It is straightforward to incorporate historical ratings in the graph. Each visitor is represented by a dedicated *visitor* node, linked to *presentation* nodes over directed *viewed* edges. One may link a visitor to all of the presentations that she viewed, or, only to presentations that she is known to have liked. We follow the first option in this work, in light of data sparsity. Node pairs linked over the *viewed* relation, are linked also over inverse edges of the type *viewed*$^{-1}$.

Recommendation with Personalized PageRank

We wish to recommend presentations to a user based on her historical feedback. This task corresponds to the following objective: given a distribution over the graph nodes that represents the user's preferences, *presentation* nodes are to be ranked by their relatedness to that distribution. Various measures exist that evaluate node relatedness in graphs (Fouss et al., 2007). We employ the popular PPR random walk metric (Page, Brin, Motwani, & Winograd, 1998; Richardson & Domingos, 2002), sometimes referred to as Random Walk with Restart (RWR) (Tong, Faloutsos, & Pan, 2006).

PPR applies a Markovian random walk process that follows the well-known non-personalized PageRank algorithm (Page et al., 1998). Given that the walker is at node *i*, with probability α the walker follows an outgoing link from *i*, and with probability $(1-\alpha)$ the walker *resets* randomly to some graph node. The probability distribution of finding the walker at each of the graph nodes at time *d*, V_d, is defined recursively as:

$$V_{d+1}=(1-\alpha)V_u+\alpha \mathbf{M} V_d \tag{1}$$

where the total number of nodes is *N*, and the transition matrix $\mathbf{M}$ encodes the probability that the walker moves to page *j* from page *i* following an outgoing edge. As default, $\mathbf{M}$ distributes a node's probability uniformly among the pages it links to. Importantly, PPR preserves an association between node rankings and user preferences, or a "query," having V_u denote the query distribution. That is, PPR limits the reset operation to the query nodes.

This random walk process is guaranteed to converge to a unique stationary distribution. The resultant node scores reflect their structural similarity, or relevancy, with respect to the query V_u. It has been shown that the PPR score for a target node *z* and a query node *x* equals a summation over all the paths between *x* and *z*, weighted by path traversal probabilities (Jeh & Widom, 2003). Because of the reset operation, the paths between *x* and *z* are weighted exponentially lower as their length increases. This means that items that are connected over short paths to the query nodes are considered more relevant by the PPR method; similarly, items reached over multiple paths from the query nodes are also considered more relevant.

In our work, the query distribution V_u spans over the set of *presentation* nodes that the visitor is known to have liked, weighted by the respective visitor's feedback scores. The computed PPR scores reflect structural relatedness (or, similarity) of the graph nodes to this visitor's profile. We present a list of *presentation* nodes ranked by their estimated relevancy to the user. Likewise, one may generate rankings of other entity types, for example, *positions*.

Edge weight tuning. The graph walk process is determined by the graph's topology, captured by the transition matrix **M**. It is reasonable to assume that specific edges reflect more meaningful relations. We will assume that edge importance is derived from its type (Minkov & Cohen, 2010; Shang et al., 2012). Concretely, a set of edge weight parameters Θ determines for every edge of type ℓ in the graph, a fixed weight $\theta_\ell \in \Theta$. The transition probability from node x to node y over a single time step, $\mathbf{M}_{x,y}$, is defined accordingly as:

$$\mathbf{M}_{x,y} = \frac{\theta_\ell}{\sum_{y\prime \in ch(x)} \theta_{\ell\prime}} \tag{2}$$

where $ch(x)$ denotes the set of children of x (the nodes reachable from x in one time step), θ_ℓ is the weight of the outgoing edge from x to y, and similarly $\theta_{\ell\prime}$ is the weight of any outgoing edge of type $\ell\prime$ leading from x to a child node $y\prime$. In words, the probability of reaching node y from x is defined as the proportion of the edge weight from x to y out of the total outgoing weight from the parent x.

The edge weights Θ can be set manually, according to prior beliefs; tuned empirically; or, learned from labeled examples (Minkov & Cohen, 2010). Here, we empirically tune the edge weights using exhaustive search. We leave the exploration of edge weight learning, as well as path-based node ranking schemes (Lao, Minkov, & Cohen, 2015) for future work.

Discussion. The graph walk process naturally integrates multiple types of evidence. Initiating at some *presentation x* that is included in the user's profile V_u, it will reach a related *presentation z* via a shared *theme* node, or directly, because of the modeling of text-based similarity edges, over the following paths: $\mathrm{x} \xrightarrow{has-theme} \mathrm{y} \xrightarrow{has-theme^1} \mathrm{z}$, and $\mathrm{x} \xrightarrow{similar-to} \mathrm{z}$. Importantly, the PPR measure is transitive—as the random walk continues, similarity propagates between pairs of related *presentations* continuously. Accordingly, immediate physical proximity (same position) is expressed via the following 2-hop path: $\mathrm{x} \xrightarrow{located-in} \mathrm{y} \xrightarrow{located-in^1} \mathrm{z}$; and extended physical proximity corresponds to a 3-hop path $\mathrm{x} \xrightarrow{located-in} \mathrm{y} \xrightarrow{nearby} \mathrm{s} \xrightarrow{located-in^1} \mathrm{z}$., or 4-hop path $\mathrm{x} \xrightarrow{located-in} \mathrm{y} \xrightarrow{nearby} \mathrm{s} \xrightarrow{nearby} \mathrm{t} \xrightarrow{located-in^1} \mathrm{z}$., etc.

We note that in case that one is interested in biasing node rankings in favor of the current location of the visitor, this can be achieved by including the respective *position* node in the query, or, by weighting the queried *presentation* nodes by recency, assigning higher weight to the *presentation(s)* associated with the position visited last.

In addition, collaborative aspects are modeled through the path: $\mathrm{x} \xrightarrow{viewed^{-1}} \mathrm{y} \xrightarrow{viewed} \mathrm{z}$. Collaborative and content-based similarities are naturally integrated by mixed paths like $\mathrm{x} \xrightarrow{viewed^{-1}} \mathrm{y} \xrightarrow{similar-to} \mathrm{s} \xrightarrow{viewed} \mathrm{z}$.

Performing the random walk for a sufficient number of steps propagates and accumulates similarity along these paths, integrating content-based, collaborative and location-based similarities. Due to the exponential decay over path length, infinite graph walk probabilities can be approximated by limiting the graph walk to a finite number of steps k (Minkov & Cohen, 2010).

Experimental Setup

This section introduces our experimental data and defines the evaluation methodology.

Data

Collection procedure. We experiment with authentic data collected during visits to the Hecht Museum, a small size museum located on the campus of the University of Haifa (http://mushecht.haifa.ac.il/). The museum presents both archeological and art exhibits, as described in more detail elsewhere (Kuflik et al., 2014).

