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

Einat Minkov
Information Systems Dep.
University of Haifa
Haifa, Israel, 31905
einatm@is.haifa.ac.il

Keren Kahanov
Computer Science Dep.
University of Haifa
Haifa, Israel, 31905
kahanovk@gmail.com

Tsvi Kuflik
Information Systems Dep.
University of Haifa
Haifa, Israel, 31905
tsvikak@is.haifa.ac.il

Abstract

Visitors to museums and other cultural heritage sites encounter a wealth of items in a variety of subject areas, but can explore only a minority of these items. Recommender Systems may help visitors 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, as items are located in a physical space and they may have semantic links between them. We propose a personalized graph-based recommender framework, representing diverse multi-source information in a relational graph. A random walk measure is applied to rank items of interest by their relevancy to a visitor profile, integrating the various dimensions. We report the results of extensive 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 clear superiority of the graph-based approach.

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

1 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. In many museums, there exist numerous exhibits, which are associated with a broad range of topics and are physically spread over large spaces, making it impossible to view all exhibits in one visit (Davey, 2005). Visitors may therefore need assistance in getting the best experience from their visit. Obviously, visitors differ in their preferences, knowledge and expectations, as they come to the museum with their individual ‘identity’ (Falk, 2009). CH recommender systems aim at generating personalized recommendation that fit the visitors specific preferences and needs. Such personalized services can be implemented using dedicated mobile applications (Ardissono, Kuflik, & Petrelli, 2012).

Mobile devices are typically available at CH sites nowadays, 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–feedback on viewed items may be collected explicitly, or in a non-intrusive manner (Stock et al., 2007). For example, it is possible to track the users’ behavior over time, across locations and interaction contexts by analyzing signals transmitted by the visitor’s mobile device (Kuflik, Kay, & Kummerfeld, 2012; Dim & Kuflik, 2014).

Nevertheless, making personalized recommendations at the museum is a challenging problem in several respects. Crucially, the collected feedback information is sparse: every user gets to view and provide feedback for a small set of items out of the plethora of items available, where in the majority of cases, the visitor is introduced to the museum for the first time (Biran, Poria, & Oren, 2011; Snijders, 2014). 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; ideally, it would prioritize exhibits residing in high proximity to those previously visited by the user. It is therefore desired to model *background knowledge* about the museum environment. In addition to physical layout information, relevant background knowledge may map the museum’s environment and items into a semantic space; for example, exhibits are often associated with specific themes. The modeling of semantic aspects is especially important considering the sparsity of historical ratings. In order to integrate visitors feedbacks with physical and semantic contexts, the recommender system must be able to effectively consolidate such heterogenous information.

In this work, we describe an adaptive system designed to present individual museum visitors with personalized recommendations on their mobile device. Concretely, we target the recommendation of *multimedia presentations*, which are available for viewing on the mobile device, providing complementary material to the museum's exhibits. Assuming that some feedback is available for presentations already viewed, our goal is to recommend the visitor on additional presentation items of interest.

We outline and evaluate a graph-based recommendation approach that handles the above-mentioned 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* items, as well as physical *positions* and semantic *themes*. The graph edges denote structured relations, e.g., *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 textual descriptions. In this fashion, past visits history and information about the museum's physical and thematic environment are encoded in a joint graph.

The graph-based recommendation process involves inference of multi-facet node relevancy with respect to a *query*, defined as a distribution of interest over the graph nodes. We apply the Personalized Page Rank (PPR) algorithm (Haveliwala, 2002; Tong, Faloutsos, & Pan, 2006) to rank *multimedia presentations* by their relatedness to a user profile, corresponding to the set items already viewed and liked by the user. PPR applies a random walk procedure, which captures transitive associations between entities, thus assessing inter-node relatedness from a global perspective. Consequently, graph-based recommendation alleviates the sparsity problem.

This paper 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 visitors guide system at the museum, data has been collected in the form of visit logs for research purposes (Kuflik, Wecker, Lanir, & Stock, 2014). Given user feedback on viewed *multimedia presentations*, our goal is to rank the remaining presentation items according to the user's tastes. We report a set of comparative experiments, showing that the graph-based recommendation approach significantly outperforms popular content based and collaborative filtering recommendation approaches, including a state-of-the-art matrix factorization method.

There are several main contributions of this work:

- We show that the graph-based framework delivers accurate recommendations in the challenging cultural heritage domain. Compared with alternative methods, this approach models historical ratings jointly with diverse background knowledge, including contextual physical proximity and semantic aspects. The proposed approach can be readily applied to other problems with similar characteristics.

- There exist relatively few works that employed graph based similarity in general, and the Personalized PageRank measure in particular, for recommendation purposes. We report the results of a comprehensive set of comparative experiments, demonstrating the potential of graph-based approach and its superiority over alternative, popular and state-of-the-art, methods in a contextual recommendation setting.
- We empirically evaluate and discuss in detail issues related to graph design, considering several plausible graph variants, as well as evaluate the impact of tuning parametric edge weights on recommendation performance.

The remainder of the paper is organized as follows. Section 2 provides necessary background, and is followed by a review of related research in Section 3. The graph-based recommendation framework and the proposed graph schema are described in Section 4. Section 5 describes our experimental data and defines the experimental setup. In Section 6, we outline the various recommendation methods evaluated in this work. Our main set of results is presented in Section 7. Section 8 further discusses issues related to graph design, including the impact of edge weight tuning. Section 9 concludes this paper, and discusses directions for future research.

2 Background

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}. Ratings may be given in multiple forms, such as numerical scores, e.g. $[1, 5]$, or binary indications, e.g., $\{$like, dislike$\}$. 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, $\{r_{ui}, i \in I_u\}$, serves to represent her tastes. Given the rating history, it is desired to predict ratings for the remaining items that the user has yet to experience, $\{I - I_u\}$.

The main methods used to solve the recommendation problem can be roughly categorized into content-based (CB) and collaborative filtering (CF) approaches (Adomavicius & Tuzhilin, 2011). *Content-based* methods rely on available descriptions of items, representing them in a pre-defined feature space; for instance, given textual descriptions, items may be represented as a vector of weighted term counts, or topics. A *user profile* is constructed in this paradigm as the aggregation of the vectors of items that user is known to have liked (or disliked). The constructed user profile is compared with candidate items in the joint feature space to generate recommendations.

There are several drawbacks of content based methods. They are sensitive to content sparsity; in particular, textual descriptions are often short and sparse. It is further claimed that content-based models tend to over-specialize, as they are inherently biased to favor items similar to the historical ones, whereas in practice, different items may be highly useful to the user. In this work, we experiment with a variant of the Rocchio algorithm (Salton, 1971; Lops, de Gemmis, & Semeraro, 2011), representing *multimedia presentation* items based on their textual content.

Collaborative filtering methods rely solely on the historical record of item ratings by users. These methods model user preferences collaboratively, detecting rating patterns across users and items. Accordingly, CF can make cross-genre or 'outside the box' recommendations. Primary CF approaches include neighborhood-based methods and latent factor models (Adomavicius & Tuzhilin, 2005). The latter approach projects the users and items matrix into a smaller dimensional space, thereby clustering similar users and items. The latent factors represent latent characteristics of the users and items in the system, where high correspondence between item and user factors leads to a recommendation. These methods have become popular in recent years due to good scalability and high predictive accuracy. We experiment in this work with neighborhood-based CF methods, as well as with a state-of-the-art matrix factorization method (Koren, Bell, & Volinsky, 2009).