A location-aware mobile device is provided to the museum visitors. As the visitor tours the museum and is detected at some point of interest, she is prompted to select an exhibit out of the nearby exhibits using a graphical interface. A list of all relevant presentations is then listed on the mobile screen. This procedure is illustrated in Figure 2. As shown in the middle part of the figure, the presentations are introduced by title, typically phrased as a question, for example, "What are the characteristics of Egyptian anthropoids?" Once the visitor selected a title of interest, the respective multimedia presentation is displayed (see the right part of Figure 2). All of the offered presentations are 1 to 2 minutes long. The user can quit browsing a multimedia representation at any time, and can view multiple presentations for a selected exhibit. (See also Kuflik, Stock, et al., 2011; Kuflik et al., 2014).

The mobile device is designed to present and receive user feedback. Colored smiley emoticons may appear next to some of the offered presentations, reflecting high average rating received for those presentations by past visitors. (Emoticons were presented for 72% of the presentations at the time of this research; they are not included in the screenshots in Figure 2.) Having viewed a presentation, the user is requested to provide her own feedback.

Our data set consists of visit logs collected over a period of several months. The logs include the POIs that the visitor passed through, the exhibits selected, and the feedback

FIG. 2. Mobile visitors guide screenshots. [Color figure can be viewed at wileyonlinelibrary.com]

scores (ratings) assigned to the presentations viewed. Presentations that were terminated early by the user and presentations for which no feedback was provided were excluded from the data set.

Data statistics. Overall, more than 300 presentations are available for viewing at the Hecht museum that correspond to 76 exhibits, displayed at 49 positions, across multiple rooms and two floors.

Figure 3(a) details the number of presentations offered per exhibit. As shown, for nearly half of the exhibits, five or more presentations are offered. Figure 3(b) further describes the number of presentations viewed in the discourse of a single visit.[3] As shown, the median number of presentations viewed is 10, where visitors most often viewed between five and nine presentations during their visit. (Log files of visits in which a single presentation has been viewed were excluded from the data set.) This means that a typical visitor only views (and provides feedback) for a small fraction of the multimedia presentations available.

The distribution of collected feedback is highly skewed. About 1% of the presentations were viewed by about a third of the visitors in our data sets; most of these presentations are associated with exhibits located near the museum entrance. On the other hand, approximately 17% of the presentations were viewed by a single visitor, or none. For a given exhibit, there is further bias in favor of certain presentations. It has been previously shown that visitors tend to favor general-themed presentations over aspect-specific ones (Kuflik et al., 2014). In the Hecht museum, general introductory presentations are offered for about a third of the exhibits. In addition, it is known that whenever presented with a list of items, human users tend to focus on the topmost ranked items. During data collection, presentations were arbitrarily ordered by their IDs, that is, their ordering was random but fixed. While some of these biases may be softened with proper experimental design and targeted data collection, we believe that our data set captures authentic visitor behavior, reflecting real-world challenges such as data sparsity and unbalanced data distributions.

Pre-processing. The visit logs that constitute our data set were obtained at several different points in time, and therefore exhibit some variance with respect to rating scales. Some of the feedback is binary (*like, dislike*), but the majority of feedback is given in 3-point or 5-point scale. We converted the different feedbacks to a uniform 5-point scale: binary feedbacks were converted to integer values of {1,5}, and scores on a 3-point scale were represented using the values {1,3,5}.

Figure 3(c) shows the distribution of the processed feedback scores. As shown, the ratings tend to be positive–about 65% of the feedback scores are very high (score of 5), and very few ratings are low. In order to alleviate sparsity, we include in the user profile all of the presentations for which she assigned feedback scores in the range 3–5, weighted by their normalized feedback score.

Graph design. The proposed graph schema directly links similar *presentation* node pairs, as well as similar *theme* node pairs. The presentations are generally short; most of them correspond to a textual description of between 100–200 words. We consider the full textual content of the presentations, including title and transcript of the spoken commentary, in assessing pairwise similarities. We used WEKA (Hall et al., 2009) to compute cosine similarity between the respective TF-IDF weighted term vectors, having stop words removed, content words stemmed and lower-cased, and word weights normalized by document length. *Theme* nodes are represented as the centroid of vectors of the *presentations* associated with each theme. The centroids were computed using Euclidean distance, with inter-theme relatedness evaluated using cosine similarity.

It is generally desired to avoid the modeling of weak associations as graph edges–weak links are uninformative, and can increase the cost involved in computing the PPR measure. We therefore selectively link only those entity pairs for which the computed similarity score exceeds some manually tuned threshold. The threshold values for the

[3]Please note that in this section, and in the rest of the paper, we only refer to viewed presentations for which feedback was provided.

DOI: 10.1002/asi

similar-to and *similar-theme* edge types were set based on the training data to 0.2 and 0.4, respectively. Consequently, *presentation* nodes are linked over *similar-to* edges to 2.9 other *presentation* nodes on average; and, most of the *theme* nodes are connected over the *similar-theme* relation to three to four other *theme* nodes. Similarly, we link each *position* node to its three nearest *positions* in terms of walking distance (and up to six positions in case of a tie). Note that the resulting similarity-based edges are asymmetric, for example, node x may point to node y over a *similar-theme* edge, whereas y may not be point to x.

In summary, the experimental data set is sparse. The resulting graph is small–it consists of several hundreds of nodes (see Table 1) and less than 10K directed edges. The respective user-item matrix includes about 4% populated cells. Only a small number of feedbacks is available for individual visitors. Recommendation at the museum must therefore address constant “cold start” conditions.

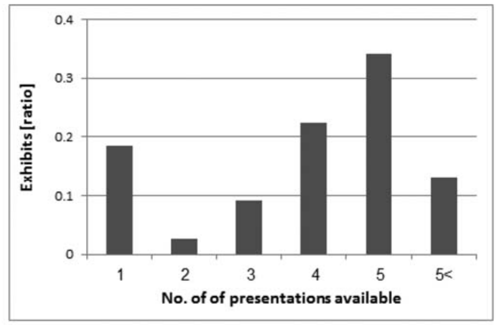
Experimental Design

We perform a set of *prediction* experiments using the visit logs collected at the Hecht museum. Given the set of ratings provided by user u, we target in each experiment one of the rated presentations, $i^* \in I_u$, having the remaining presentations $\{I_u - i^*\}$ serve as the user’s profile V_u. In case that user u positively appreciated item i^*, we expect to find it among the top items of the generated ranked list of *presentation* nodes. This experimental setting is imperfect, mainly, other highly ranked presentations may be of interest to the user, for which we do not possess relevancy judgements. Yet, such a setting is often applied for the purpose of comparing the performance of multiple ranking methods; importantly, it is unbiased as all systems are assessed under the same conditions (Baluja et al., 2008).