A well-known weakness of both CF and CB systems is the 'cold start' problem, handling new users (both approaches) and new items (CF), for which there is little rating history available. In addition, rating sparsity is a significant known problem, as in general there are many items available and the odds are that users only share a small number of rated items. This problem is somewhat alleviated by model-based approaches such as matrix factorization (Koren et al., 2009). Finally, the modeling of additional dimensions, such as physical proximity or other contextual aspects, in the recommendation process is non-trivial. For a more detailed discussion of recommendation techniques and their pros and cons, see for instance (Burke, 2002) and (Adomavicius & Tuzhilin, 2005).

3 Related Work

3.1 Context aware recommendation

There has been rising interest recently in context aware recommender systems (CARS) (Adomavicius & Tuzhilin, 2011; Panniello, Tuzhilin, & Gorgoglione, 2014) that adapt recommendations to the specific situation in which items are consumed. Ideally, context information should be integrated directly into the recommendation model. However, state-of-the-art CF approaches such as Matrix Factorization do not provide a straightforward way of doing so (Konstas, Stathopoulos, & Jose, 2009). In order to model contextual information, 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 instead of the traditional two-dimensional user-item matrix. Their model describes context information, such as user age and gender, using categorical features. This approach comes with the cost of increased sparsity–as now information is spread over a multidimensional space instead of the classical (already sparse) two dimensional one. Others have suggested to handle context information by partitioning the original user-item rating matrix, such that ratings with similar contexts are grouped together (Liu & Aberer, 2013). None of these extensions however readily accommodates complex, structured and semi-structured, relational background information, which we model here using the graph-based representation.

3.2 Graph-based recommendation

Several studies have previously explored related graph-based methods for recommendation. Early works used homogenous graphs, in which nodes denoted items and edges represented inter-item similarity (Gori & Pucci, 2007; Yildirim & Krishnamoorthy, 2008). Other works modeled bipartite graphs, consisting of *users* and *items*, having edges represent ratings assigned by users to viewed items (Craswell & Szummer, 2007; Fouss, Pirotte, Renders, & Saeren, 2007; Baluja et al., 2008). Recent studies have already started to explore the potential of graph representation and reasoning techniques for context-aware recommendation. These works construct heterogeneous graphs, consisting of multiple node and edge types. In addition to user-item ratings, the heterogeneous graph structure has been used to model social context, drawn from location-based and event-based social networks (Konstas et al., 2009; Bu et al., 2010; Noulas, Scellato, Lathia, & Mascolo, 2012; Shang, Kulkarni, Cuff, & Hui, 2012; Wang, Terrovitis, & Mamoulis, 2013; Tiroshi, Berkovsky, Kaafar, Chen, & Kuflik, 2013; Tiroshi et al., 2014; Pham, Li, Cong, & Zhang, 2015; Bagci & Karagoz, 2015).

Several works have applied the Personalized PageRank measure to perform item ranking in contextual graphs, reporting superior performance over classical recommendation methods. Konstas *et al.* (Konstas et al., 2009) targeted the recommendation of music tracks to users, modeling social friendships and tags in the graph; in their experiments, PPR outperformed a user-based kNN method. The work by Noulas *et al.* (Noulas et al., 2012) targeted location recommendation–the graph in this case included users and venues as nodes, having edges denote historical user-location visits and user friendships. Recommendation experiments using the PPR measure showed it to be preferable to a set of alternative approaches, including collaborative filtering methods. Due to skewed distribution of visits across venues, simply recommending venues by popularity was shown to perform well, having the graph-based PPR be the only method to beat this popularity baseline. More recently, Bagci and Karagoz (Bagci & Karagoz, 2015) used PPR for recommending locations to visit to users, based on popularity and past 'check-ins' of friends. They too report that the PPR algorithm outperforms classical recommendation approaches. Yao et al. (Yao et al., 2015) proposed a a multi-layer graph-based recommendation framework, incorporating a variety of implicit contextual information into the recommendation process, and showed contextual PPR schemes to perform best out of a set of alternative approaches.

Our work corroborates these previous findings. While the previous works focused on the modeling of social information, we show that the graph approach can effectively leverage relational background knowledge, representing aspects such as physical and thematic relatedness. Interestingly, none of the abovementioned works has examined graph adaptations by means of edge weight tuning (e.g., (Minkov & Cohen, 2011)) for recommendation purposes. We further show that tuning parametric task-specific edge weights substantially affects the global similarity measure, leading to performance gains; as a result, in contrast with previous findings (Konstas et al., 2009), recommendation performance does not degrade but rather benefits from expanding the graph with additional dimensions. A main contribution of this study is in providing a comprehensive comparison of graph-based recommendation with a variety of alternative popular methods, thus providing strong evidence about the advantages and applicability of graph-based contextual recommendation.

3.3 Recommendation systems for museum visitors

This paper investigates graph-based recommendation to enhance the museum visit experience. There exist several research works that consider recommender systems for museum visitors. Stock *et al.* (Stock et al., 2007) proposed an overlay user model, having the user model first 'overlaid' over a domain ontology, rating parts of the ontology according to the user's preferences, and then propagating these ratings over the ontology network to recommend other exhibits of interest. Grieser *et al.* (Grieser et al., 2007) aimed at predicting the exhibit that a visitor would visit next based on available visit history. They applied a Naive Bayes learning model, considering exhibit proximity, textual description of the exhibit, and exhibit popularity. In their experiments, the baseline of exhibit prediction based on popularity was found to be most successful. The graph-based approach proposed here is more comprehensive than these works, as it models collaborative user history information jointly with content-based and physical proximity aspects. We believe that the graph based approach is generally advantageous compared with statistical learning in conditions of data sparsity.

Bohnert *et al.* (Bohnert et al., 2012) suggested 'Gecko mmender', a system that takes as input visitors' explicit ratings of exhibits, and uses a nearest-neighbor CB approach to predict ratings for unvisited exhibits. These ratings form the basis for theme/tour recommendations. Gecko mmender was evaluated in a field study at Melbourne Museum (Melbourne, Australia), focusing mainly on assessing different display modes of the predicted ratings. In a recent work by Bohnert *et al.* (Bohnert & Zukerman, 2014), they target the prediction of museum visitors viewing times of exhibits. They experiment with Spatial Process Model (SPM), a collaborative model based on the theory of spatial processes. SPM represents visitors viewing-times as Gaussian random vectors. It is assumed that the correlation between observations increases with decreasing exhibit conceptual distance, encoded in a covariance matrix. The method is claimed to alleviate the cold-start new-user problem through the modeling of correlation between the exhibit areas viewed by the visitor and the other exhibit areas. Several types of exhibit distances are evaluated, including viewing-time similarity, semantic similarity and walking distance. The results, evaluated using a dataset of 157 visitor histories at the Melbourne Museum, indicate that physical distance correlations yield the best predictions of viewing times. The graph-based approach described here is in fact complementary to the SPM model–one may generate multi-facet correlation scores using the global graph-based relatedness measure, and use them for viewing time prediction and related problems.

Finally, another recent work by Bartolini *et al.* (Bartolini et al., 2014) addresses recommendation of diverse multimedia materials across cultural heritage sites. They use a graph representation to propagate semantic similarity between items, based on available 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 the assessment of visitor satisfaction using the system in field conditions, whereas we focus in the evaluation of the quality of multi-facet graph-based recommendation compared with alternative recommendation approaches.