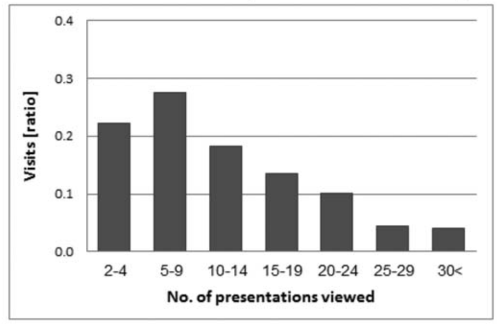
We dedicate a *held-out* portion of the example set for tuning of the graph edge weights. We selected all of the queries generated per 10% of the users (32 complete visit logs) for this purpose. The examples derived from the remaining 287 (90%) visit logs serve for evaluation using the leave-one-out procedure.

Two modes of recommendation are considered in our experiments:

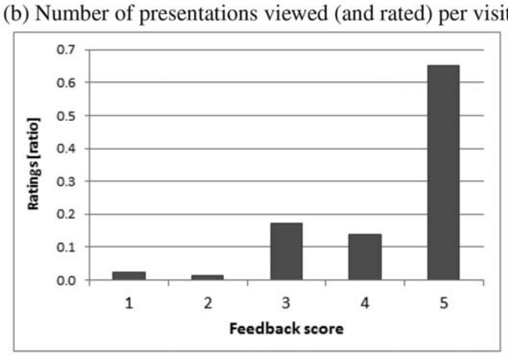
- *General Recommendation*: We are interested in assisting the visitor in choosing *multimedia presentations* of interest while touring the museum. All presentations (except those already viewed by the user) are considered as candidate items in this setting. Ideally, the museum’s physical layout should be taken into account in this mode. Physical distances are modeled in the graph, and affect the generated rankings.
- *Per-Exhibit Recommendation*: In this setting, we consider only *presentations* that are associated with an attended exhibit as candidate items for recommendation. (Technically, per-exhibit differs from the general recommendation mode in the final ranking step, having presentations that are not associated with the current exhibit filtered out.) As indicated in Figure 2(a), about fifth of the exhibits are associated with a single presentation. We omit queries pertaining to these exhibits from the evaluation data set in the per-exhibit evaluation mode, as single-choice recommendation is trivial.



(a) Distribution of the number of presentations available per exhibit



(b) Number of presentations viewed (and rated) per visit



(c) Distribution of the feedback scores

FIG. 3. Data set statistics. (a) Distribution of the number of presentations available per exhibit. (b) Number of presentations viewed (and rated) per visit. (c) Distribution of the feedback scores.

Evaluation Measures

All of the recommendation methods considered in this paper generate scores for candidate items, which are then processed into a ranked list to be presented on the mobile screen. We accordingly assess performance using measures used for the evaluation of ranked lists. Notably, each evaluated query has a single known correct answer in our experiments.

Recall-at-k. This measure estimates the probability of retrieving the correct answer within the top k ranks (Minkov & Cohen, 2010). For example, recall@3 = 0.7 means that

DOI: 10.1002/asi

for 70% of the queries, the correct answer appears among the top three ranks of the retrieved lists.[4] The screen of handheld devices is limited in size, and it expected that the highest-ranking items will receive most of the user's attention. For this reason, we report *recall@k*$=[1-5]$ in the general recommendation setting, and $k=[1-3]$ in the per-exhibit mode. In the latter case, the number of candidate items is small, so that the added value of recommendation is in pointing out the very few items that are of highest interest to the user.

Mean reciprocal rank (MRR). The reciprocal rank of a response to a single query is defined as the multiplicative inverse of the rank of the correct answer: $\frac{1}{k_i}$, and the mean reciprocal rank is the average of the reciprocal ranks for all of the test queries.[5]

Ratio of failed tests (RFT). Occasionally, a recommendation method may fail to predict scores for the target presentation. In such cases, we assign this presentation a de-facto score of zero, appending it to the bottom of the output ranked list. This means that impaired coverage affects recall-at-N and MRR performances. For completeness, we also report the *ratio of failed tests*, in which the test presentation failed to receive a score.

Experiments

This section includes a description and implementation details of the evaluated recommendation methods.

Graph-based recommendation. In order to assess the utility of combining historical ratings with background knowledge, we experiment with several graph variants illustrated in Figure 4:

1. *The Museum graph (G:M).* This graph describes background knowledge about the museum, modeling thematic and physical similarities. Specifically, it includes *presentations*, *themes* and *positions* nodes, connected over the edge types *similar-to*, *similar-theme*, *nearby*, *has-theme*, *located-in*, and the respective inverse edges.
2. *Visitors graph (G:V).* This bi-partite graph variant represents ratings history, including *presentations* and *users* as entities, linked over *viewed* and *viewed*$^{-1}$ edges.
3. *Unified graph (G:U).* The *museum* and *visitors* graphs contain complementary information. The *unified* graph forms the union of the two graphs, as demonstrated in Figure 4(c).
4. *Combined graphs (G:MV).* This approach integrates the scores produced by the *visitors* and *museum* graphs using a linear combination of the scores:

$$\hat{r}_{ui}(v_u, G:MV)=(1-\beta)\cdot\hat{r}_{ui}(v_u, G:M)+\beta\cdot\hat{r}_{ui}(v_u, G:V) \quad (3)$$

The weighting coefficient β was tuned empirically in our experiments using grid search over the range [0.1,0.9] with step size 0.1, optimizing performance on the held-out examples. The graph edge weight parameters Θ were similarly tuned using the held-out examples (see Impact of Graph Tuning). We set the damping factor of the random walk process (Equation [1]) to $\alpha=.85$ following previous work (Minkov & Cohen, 2010).[6]

$$userSim(u,v)=\frac{\sum_{i\in I_{uv}}(r_{ui}-\bar{r}_u)(r_{vi}-\bar{r}_v)}{\sqrt{\sum_{i\in I_{uv}}(r_{ui}-\bar{r}_u)^2}\sqrt{\sum_{i\in I_{uv}}(r_{vi}-\bar{r}_v)^2}} \quad (4)$$

$$\hat{r}_{ui}=\bar{r}_u+\frac{\sum_{v\in N_i(u)} userSim(u,v)\times(r_{vi}-\bar{r}_v)}{\sum_{v\in N_i(u)}|userSim(u,v)|} \quad (5)$$

$$itemSim(i,j)=\frac{\sum_{v\in U_{ij}}(r_{vi}-\bar{r}_v)(r_{vj}-\bar{r}_v)}{\sqrt{\sum_{v\in U_{ij}}(r_{vi}-\bar{r}_v)^2}\sqrt{\sum_{v\in U_{ij}}(r_{vj}-\bar{r}_v)^2}} \quad (6)$$

$$\hat{r}_{ui}=\bar{r}_i+\frac{\sum_{j\in N_u(i)} itemSim(i,j)\times(r_{uj}-\bar{r}_j)}{\sum_{j\in N_u(i)}|itemSim(i,j)|} \quad (7)$$

We approximate PPR scores using finite random walk repeated for six iterations. As discussed earlier, the impact of additional steps on the generated rankings is negligible.