4 Graph-based recommendation

We represent background knowledge alongside historical rating information in a joint graph. The graph is heterogeneous, consisting of typed nodes and directed and typed edges, and can therefore accommodate knowledge in a relational form. Dedicated graph-based measures may be employed to rank the graph nodes by their similarity, or relevancy, to a *query* of interest, represented as a distribution over the graph nodes. We construct queries that correspond to a user's profile, including items (*multimedia presentations*) that she is known to have liked, having items ranked by their relevancy to this profile. Following previous works, we employ the Personalized PageRank random walk algorithm to assess inter-node similarity in the graph. We further exploit the relational structure of the graph, and tune parametric edge weights to optimize performance.

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

4.1 The museum as a graph

Let us first define the underlying graph representation. A graph $G =< V, E >$ consists of a set of nodes V, and a set of labeled directed edges E. We will 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 now turn to describe the museum's environment in the form of such a relational graph. We distinguish between the following entity classes, representing them as distinct node types:

- *Positions*. A position is a *point of interest* (POI), a physical point in which multimedia information is available about exhibits nearby. The POIs are spread in the museum environment over multiple rooms and floors.

- *Presentations*. These nodes represent multimedia presentations offered for viewing on the visitor's mobile device. Presentations are associated with

source type	**edge type**	**target type**
presentation	located-in	*position*
	has-theme	*theme*
	similar-to	*presentation*
	viewed^{-1}	*visitor*
position	nearby	*position*
	located-in^{-1}	*presentation*
theme	has-theme^{-1}	*presentation*
	similar-theme	*theme*
visitor	viewed	*presentation*

Table 1: Relation types in the museum graph

concrete exhibits. As of today, once the user is tracked at some point of interest, she is offered to view the presentations pertaining to the exhibits associated with that position.

- *Themes*. The multimedia presentations in Hecht Museum have been associated with a set of nine specific themes, e.g. Religions, Art symbols and Maritime, following the process described in Katz *et al* (Katz et al., 2006). We represent each of these themes as a node in the graph.
- *Visitors*. These nodes represent individual visitors to the museum. While the other node types describe static aspects of the museum's layout, historical visit information is dynamic, being accumulated over time. The *visitors* will be associated with the museum entities through dedicated edges that describe their visit experience.

The full set of graph edge types is detailed in Table 1. As mentioned before, *multimedia presentations* are offered per specific museum exhibits. Thematically related exhibits are grouped together physically in separate physical *positions*. We therefore directly link each *presentation* with the *position* with which it is associated over an edge of type *located-in*. *Has-theme* edges further associate 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 semantic relations *located-in*$^{-1}$ and *has-theme*$^{-1}$.

We further model a set of edges describing various aspects of inter-entity similarity in the museum's environment. *Positions* that reside in high physical proximity are linked over *nearby* edges. Physical proximity was measured in terms of

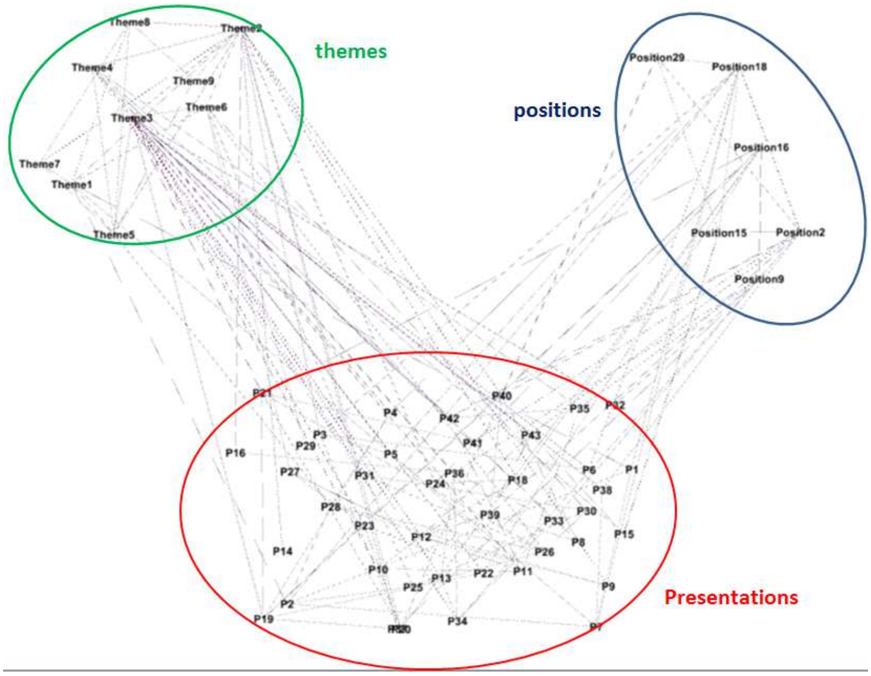


Figure 1: An illustration of the museum graph: *multimedia presentation* items are inter-linked via common physical *positions* and annotated semantic *themes*.

walking steps for this purpose. In addition, *presentations* that exhibit high content similarity are inter-linked over the *similar-to* relation. We compute term-based content similarity: having each *presentation* represented as a TF-IDF term-weighted vector, cosine similarity is computed between every *presentation* node pair. Finally, *similar-theme* edges connect semantically related *themes*. Each theme is represented for this purpose as the centroid of the TF-IDF vectors of the *presentations* associated with that theme, having cosine similarity computed between the theme centroids. More details about the representation of inter-item similarity as edges are provided in Section 8.

The graph schema described thus far represents *multimedia presentations* alongside *location* and *theme* entities, linking these objects over relations that denote thematic relations and physical proximity. Figure 1 illustrates the resulting graph for a subset of our experimental data.

It is straightforward to further incorporate historical ratings in the graph. We represent individual visitors by dedicated *visitor* nodes, linking every *visitor* to *presentation* nodes over directed *viewed* edges. One may link a visitor node to all of the presentations that she viewed; or, it is possible to link the visitor node only to presentations that she is known to have liked. As described in Section 5.1, most of the feedback scores in the visit logs in our datasets are positive. For this reason, and considering data sparsity, we follow the first option in this work. Node pairs linked over the *view* relation, are linked over edges of the inverse semantic type, *viewed*$^{-1}$, pointing in the opposite direction. In this fashion, the graph models *presentations* and *visitors* association patterns based on available ratings history.

4.2 Recommendation with Personalized PageRank

We are interested in recommending presentations to a user based on her historical feedbacks. This corresponds to the following objective: given a distribution over the graph nodes that represents the user's tastes, *presentation* nodes are to be ranked by their relatedness to that distribution. Various measures exist that evaluate structural node relatedness in graphs (Fouss et al., 2007). We employ here the popular Personalized PageRank (PPR) random walk metric, sometimes referred to as Random Walk with Restart (RWR) (Tong et al., 2006).

PPR is often described as a variant of the well-known PageRank algorithm (Page, Brin, Motwani, & Winograd, 1998), originally designed to measure the relative importance of individual Web pages. Consider the Web graph, in which nodes denote Webpages and directed edges denote hyperlinks. The non-personalized PageRank algorithm models the behavior of a Web surfer, who at any given time, chooses either to follow a hyperlink to a related Webpage, or "resets" randomly to one of the pages on the Web. Formally, given that the surfer is at node (page) i, with probability α the surfer is expected to move to node j following an outgoing link from i, and with probability $(1 - \alpha)$ the user resets randomly to some page. The probability distribution of finding the surfer at each of the graph nodes at time d, V_d, is defined recursively as:

$$V_{d+1} = (1-\alpha)[\frac{1}{N}]_{1\times N} + \alpha \mathbf{M} V_d \tag{1}$$

where the total number of nodes (pages) is N, and the transition matrix $\mathbf{M}$ encodes the probability that the surfer moves to page j from page i following a hyperlink. As default, $\mathbf{M}$ distributes a node's probability uniformly among the pages it links to, i.e.

$$\mathbf{M}_{ij} = \begin{cases} \frac{1}{|ch(i)|} & \text{if there is an edge from } i \text{ to } j \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $ch(i)$ is the set of nodes that can be reached over an outgoing link from i (the 'children' of i).

Due to the reset operation, the unified matrix is stochastic and irreducible. This guarantees that the random walk process converges to a unique stationary distribution, V^*. The *PageRank score* of node j, p_j, is its probability in the stationary state V^*, giving a measure of document centrality in the network.

The PageRank algorithm computes 'universal' node importance scores, ignoring user preferences. The Personalized PageRank variant (Page et al., 1998; Richardson & Domingos, 2002) preserves an association between node rankings and user preferences, or a 'query'. Rather than assume that the user is equally interested in, and would reset to, any graph node uniformly at random, the enhanced random walk scheme limits the reset operation to the query nodes, as follows:

$$V_{d+1} = (1-\alpha)V_u + \alpha \mathbf{M} V_d \tag{3}$$

where V_u denotes the query distribution, including nodes that are known to be of interest to user u. The Personalized PageRank scores are derived from the corresponding stationary state distribution.

The generated PPR scores reflect structural similarity, or relevancy, of the graph nodes 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 (including cyclic paths, and paths that cross z multiple times), weighted by path traversal probabilities (Jeh & Widom, 2003; Fogaras, Rácz, Csalogány, & Sarlós, 2005). Importantly, due to the reset probability $(1-\alpha)$, the paths between x and a destination node z are weighted exponentially lower as their length increases. Intuitively, this means that items that are connected over short paths to the query nodes are considered more relevant by the PPR method; similarly, items that can be reached over multiple paths from the query nodes are also considered more relevant.

Edge weights. The graph walk process is determined by the graph's topology, captured by the transition matrix $\mathbf{M}$. If edges are assumed equal importance, then $\mathbf{M}$ distributes evenly the transition probabilities across all of the outgoing edges from node i (Eq. (2)). It is reasonable to assume however that the random surfer is inclined to traverse specific edges, which reflect a stronger, or more meaningful, semantic relations. In this work, we will assume that edge importance is derived from its type (Minkov & Cohen, 2011; 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 as:

$$\mathbf{M}_{x,y} = \frac{\theta_\ell}{\sum_{y' \in ch(x)} \theta_{\ell'}} \tag{4}$$

where θ_ℓ is the weight of the outgoing edge from x to y. 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 thus directly affect the probability flow in the graph.

The graph edge weights Θ can be set manually, according to prior beliefs; tuned empirically; or, learned from labeled examples (Minkov & Cohen, 2011). In this work, 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 & Cohen, 2010; Lao, Minkov, & Cohen, 2015) for future work.

4.3 Recommendation at the museum

We use Personalized PageRank to generate recommendations for individual visitors. A user profile is constructed based on available feedback scores; the query distribution V_u spans over the set of *multimedia presentation* nodes that the visitor is known to have seen and liked, weighting them by the respective feedback scores. Applying PPR yields a score distribution over the graph nodes, reflecting graph-based relatedness (or, similarity) to the visitor's profile. It is straight-forward to filter this distribution by node type, and present a list of *presentation* items ranked by their estimated relevancy to the user. Likewise, one may generate rankings of other entity types, e.g., *positions*.

An advantage of the graph walk is that it integrates multiple types of evidence. A random walk process that is initiated at some *presentation* x of interest will reach a related *presentation* z by passing via a shared *theme* node, or directly, due to the modeling of text-based similarity edges, over the following paths: $\text{x} \overset{has-theme}{\longrightarrow} \text{y} \overset{has-theme^1}{\longrightarrow} \text{z}$, and $\text{x} \overset{similar-to}{\longrightarrow} \text{z}$. Importantly, the PPR measure is transitive–as the random walk continues, similarity propagates between pairs of related *presentations* continuously.

The graph walk further models physical proximity, giving preference to presentations of exhibits that are located nearby. Immediate physical proximity (same position) is expressed via the 2-hop path $\text{x} \overset{located-in}{\longrightarrow} \text{y} \overset{located-in^1}{\longrightarrow} \text{z}$; presentations of exhibits at nearby positions are reached via the 3-hop path $\text{x} \overset{located-in}{\longrightarrow} \text{y} \overset{nearby}{\longrightarrow} \text{s} \overset{located-in^1}{\longrightarrow} \text{z}$. Presentations at yet further positions are reached over longer paths, e.g., $\text{x} \overset{located-in}{\longrightarrow} \text{y} \overset{nearby}{\longrightarrow} \text{s} \overset{nearby}{\longrightarrow} \text{t} \overset{located-in^1}{\longrightarrow} \text{z}$.

Collaborative aspects are modeled through the path: $\text{x} \overset{viewed^{-1}}{\longrightarrow} \text{y} \overset{viewed}{\longrightarrow} \text{z}$. Collaborative and content-based similaritiea are naturally integrated by mixed paths like $\text{x} \overset{viewed^{-1}}{\longrightarrow} \text{y} \overset{similar-to}{\longrightarrow} \text{s} \overset{viewed}{\longrightarrow} \text{z}$.

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 effectively approximated by limiting the graph walk to a finite number of steps k (Toutanova, Manning, & Ng, 2004; Fogaras et al., 2005; Minkov & Cohen, 2011).

5 Experimental Setup

This section first introduces our experimental data. It then defines the evaluation methodology, presenting two variants of the recommendation tasks and the evaluation measures used in this work.

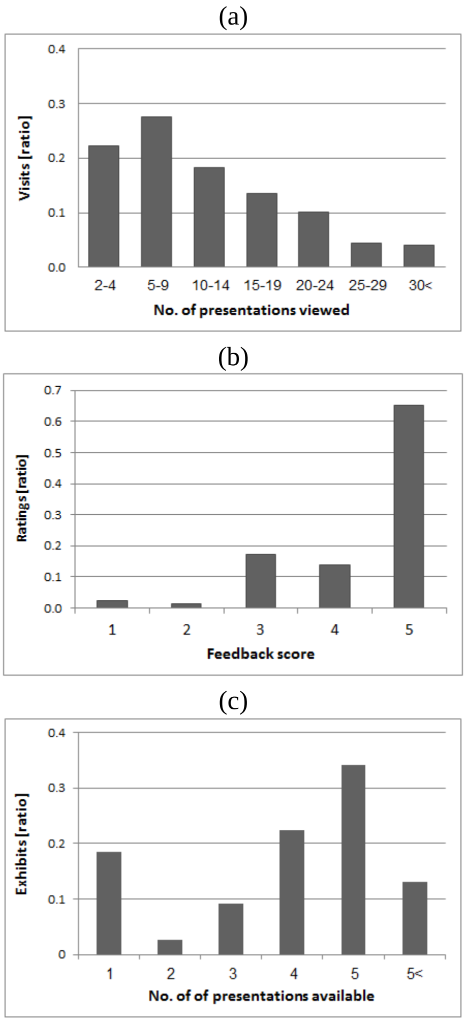
5.1 Data

We experiment with authentic data collected during visits to the Hecht Museum. A mobile device is provided to the museum visitors, on which they are offered to view relevant multimedia presentations at each POI visited (for details see Kuflik *et al* (Kuflik et al., 2011; Kuflik, Wecker, joel Lanir, & Stock, 2014)). The multimedia presentations include limited textual content; most of them contain between 100–200 words, corresponding to a paragraph or two. Overall, more than 300 presentations are available that correspond to 76 exhibits, displayed at 49 positions, spread across multiple rooms and two floors. The mobile device is designed to present and receive user feedback. Colored smiley emoticons 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 presentation at the time of this research.) Once viewed a presentation, the user is requested to provide her own feedback.