Content-based recommendation (CB). We experiment with a version of the Rocchio algorithm (Lops, de Gemmis, & Semeraro, 2011; Salton, 1971). This method computes a "prototype" vector for user u by averaging vectors of documents known to be of interest to u, and subtracting away the weighted fraction of vectors of uninteresting documents.

We represent each candidate *presentation* as a vector of TF-IDF weighted terms, describing its textual contents. Since in this study, only a small fraction of ratings are below the median score, we only model positive feedback. We also weight the individual presentation vectors included in the user's profile by the respective feedback scores so that highly liked presentations contribute more to the profile. Item relevancy is estimated using cosine similarity in this vector space.

User-based k-nearest neighbor (CF:U-kNN). We follow closely on Desrosiers and Karypis (2011) in our implementation of this well-known CF method. This method generates a rating prediction $\hat{r}_{ui}$ based on the ratings for item i by a set of k users most similar to the target user u. The similarity between users u and v, $userSim(u, v)$, is computed as the Pearson's correlation bewteen their historical ratings, as defined in Equation (4), where I_{uv} denotes the set of items co-rated by users u and v, and $\bar{r}_u$ and $\bar{r}_v$ denote the average

[4]We consider the *effective* rank of the target item, which may be a real number, for example, if the 5th and 6th ranked items are assigned identical scores, the rank of both items is 5.5, computed as (5 + 6)/2. In our experiments, item scores are often on par using random recommendation, and to a lesser extent, using CF kNN and some graphs variants. The evaluation is strict, having ranks rounded up in evaluating recall; for example, rank 5.5 contributes to recall@6 and downwards.

[5]In computing the MRR measure, half ranks were maintained.

[6]The produced PPR rankings are generally insensitive to α value, for example, (Minkov & Cohen, 2010).

DOI: 10.1002/asi

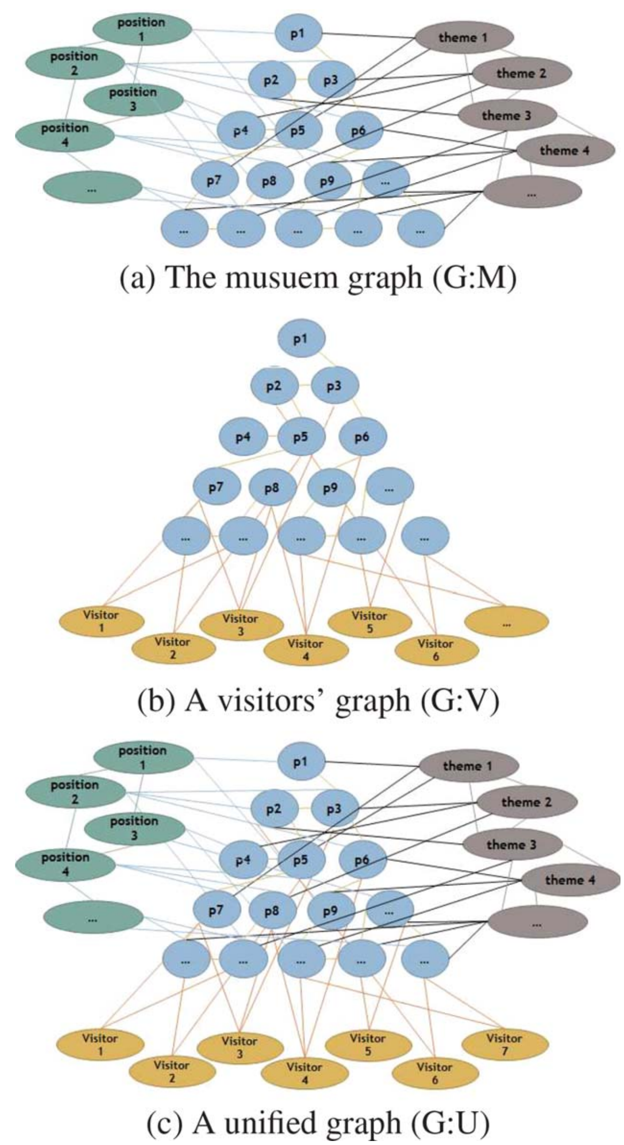


FIG. 4. Illustration of the structure of several graph variants. (a) The museum graph (G:M). (b) A visitors' graph (G:V). (c) A unified graph (G:U). [Color figure can be viewed at wileyonlinelibrary.com]

ratings of users u and v, respectively. This formula applies mean-normalization of rating scores per user, so as to account for variance in rating scales across individuals. Once the set of neighbors $N_i(u)$ is identified, the predicted rating is computed according to Equation (5), weighting the contribution of each neighbor by its similarity to u.

Item-based kNN (CF:I-kNN). This method evaluates the recommendation score by analyzing the ratings of similar items (Desrosiers & Karypis, 2011). As defined in Equation (6), the similarity between items I and j, *itemSim*(i, j), is determined by the extent to which other users assigned similar ratings to the two items, where U_{ij} denotes the set of users who have rated both items i and j. The predicted rating $\hat{r}_{ui}$ is computed as a weighted average of the ratings assigned by user u to $N_u(i)$, a set of up to k items that are found to be most similar to item i using Pearson correlation, as defined in Equation (7).

In our experiments with both kNN methods, we tune the neighborhood size k. We apply the common practice of *negative filtering*, discarding neighbors with a negative correlation score (Herlocker, Konstan, Borchers, & Riedl, 1999). Several additional threshold types were evaluated in order to identify a high-quality set of neighbors: discarding neighbors for which the correlation score is below a minimum similarity value, or below a minimal number of common ratings. The various combinations of threshold values and types, as well as the neighborhood size, were evaluated exhaustively. We report results using the best parameters per setting.

Matrix factorization (CF:MF). Given a user-item ratings matrix $M=(r_{ui})$, matrix factorization maps the users and items into a joint latent factor space of dimensionality k, having every item i and user u represented as vectors $q_i, p_u \in \mathbb{R}^k$. The rating $\hat{r}_{ui}$ is approximated by the dot product of the item and user vectors, capturing the user's overall interest in the item's characteristics. We experiment with a state-of-the-art MF formulation outlined by Koren, Bell, and Volinsky (2009), as implemented in the GraphLab software package[7] (Low et al., 2012) using alternating least squares optimization for minimizing the cost function. We tuned the algorithm parameters using grid search, setting the number of latent factors to 110 in the general recommendation mode, and to 100 in the per-exhibit mode. The regularization coefficient was set to $\lambda = 1$, and the number of iterations to convergence was set to a maximum of 1,000. Stochastic optimization is prone to converge to a local optimum. We therefore report average results of five runs with randomized initialization.