The dataset used in the experiments is composed of 319 visit logs. The recorded information includes the POIs that the visitor passed through, presentations viewed and the feedback scores (ratings) assigned to these presentations. Log files of visits in which a single presentation has been viewed are not included in our dataset. The visit logs are anonymous, as we do not have access to personal details about the visitors. We will assume that each visit log was generated by a different user, as the ratio of repeated visits at the Hecht Museum is generally low.

Figure 2 presents various dataset statistics. Specifically, Fig. 2(a) describes the number of presentations viewed per visit. As shown, the median number of presentations viewed is 10, where visitors most often viewed between 5-9 presentations during their visit. Considering the limited number of viewed items, it is desired to direct the visitors to those items that are most interesting for them. At the same time, it is challenging to make recommendation as little is known about the visitor.

The visit logs that constitute our dataset were obtained at several different



(a)

(b)

(c)

Figure 2: Dataset statistics: (a) Number of presentation viewed per visit; (b) Distribution of feedback scores; (c) Number of presentations available per exhibit

points in time, and therefore exhibit some variance with respect to rating scales. Some of the feedbacks are binary (*like, dislike*), but the majority of feedbacks are 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 $\{$1,3,5$\}$.

Fig. 2(b) 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 negative. In order to alleviate sparsity, we choose to model all of the presentations viewed by a visitor, for which the feedback scores are positive (in the range 3–5), as her user profile, weighting the items in the profile by their normalized feedback score.

Fig. 2(c) presents statistics about the number of presentations available per exhibit. There is a single relevant presentation for 18% of the exhibits. For about half of the exhibits, 5 or more presentations are available. As discussed below (Sec. 5.2), we evaluate in our experiments a setting in which the set of presentations available for a given exhibit are ranked by their relevancy to the individual visitor.

We find that the 'popularity' of the various presentations, estimated in terms of the number of times viewed, varies greatly. About 1% of the presentations were viewed in 100-119 visits; 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. There are about 14 feedbacks available per each viewed presentation item on average in our dataset.

In summary, the experimental dataset is sparse. 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. Content-wise, the textual descriptions of target items are short and sparse.

5.2 Experimental design

We perform a set of *prediction* experiments using the authentic visit logs collected at the Hecht museum.

Given the set of ratings provided by user u, we consider in each experiment one of the rated presentation items, $i^* \in I_u$, having the remaining items $\{I_u - i^*\}$ 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 recommended *presentation* items. This experimental setting is imperfect, mainly, other highly ranked items may be of high interest to the user as well, 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).

Since the experimental dataset is limited in size, we perform exhaustive leave-one-out prediction experiments–for every user u in the dataset, up to $|I_u|$ labeled example queries are generated. Since we are interested in predicting items that the visitor is known to have liked, presentations for which u assigned low ratings are excluded from the example pool. (Specifically, we discarded presentations for which the feedback score was below the median score of 3, constituting 3.7% of all user feedbacks.)

We dedicate a *held-out* portion of the resultant example set for parameter tuning purposes, mainly, 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.[2] The examples derived from the remaining 287 (90%) visit logs serve for evaluation using the leave-one-out procedure.

Finally, we consider two modes in which personalized recommendation of *multimedia presentations* can enhance the museum visit experience:

- *General Recommendation:* We are interested in assisting the visitor in choosing items of interest while touring the museum. All of the *multimedia presentations* (except those already viewed by the user) are considered as candidates for recommendation in this setting. Ideally, the museums physical layout should be taken into account in this mode. Physical distances are modeled in the graph via the *nearby* edges, connecting adjacent *positions*, and thus affect the generated rankings. The explicit design of a route planner module is out of scope of this work however.

- *Per-Exhibit Recommendation:* As visitors are moving in the museum space, they may be most interested in presentations about exhibits in their vicinity. The set of relevant multimedia presentations may be small, however the size of the mobile screen is limited, and the top listed items draw most of the attention of a human user. We therefore wish to rank the relevant presentations according to the visitor's personal preferences. In this case, only *presentations* that are associated with an attended exhibit are the candidate items for recommendation.

5.3 Evaluation measures

All of the recommendation methods considered in this paper generate scores for candidate items given the profile of target user u. These scores are then processed

[2] As indicated previously by Minkov and Cohen (Minkov & Cohen, 2011), the graph edge weights parameters can be effectively tuned using a small number of examples.

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. Formally, the non-interpolated recall at rank k of a given list is defined to be 0 for each rank $k = 0, ..., k_{i-1}$, where k_i is the the rank that holds the item that should be predicted, and 1 for ranks $k \geq k_i$. The (mean) recall@k averages the recall scores at each rank k of multiple queries. For example, recall@3=0.7, means that for 70% of the queries, the correct answer appears among the top 3 ranks of the retrieved lists.[3]

Given the limited size of the screen of handheld devices, only a limited number of items can be presented at a time. It is further expected that the highest ranking items will receive most of the user's attention. For this reason, we report $recall@k$ for the topmost ranks: k =[1–5] for the general recommendation setting, and k =[1–3] in the per-exhibit mode. In the latter case, the number of candidate items is small to begin with, 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). This measure considers the full ranked lists generated. The reciprocal rank of a response to a single query is defined as the multiplicative inverse of the rank of the correct answer: $\frac{1}{k_i}$. The mean reciprocal rank is the average of the reciprocal ranks for all of the test queries.[4]

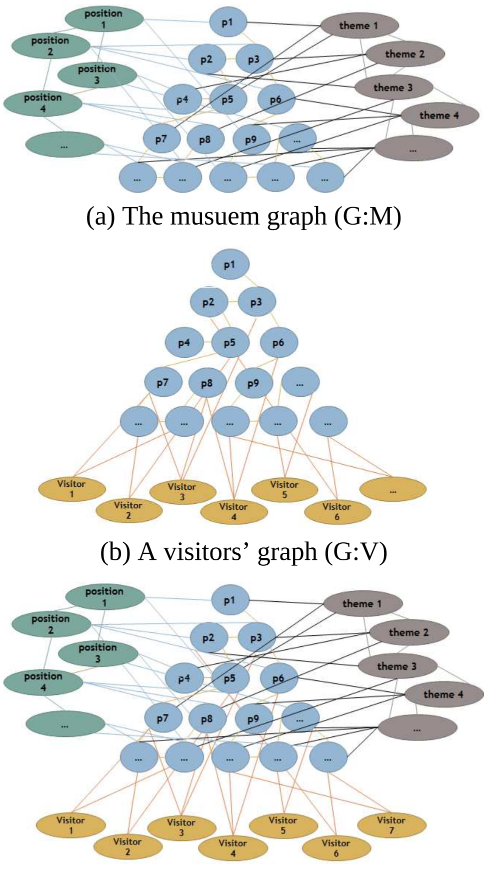
Ratio of failed tests (RFT). Occasionally, a recommendation method may fail to predict scores for some items, possibly including the target item. In such cases, we assign these items a de-facto score of zero, appending them to the bottom of the output ranked list. Impaired coverage therefore directly 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.