Hybrid recommendation (Hyb). Hybrid systems combine multiple techniques, ideally compensating for the weakness of the individual methods (Berkovsky, Heckmann, & Kuflik, 2009). We experiment with a combination of two methods: content-based (CB) and item-based kNN (CF:I-kNN).[8] Concretely, we first re-scale the scores produced by the two methods, and then compute a weighted average of the normalized item scores. The weighting coefficient was tuned using grid search over the range [0.1,0.9] with step of 0.1. The selected weights values were (0.3, 0.7) in the general recommendation scenario, and (0.1, 0.9) in the per-exhibit scenario, assigning in both cases a higher weight to the item-based kNN method.

Random (B:R). This naive nonpersonalized baseline selects one of the candidate *presentations* uniformly at random. Comparing against this baseline demonstrates the contribution of informed recommendation systems over mere chance.

[7]http://graphlab.org/

[8]As discussed later, item-based kNN performed best among the CF methods in our experiments.

DOI: 10.1002/asi

TABLE 2. MRR and RFT performance of the various recommendation methods.

	General		Per-exhibit	
	MRR	RFT	MRR	RFT
G:M	.110*	.002	.627*	.001
G:V	.143	.009	.788	.001
G:U	.151	.002	.788	.001
G:MV	**.152**	.002	.789	.001
CB	.048*	.002	.585*	.001
CF:U-kNN	.028*	.508	.600*	.399
CF:I-kNN	.084*	.874	.648*	.293
CF:MF	.065*	.000	.658*	.000
Hyb	.104*	.002	.659*	.001
B:PR	.018*	.117	-	-
B:P	.140	.000	**.801**	.000
B:R	.007*	.000	.506*	.000

Note: An asterisk denotes statistically significant difference in MRR compared with G:MV, the best performing graph-based method (p-val <0.005).

Proximity (B:PR). This is a stricter version of random recommendation, which limits the set of candidate presentations in terms of distance from the user's whereabouts. Given V_u, we obtain the list of *presentations* associated with already visited *positions*; that is, those that can be reached from any *presentation* $x \in V_u$ over the path $x \xrightarrow{located-in} y \xrightarrow{located-in^{-1}} z$. We also include presentations relevant for nearby positions, reached over the path: $x \xrightarrow{located-in} y \xrightarrow{nearby} q \xrightarrow{located-in^{-1}} z$.

Popularity (B:P). The *popularity* baseline ranks the candidate *presentations* according to their popularity score, computed as the number of users who viewed each presentation. This method is nonpersonalized yet informed and often hard-to-beat (Lucchese, Perego, Silvestri, Vahabi, & Venturini, 2012). A challenge of any personalized recommendation is to outperform this one-fits-all approach.

Main Results

This section reports our results using the different methods. We remind the reader that the parameters of the graph-based methods were tuned using held-out examples; in contrast, all other methods have been optimized directly on the *test* data. Despite the comparison being strict in this fashion, graph-based recommendation is shown to give preferable results.

General Recommendation

Table 2 includes MRR and RFT results for the general recommendation scenario, and Figure 5(a) shows the respective recall-at-rank performances. As shown, the graph variants G:U and G:MV, which model ratings history together with background knowledge, yield the best overall performance with respect to all measures. MRR results using G:U and G:MV are 0.151 and 0.152, respectively. In terms of recall, G:MV gives slightly better performance, yielding recall of 0.063 at the topmost rank, and 0.220 recall at rank 5.

It is informative to contrast these results with the prediction quality of the non-personalized baselines. As one might expect, the popularity-based method (B:P) shows strong performance, yielding MRR score of 0.140. The more naive baselines—*random* and *proximity* based recommendation—result in very low MRR performance, as well as negligible recall at the top levels. The evident strength of popularity-based recommendation correlates with our data set statistics. As discussed earlier, the data set is characterized with a large number of candidates, for which the distribution of available feedbacks is highly skewed.

The evaluated CF methods show relatively weak performance in our experiments, all failing to beat the popularity-based baseline. Item-based kNN recommendation achieved MRR of 0.084, MF—0.065, and user-based kNN—a low 0.028. The weakness of these methods can be attributed in part to data sparsity. As shown in Table 2, RFT is very high using the kNN CF methods. We found that in many cases, there were no relevant "neighbors" identified due to rating sparsity. In fact, many visitors in our data set (30%) assigned the same feedback score to all of the presentations that they have viewed, and therefore did not contribute to the recommendation process (see Equations [4] and [6]). MF is generally more robust to sparsity issues, but deliver mediocre results here. It is likely that its performance would improve should a larger set of ratings be provided.

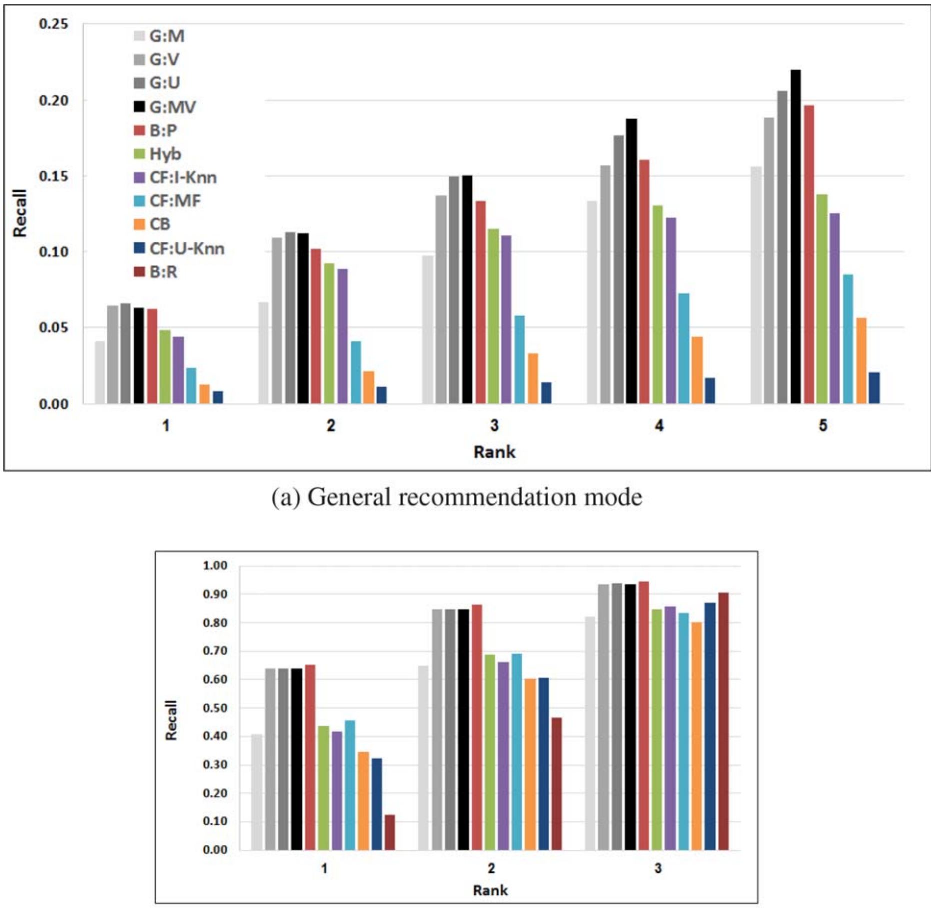
Interestingly, recommendation using the graph variant G:V, which similarly to the CF methods, models ratings history only, delivers strong performance and outperforms the popularity baseline. This suggests that the transitive graph relatedness measure is advantageous in conditions of sparsity. Another factor that may positively affect the graph-based recommendation is that random walk measures like PPR exhibit bias towards highly connected nodes (Tong & Faloutsos, 2006), thus implicitly modeling item popularity information.

Finally, CB recommendation performance falls behind its graph counterpart, G:M, which models background information (0.048 vs 0.110 in MRR). The hybrid method (Hyb) improves upon each of its component systems, but still falls short of graph-based recommendation. Again, we conjecture that the graph-based techniques are preferable in conditions of sparse data. Also, the graph method integrates physical proximity aspects, which are missing from either the CF or CB approaches.

Per-exhibit recommendation

Since the set of candidate items for recommendation in the per-exhibit setting is highly limited (Figure 3[a]), random recommendation achieves high recall levels at the top

DOI: 10.1002/asi



(a) General recommendation mode

(b) Per-exhibit recommendation mode

FIG. 5. Recall-at-rank performance of the evaluated methods: graph-based recommendation using the museum graph (G:M), the visitors ratings graph (G:V), hybrid unified graph (G:U) and integrated variant scores (G:MV); item-based and user-based collaborative filtering (CF:I-KNN, CF:U-KNN), matrix factorization (CF:MF), content-based recommendation (CB), a hybrid combination of CB and CF:I-KNN (HYB), popularity-based baseline (B:P) and random recommendation baseline (B:R). (a) General recommendation mode. (b) Per-exhibit recommendation mode. [Color figure can be viewed at wileyonlinelibrary.com]

three ranks (0.12, 0.47 and 0.91) (Figure 5[b]), and a high MRR score of 0.506 (Table 2). The *popularity* baseline achieves the strongest performance overall in this setting–yielding MRR of.801. We find that this result correlates with user bias towards presentations displayed at the top of the mobile screen during data collection. As previously shown, the mobile guide screen accommodates about four presentation titles (Figure 2), and for about half of the exhibits five or more presentations are offered (Figure 3[a]). As users tend to view (and like) the first presentation on the list, and generally focus on the top listed items, simply ranking the presentations by their popularity is somewhat trivially successful.[9] We believe that given a larger number of candidate presentations per exhibit, and interaction with users that encourages them to follow their own interests, the effect of personalization will be more noticed.

[9]As noted by Kuflik et al. (2014), the second presentation viewed matched the user profile, while the first one was a popular, generic one.

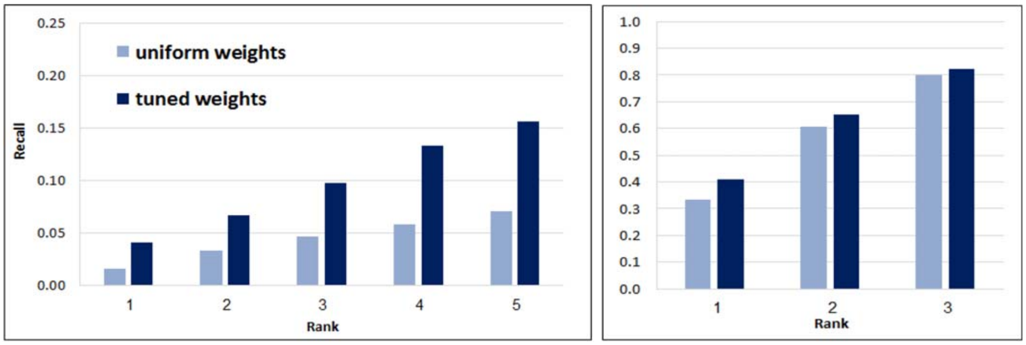
TABLE 3. Evaluation of graph variants.

	General		Per-exhibit	
	MRR^U	MRR	MRR^U	MRR
G:M	.058	.110$^{(+89.6\%)}$	.582	.627$^{(+7.7\%)}$
G:V	-	.143	-	.788
G:U	.150	.151$^{(+4.9\%)}$	.754	.788$^{(+4.5\%)}$
G:MV	.150	**.152**$^{(+1.3\%)}$	.735	**.789**$^{(+7.3\%)}$

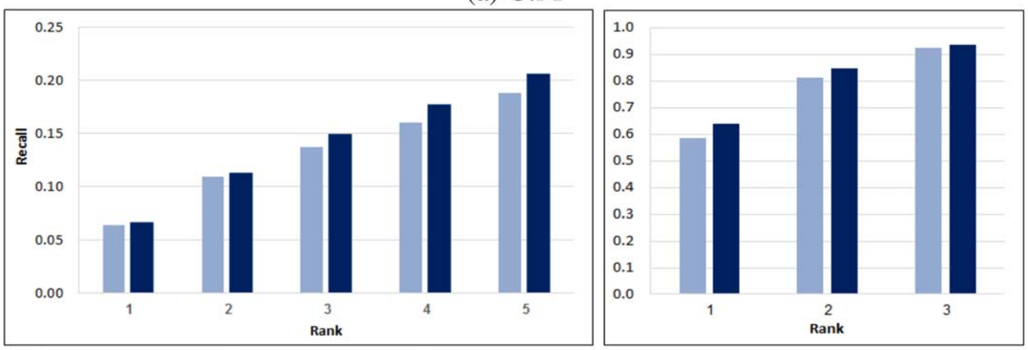
We now turn to discuss the personalized recommendation methods, which ideally, would increase visitor satisfaction beyond the one-fits-all rankings.

Consistently with the previous findings, the best-performing approaches are the integrative graph-based methods, G:U and G:MV, yielding MRR of 0.788 and 0.789, respectively. The visitors graph G:V gives roughly equal performance (MRR of 0.788). The contribution of background knowledge seems negligible in this case. Indeed, physical proximity is irrelevant in the per-exhibit setting.