6 Experiments

We compare the graph-based approach against a set of popular recommendation methods, including content-based, memory-based nearest-neighbor collaborative filtering (CF) algorithms, and state-of-the-art matrix factorization. In addition,

[3] We consider the *effective* rank of the target item, which may be a real number, e.g., 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; e.g., rank 5.5 contributes to recall@6 and downwards.

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



(a) The musuem graph (G:M)

(b) A visitors' graph (G:V)

(c) A unified graph (G:U)

Figure 3: Illustration of the structure of several graph variants

several baselines are evaluated that are either naive or well-informed, such as recommending items by their popularity. This section includes a description of the evaluated approaches and implementation details.

6.1 Graph variants

In order to assess the utility of combining historical ratings with background knowledge, we experiment with several graph variants, modeling these aspects independently and in combination. Figure 3 illustrates the structure of these graph variants.

1. *The Museum graph (G:M)*. This graph describes background knowledge about the museum's environment. It models thematic and physical simi-

larities. Specifically, the graph includes *presentation*, *theme* and *position* entities, connected over the relations *similar-to*, *similar-theme*, *nearby*, *has-theme*, *located-in* and the respective inverse edges.

2. *Visitors graph (G:V).* This graph variant represents ratings history. The graph is bi-partite–it includes *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 3(c).

4. *Combined graphs (G:MV).* Another approach for combining the two information sources is to integrate the scores produced using the *visitors* and *museum* graph variants. We experiment with a linear combination of the scores:

$$\widehat{r}_{ui}(v_u, G:MV) = (1-\beta)\cdot \widehat{r}_{ui}(v_u, G:M) + \beta \cdot \widehat{r}_{ui}(v_u, G:V) \quad (5)$$

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, as described in detail in Section 8.

We set the damping factor of the random walk process (Eq. (3)) to α =.85 following previous works (Page et al., 1998; Minkov & Cohen, 2011).[5] We approximate PPR scores using finite random walk repeated for 6 iterations. As discussed in Section 4.2, the impact of additional steps on the generated rankings is negligible.

6.2 Content-based Recommendation (CB)

We experiment with a version of the Rocchio algorithm (Salton, 1971; Pazzani & Billsus, 1997; Lops et al., 2011). 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. Item relevancy is then estimated using cosine similarity in this vector space.

We represent each candidate *multimedia presentations* as a vector of TF-IDF weighted terms, describing its textual contents. Since in this study, only a small fraction of item ratings are below the median score, we only model positive feedbacks. In computing the user profile vector, we weight the individual presentation vectors by the respective feedback scores. In this fashion, highly liked presentation

[5] The produced PPR rankings are generally insensitive to α value, e.g., (Minkov & Cohen, 2011).

$$userSim(u,v) = \frac{\sum_{i \in I_{uv}} (r_{ui} - \overline{r}_u)(r_{vi} - \overline{r}_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - \overline{r}_u)^2} \sqrt{\sum_{i \in I_{uv}} (r_{vi} - \overline{r}_v)^2}} \tag{6}$$

$$\widehat{r}_{ui} = \overline{r}_u + \frac{\sum_{v \in N_i(u)} userSim(u,v) \times (r_{vi} - \overline{r}_v)}{\sum_{v \in N_i(u)} \mid userSim(u,v) \mid} \tag{7}$$

$$itemSim(i,j) = \frac{\sum_{v \in U_{ij}} (r_{vi} - \overline{r}_v)(r_{vj} - \overline{r}_v)}{\sqrt{\sum_{v \in U_{ij}} (r_{vi} - \overline{r}_v)^2} \sqrt{\sum_{v \in U_{ij}} (r_{vj} - \overline{r}_v)^2}} \tag{8}$$

$$\widehat{r}_{ui} = \overline{r}_i + \frac{\sum_{j \in N_u(i)} itemSim(i,j) \times (r_{uj} - \overline{r}_j)}{\sum_{j \in N_u(i)} \mid itemSim(i,j) \mid} \tag{9}$$

contribute more to the user profile compared with a presentation that was assigned lower scores.

6.3 Collaborative filtering methods

We experiment with two well-known variants of neighborhood-based CF methods, namely *user-based* and *item-based* k-nearest-neighbor (kNN) recommendation. We follow closely on Desrosiers and Karypis (Desrosiers & Karypis, 2011) in our implementation of these methods. We further experiment with a state-of-the-art matrix factorization algorithm (Koren et al., 2009), as detailed below.

User-based kNN (CF:U-kNN) 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 based on their historical ratings. Here, inter-user similarity is evaluated using Pearson's correlation as defined in Eq. (6), where I_{uv} denotes the set of items co-rated by users u and v, and $\overline{r}_u$ and $\overline{r}_v$ denote the average 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 Eq. (7), weighting the contribution of each neighbor by its similarity to u.

Item-based kNN (CF:I-kNN) This method evaluated the recommendation score by analyzing the ratings of similar items. As defined in Eq. (8), 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 $\widehat{r}_{ui}$ 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 Eq. (9).

In our experiments, we tune the neighborhood size parameter k so as to optimize performance. We apply *negative filtering*, discarding neighbors with a negative correlation score, as negative Pearson correlation has been argued to be unuseful for recommendation (Herlocker, Konstan, Borchers, & Riedl, 1999). Several additional threshold types were evaluated in order to identify a high-quality set of neighbors. The first discards neighbors for which the correlation score is lower than a *minimum similarity* threshold value. Another threshold type requires some minimal number of common ratings to establish neighborhood relationship: in user-based kNN, a neighbor must have at least θ_i items co-rated with the test user; using item-based kNN, neighbor items are required to have θ_u users who co-rated items i and j. The various combinations of threshold values and types, as well as the neighborhood size, were evaluated exhaustively. We report results using the best joint parameter assignments per setting.

Matrix factorization (CF:MF) We experiment with a matrix factorization formulation outlined by Koren *et al* (Koren et al., 2009). In general, 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. The factor vectors are learned by minimizing the regularized squared error on the set of known ratings:

$$min \sum_{(u,i)\in d} (r_{ui} - q_i^T p_u)^2 + \lambda(\|q_i\|^2 + \|p_u\|^2) \tag{10}$$

Here, d is the set of the (u, i) pairs for which r_{ui} is known. The constant λ controls the extent of regularization.

We report MF results using the implementation included in the GraphLab software package[6] (Low et al., 2012), applying 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 has been 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.

[6] http://graphlab.org/

6.4 Hybrid method (Hyb)

Hybrid recommender system combine multiple techniques, ideally compensating for the weakness of the individual methods (Burke, 2002; Berkovsky, Heckmann, & Kuflik, 2009). We experiment with a combination of two recommendation techniques: content-based (CB) and item-based kNN (CF:I-kNN).[7] 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.

6.5 Baselines

We further consider several baseline approaches:

Random (B:R) This naive non-personalized baseline selects one of the candidate *presentations* uniformly at random. Comparing against this baseline, we will demonstrate the contribution of informed recommendation systems over chance.

Proximity (B:PR) This method models the assumption that museum visitors prefer to view presentations for exhibits located nearby. It is implemented as a stricter version of random recommendation, limiting the set of candidate presentations in terms of distance from the user's whereabouts. The set of candidate presentations is constructed as follows. Given V_u, we obtain the list of items associated with already visited, i.e., *presentations* that can be reached from any *presentation* $x \in V_u$ over the path $\text{x} \overset{located-in}{\longrightarrow} \text{y} \overset{located-in^{-1}}{\longrightarrow} \text{z}$. We also consider presentation relevant for nearby positions, reached over the path: $\text{x} \overset{located-in}{\longrightarrow} \text{y} \overset{nearby}{\longrightarrow} \text{q} \overset{located-in^{-1}}{\longrightarrow} \text{z}$.

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 non-personalized yet informed and often hard-to-beat (Lucchese, Perego, Silvestri, Vahabi, & Venturini, 2012). A challenge of any personalized recommendation is whether it can outperform this one-fits-all approach.

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

	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

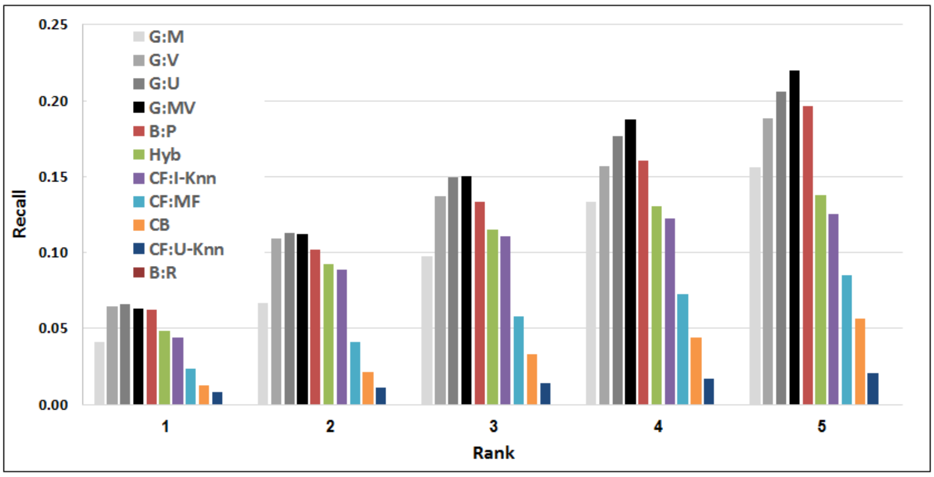
Table 2: MRR and RFT performance of the various recommendation methods. An asterisk denotes statistically significant difference compared with the graph-based method (p-val<0.0045).

7 Main results

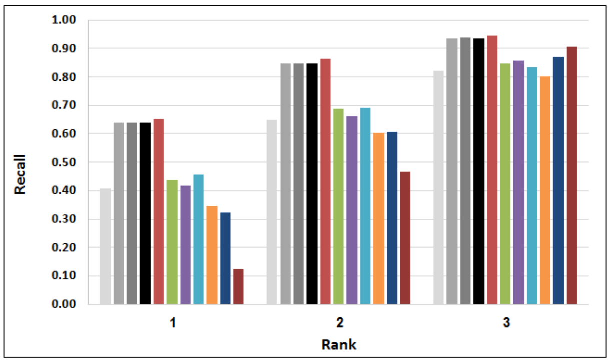
This section reports our results using the different methods. As mentioned before, parameter values have been tuned so as to optimize performance. Importantly, the parameters of the graph-based methods were tuned using the held-out examples; in contrast, all other methods have been optimized directly on the *test* data. Despite the comparison being strict in this fashion, graph-based recommendation is shown to give preferable results. In what follows we first describe our findings in the general recommendation setting, and then discuss per-exhibit recommendation.

General recommendation Table 2 includes MRR and RFT results for the general recommendation scenario, and Figure 4(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 .151 and .152, respectively. In terms of recall, G:MV gives slightly better performance, yielding recall of .063 at the topmost rank, and .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 .140. The other more naive baseline–*random* and *proximity* based recommendation–result in very low MRR



(a)



(b)

Figure 4: Recall-at-rank performance of the evaluated methods in the general recommendation mode (top) and per-exhibit mode (bottom): 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).

performance, as well as negligible recall at the top levels. These results correlate with our dataset statistics, characterized with a large number of candidates, for which the distribution of available feedbacks ('popularity') 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 .084, matrix factorization–.065, and user-based kNN–a low .028. The weakness of these methods can be attributed in part to data sparsity issues. As shown in Table 2, the ratio of failed tests is very high using item-based and user-based kNN CF approaches (.874 and .508, respectively). We found that in many cases, there were no relevant 'neighbors' identified due to rating sparsity. Moreover, many visitors in our dataset (30%) assigned the same feedback score to to all of the presentations that they have viewed, and therefore did not contribute to the recommendation process (see Equations (6) and (8)). Matrix factorization techniques are generally more robust to sparsity issues, but deliver mediocre results here. It is possible that CF:MF 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. We conclude 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 some bias towards highly connected nodes (Tong & Faloutsos, 2006), thus implicitly modeling item popularity information.

Finally, performance of content based (CB) recommendation falls behind its graph counterpart, G:M, which models background information (.048 vs. .110 in MRR). The hybrid method (Hyb) improves upon each of its component systems, decreasing RFT and yielding a somewhat disappointing MRR of .104. Again, we conjecture that the graph-based techniques are preferable in conditions of sparse data. The graph method further integrates physical proximity aspects, which are missing from either the collaborative or content-based approaches.

Per-exhibit recommendation The set of candidate items for recommendation in the per-exhibit setting is limited to a small number of *multimedia presentations* directly associated with a specific exhibit, which the museum's visitor is known to be attending (see Figure 2(c)). Accordingly, random recommendation achieves high recall levels at the top three ranks (.12,.47 and .91), as shown in Figure 4(b), and a high MRR score of .506 (Table 2). The *Popularity* baseline achieves the strongest performance overall in this setting–yielding MRR of .801. This result however may indicate a bias towards the presentations displayed at the top of the mobile screen at the time of data collection, as users tend to view (and like) the top listed items. We therefore believe that the results using B:P should be ‘taken with a grain of salt’ in this case.

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 .788 and .789, respectively. The visitors graph G:V gives roughly equal performance (MRR of .788). The contribution of the modeled background knowledge seems negligible in this case. Indeed, physical proximity, which is modeled in the museum’s graph, is irrelevant in the per-exhibit setting.

Also in this setting, the CF methods yield inferior results. The best performing CF method is .658 in MRR, obtained using MF, compared with .788 using the variant G:V, which models similar information. Likewise, the results using CB recommendation are inferior to the counterpart graph-based variant G:M–MRR results are .585 vs. .627. The *hybrid* recommendation approach improves over its component methods, giving a comparable result to CF:MF (.659), and lower performance compared with any graph method that considers the historical ratings.