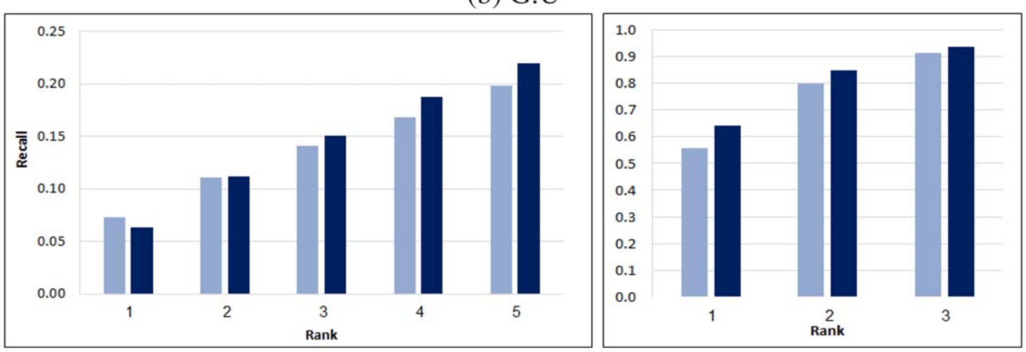
DOI: 10.1002/asi



(a) G:M



(b) G:U



(c) G:MV

FIG. 6. Impact of tuning parameters in affected graph variants, in terms of recall at the top ranks, in the general recommendation (left) and the per-exhibit (right) recommendation modes. (a) G:M. (b) G:U. (c) G:MV. [Color figure can be viewed at wileyonlinelibrary.com]

Again, CF methods yield inferior results. MF yields MRR of 0.658 compared with 0.788 using G:V, the graph variant which models equivalent information. Likewise, the results using CB are inferior to the counterpart G:M. The *hybrid* method gives a comparable result to CF:MF (0.659), and lower performance compared with any graph method that considers the historical ratings.

Impact of Graph Tuning

Representing data as a graph involves some design choices (see Data). Here we discuss the impact of the edge weight parameters Θ on recommendation performance.

We empirically tuned Θ using grid search, optimizing recommendation performance on the held-out examples, considering all combinations of edge weight values in the range [0,1] with step 0.1.[10] The *viewed* edges form an exception—these edge weights were set according to the feedback scores assigned by the visitor to each presentation. In order to avoid dominance of the *viewed* edges over other edge types, the rating scores were transformed into decimal fractions (0.1–0.5). We assigned these bi-directional edge types (e.g., *viewed* and $viewed^{-1}$) identical weights, so as to reduce the cost of edge weight tuning.

Table 3 shows MRR performance of the different graph variants using uniform edge weight parameters (MRR^U) versus the final tuned weights. Figure 6 demonstrates recall-at-rank-k results, prior to and post parameter tuning. Because the graph G:V includes a single relation type, it was not affected by weight tuning. Equal coefficients

[10]Parameters were tuned separately for the general recommendation and the per exhibit scenarios.

DOI: 10.1002/asi

(α=.5) were used in the pretuned version of the G:MV variant.

As shown, edge weight tuning was highly effective for the *museum* graph, increasing its MRR result by roughly 90% in the general recommendation setting, and by about 8% in the perexhibit setting. MRR results using G:U and G:MV were improved by means of weight tuning by 1.3–7.3%, reaching the best result overall. Similar trends are observed with respect to recall@k performance.

We found that low weights were assigned to the *similar-to* and *similar-theme* edge types, possibly reflecting ineffective modeling of content-based similarity due to text sparsity. In contrast, the structural *has-theme* edges were assigned high weights. The weight of the *nearby* edge type was set to a high value in the general setting; unsurprisingly, in the per-exhibit setting in which all candidate items are associated with a single exhibit, its weight was low. Overall, this demonstrates the flexibility of the graph approach: the very same graph is effectively optimized per recommendation task by tuning its parameters.

Conclusions

We described a graph-based framework for generating personalized recommendation to museum visitors. As visits at museums are often a one-time experience and are limited in time, recommendation must be performed in constant "cold start" conditions. The lack of sufficient rating history may be compensated by modeling of useful background knowledge. Another aspect that must be modeled at the museum is its layout—adjacent museum exhibits are typically semantically related, and items associated with exhibits nearby are likely to be visited by the user.

In an extensive set of experiments, we have showed that graph-based recommendation using the PPR measure significantly outperforms a set of classical collaborative and content-based recommendation methods. There are several main reasons for the superiority of the graph-based approach. First, the graph approach naturally models and integrates collaborative ratings together with physical layout and content aspects. Further, the structured random walk similarity measure is transitive, thus alleviating data sparsity. Moreover, the graph measure can be tuned per the specific recommendation task; we have shown that controlling the probability flow in the graph by means of edge weight tuning improves performance substantially. Finally, the random walk process favors highly connected nodes, thus implicitly modelling a bias towards popular items.

In the future, we would like to enrich the limited textual content modeled in this work using the web (Grieser, Baldwin, Bohnert, & Sonenberg, 2011) or linguistic resources (Bohnert & Zukerman, 2014). In addition to recommending *multimedia presentations*, it is possible to advise the visitor on the next *position* to visit using the described framework, and then organize these predictions into path recommendations. Finally, in future work we hope to extend our reach to additional museums and to replicate the experiments in real-time settings, providing recommendations and gathering feedback as the visitors are touring in the museum.

Acknowledgments

This work was partially supported by ISF grant No. 1108/2014.

References

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering (TKDE), 17, 734–749.

Adomavicius, G., & Tuzhilin, A. (2011). Context-aware recommender system. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), Recommender systems handbook. pp. 217–253. Springer.

Ardissono, L., Kuflik, T., & Petrelli, D. (2012). Personalization in cultural heritage: The road travelled and the one ahead. User Modeling and User-Adapted Interaction, 22, 1–27.

Bagci, H., & Karagoz, P. (2016). Context-aware location recommendation by using a random walk-based approach. Knowledge and Information Systems, 47, 241–260.

Baluja, S., Seth, R., Sivakumar, D., Jing, Y., Yagnik, J., Kumar, S., . . . Aly, M. (2008). Video suggestion and discovery for Youtube: Taking random walks through the view graph. In *Proceedings of the 17th International Conference on World Wide Web (WWW)*.

Bartolini, I., Moscato, V., Pensa, R.G., Penta, A., Picariello, A., Sansone, C., & Sapino, M.L. (2014). Recommending multimedia visiting paths in cultural heritage applications. Multimedia Tools and Applications. pp. 1–30.

Berkovsky, S., Heckmann, D., & Kuflik, T. (2009). Addressing challenges of ubiquitous user modeling: Between mediation and semantic integration. In Advances in Ubiquitous User Modelling, volume 5830, of LNCS, pages 1–19. Springer.

Biran, A., Poria, Y., & Oren, G. (2011). Sought experiences at (dark) heritage sites. Annals of Tourism Research, 38, 820–841.

Bohnert, F., & Zukerman, I. (2014). Personalised viewing-time prediction in museums. User Modeling and User-Adapted Interaction, 24, 263–314.

Bohnert, F., Zukerman, I., & Laures, J. (2012). Geckommender: Personalised theme and tour recommendations for museums. In International Conference on User Modeling, Adaptation, and Personalization, pp. 26–37. Springer Berlin Heidelberg, 2012.

Bu, J., Tan, S., Chen, C., Wang, C., Wu, H., Zhang, L., & He, X. (2010). Music recommendation by unified hypergraph: Combining social media information and music content. In *Proceedings of the International ACM Conference on multimedia (MM)*.

Davey, G. (2005). What is museum fatigue? Visitor Studies Today, 8, 17–21.

Desrosiers, C., & Karypis, G. (2011). A comprehensive survey of neighborhoodbased recommendation methods. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), In Recommender systems handbook (pp. 107–144). Springer US.

Dim, E., & Kuflik, T. (2015). Automatic detection of social behavior of museum visitor pairs. ACM Transactions on Interactive Intelligent Systems (TIIS), 4.

Fouss, F., Pirotte, A., Renders, J.M., & Saeren, M. (2007). Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. IEEE Transactions on Kknowledge and Data Engineering (TKDE), 19, 355–369.

Gori, M., & Pucci, A. (2007). Itemrank: A random-walk based scoring algorithm for recommender engines. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI)*.

Grieser, K., Baldwin, T., & Bird, S. (2007). Dynamic path prediction and recommendation in a museum environment. In *Proceedings of the Workshop on Language for Cultural Hetitage Data*.

Grieser, K., Baldwin, T., Bohnert, F., & Sonenberg, L. (2011). Using ontological and document similarity to estimate museum exhibit relatedness. ACM Journal on Computing and Cultural Heritage, 3,

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.H. (2009). The WEKA data mining software: an update. SIGKDD Exploratory Newsletter, 11, 10–18.

Haveliwala, T.H. (2002). Topic-sensitive PageRank. In *Proceedings of the International World Wide Web Conference (WWW)*.

Herlocker, J.L., Konstan, J.A., Borchers, A., & Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

Jeh, G., & Widom, J. (2003). Scaling personalized web search. In *Proceedings of the International World Wide Web Conference (WWW)*.

Karatzoglou, A., Amatriain, X., Baltrunas, L., & Oliver, N. (2010). Multiverse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the Fourth ACM Conference on Recommender Systems (RecSys)*.

Katz, S., Kahanov, Y., Kashtan, N., Kuflik, T., Graziola, I., Rocchi, C., . . . Zancanaro, M. (2006). Preparing personalized multimedia presentation for a mobile museum visitors' guide: A methodological approach. In *Proceedings of Museums and the Web*.

Konstas, I., Stathopoulos, V., & Jose, J.M. (2009). On social networks and collaborative recommendation. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. IEEE Computer, 42, 30–37.

Kuflik, T., Kay, J., & Kummerfeld, B. (2012). Challenges and solutions of ubiquitous user modeling. In Ubiquitous display environments, pp. 7–30. Springer Berlin Heidelberg, 2012.

Kuflik, T., Stock, O., Zancanaro, M., Gorfinkel, A., Jbara, S., Kats, S., . . . Kashtan, N. (2011). A visitor's guide in an active museum: Presentations, communications, and reflection. Journal on Computing and Cultural Heritage (JOCCH), 3.

Kuflik, T., Wecker, A.J., Lanir, J., & Stock, O. (2014). An integrative framework for extending the boundaries of the museum visit experience: Linking the pre, during and post visit phases. Information Technology and Tourism, 15, 17–47.

Lao, N., Minkov, E., & Cohen, W.W. (2015). Learning relational features with backward random walks. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Liu, X., & Aberer, K. (2013). Soco: A social network aided context-aware recommender system. In *Proceedings of the International World Wide Web Conference (WWW)*.

Lops, P., de Gemmis, M., & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), Recommender systems handbook. pp. 73–105. New York: Springer.

Low, Y., Gonzalez, J., Kyrola, A., Bickson, D., Guestrin, C., & Hellerstein, J.M. (2012). Distributed graphlab: A framework for machine learning and data mining in the cloud. Proceedings of the VLDB Endowment, 5.

Lucchese, C., Perego, R., Silvestri, F., Vahabi, H., & Venturini, R. (2012). How random walks can help tourism. In European Conference on Information Retrieval (pp. 195–206). Springer Berlin.

Minkov, E., & Cohen, W.W. (2010). Improving graph-walk based similarity with reranking: Case studies for personal information management. ACM Transactions on Information Systems (TOIS), 29.

Noulas, A., Scellato, S., Lathia, N., & Mascolo, C. (2012). A random walk around the city: New venue recommendation in location-based social networks. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trusts*.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The Pagerank Citation Ranking: Bringing Order to the Web. In Technical report, Computer Science Department, Stanford University.

Pham, T.N., Li, X., Cong, G., & Zhang, Z. (2015). A general graph-based model for recommendation in event-based social networks. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*.

Richardson, M., & Domingos, P. (2002). The intelligent surfer: Probabilistic combination of link and content information in PageRank. In the Conference on Neural Information Processing Systems (NIPS).

Salton, G. (Ed.). (1971). Relevance feeback in information retrieval. Englewood Cliffs, NJ: Prentice Hall.

Shang, S., Kulkarni, S.R., Cuff, P.W., & Hui, P. (2012). A random walk based model incorporating social information for recommendations. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*.

Stock, O., Zancanaro, M., Busetta, P., Callaway, C., Krüger, A., Kruppa, M., Kuflik, T, Not, E, & Rocchi, C. (2007). Adaptive, intelligent presentation of information for the museum visitor in peach. User Modeling and User-Adapted Interaction, 17, 257–304.

Tiroshi, A., Berkovsky, S., Kaafar, M.A., Vallet, D., Chen, T., & Kuflik, T. (2014). Improving business rating predictions using graph based features. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*.

Tong, H., & Faloutsos, C. (2006). Center-piece subgraphs: Problem definition and fast solutions. In *Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Tong, H., Faloutsos, C., & Pan, J.Y. (2006). Fast random walk with restart and its applications. In *Proceedings of the International Conference on Data Mining (ICDM)*.

Wang, H., Terrovitis, M., & Mamoulis, N. (2013). Location recommendation in location-based social networks using user check-in data. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL)*.

Yao, W., He, J., Huang, G., Cao, J., & Zhang, Y. (2015). A graph-based model for context-aware recommendation using implicit feedback data. World Wide Web, 18, 1351–1371.

DOI: 10.1002/asi