8 Impact of graph tuning

The representation of structured and semi-structured information as an entity-relation graph is natural, yet involves some design choices. We discuss in this section issues related to graph design, including an evaluation of the impact of the edge weight parameters Θ on recommendation performance.

Graph design. The proposed graph schema directly links similar *presentation* node pairs, as well as similar *theme* node pairs. As mentioned before, we compute inter-node similarity based on the textual content associated with the presentations and themes for all relevant node pairs.[8]

It is generally desired to avoid the modeling of weak associations as graph edges–weak links are uninformative, while increasing the cost involved in computing the PPR measure. We therefore selectively link only those entity pairs for which the computed similarity scores exceed some tuned threshold. The threshold values for the *similar-to* and *similar-theme* edge types have been set based on the training data to 0.2 and 0.4, respectively. Consequently, *multimedia presentation* nodes are linked over *similar-to* edges to 2.9 other *presentation* nodes on average;

[8] 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.

	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**$^{(+7.3\%)}$

Table 3: Evaluation of graph variants

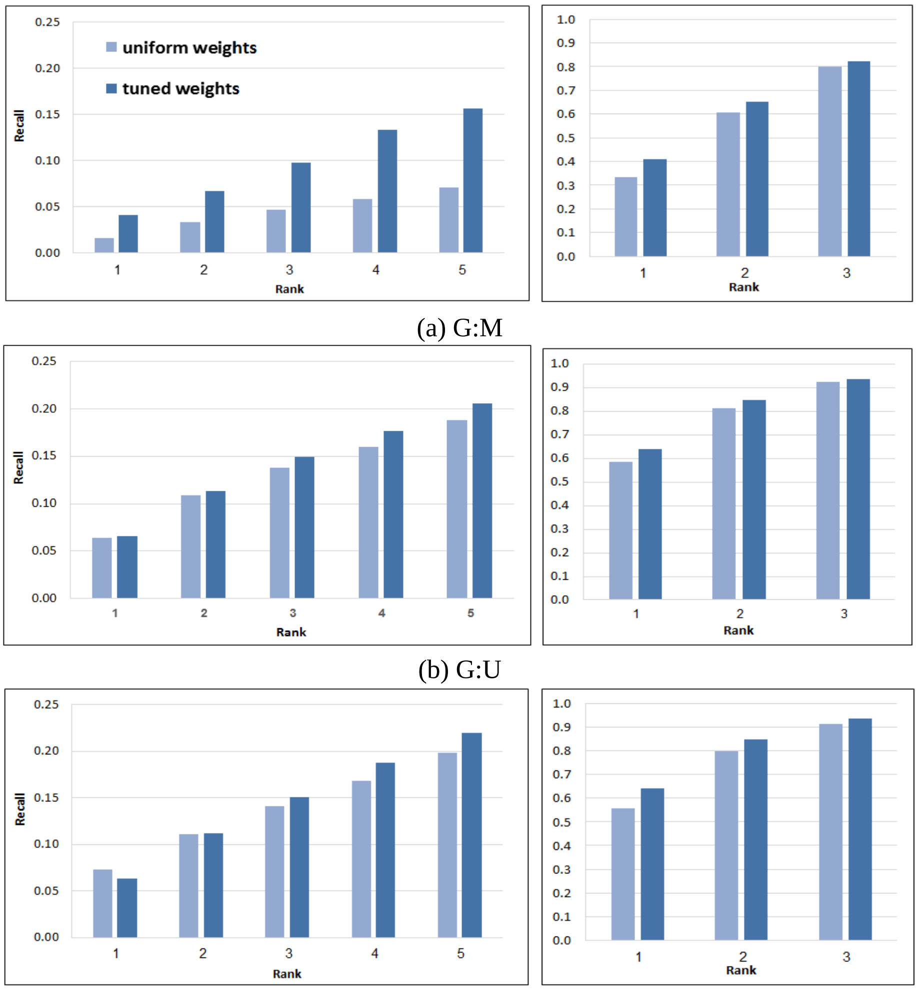
and, most of the *theme* nodes are connected over the *similar-theme* relation to 3-4 other *theme* nodes. Similarly, we link each *position* node to its 3 nearest *positions* in terms of walking distance (and up to 6 positions in case of a tie). Consequently, the similarity-based edge types are asymmetric, e.g., node x may point to node y over a *similar-theme* edge, whereas y may not be point to x.

Edge weight tuning. The parametric edge weights Θ provide another mechanism for controlling the probability flow in the graph. 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.[9] The *viewed* edges form an exception–these edge weights were set according to the feedback scores assigned by the visitor to each presentation, so that stronger association is maintained between *users* and *presentations* that they are known to have liked, compared with items that they liked less or were indifferent to. 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). Finally, we assigned bi-directional edge types (e.g., *viewed* and *viewed*$^{-1}$) identical weights, so as to reduce the cost of edge weight tuning.

It is informative to examine the effect of parameter tuning on recommendation performance. Table 3 shows MRR performance of the different graph variants using uniform pre-tuned edge weight parameters (MRR^U) against the final performance figures achieved using the tuned weights. Figure 5 further demonstrates recall-at-rank-k results, prior to and post parameter tuning. Notably, the visitors graph G:V only includes *viewed* relations, and was not affected in the tuning process. Equal coefficients ($\alpha = .5$) were used in the pre-tuned version of the G:MV graph variant.

As shown, edge weight tuning was highly effective for the *museum* graph, increasing its MRR result by roughly 90%, from low .058 to .110 in the general recommendation setting, and by about 8% in the per-exhibit setting. The performance of the G:U and G:MV variants was relatively high in their pre-tuned versions; yet,

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



(a) G:M

(b) G:U

(c) G:MV

Figure 5: 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.

edge weight tuning improved results further, increasing MRR by 1.3–7.3%, thus obtaining the best result overall. Similar trends are observed with respect to recall@k performance.

We observed that low weights were assigned to the text-based similarity relations, namely *similar-to* and *similar-theme* edge types. We attribute the ineffective modeling of content-based similarity to text sparsity. In contrast, the structural association between *presentations* and their *themes* over the *has-theme* edges was assigned high weights. The weight of the *nearby* edge type was set to a high value in the general setting; in the per-exhibit setting, however, its weight was low; this settles with the fact that physical proximity is irrelevant when all candidate items are associated with a single exhibit. Overall, this demonstrates the flexibility of the graph approach: the very same graph is effectively optimized per recommendation task by tuning its parameters.

9 Conclusion

We have described a graph-based framework for personalized recommendation of multimedia presentations 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; for example, we consider here available expert annotations of the multimedia presentations with a set of perspectives, which may characterize the user's interests. Another aspect that must be modeled at the museum is its layout information–adjacent museum exhibits are typically semantically related, and items associated with exhibits nearby are likely to be preferred by the user.

It has been demonstrated that the graph framework can readily represent relevant background knowledge, including layout information, alongside historical ratings in a joint graph. In an extensive set of experiments, we have showed that graph-based recommendation using the Personalized PageRank measure significantly outperforms a set of popular collaborative and content-based recommendation methods. In fact, the graph approach is the only one to outperform the strong one-fit-all popularity-based recommendation method.

We find that there are several main reasons for the superiority of the graph-based approach. First, the graph models collaborative ratings together with location information and content aspects, whereas the alternative methods only account for some subset of these 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 substantially affects performance. Finally, the random walk process favors highly connected nodes, thus implicitly modelling a bias towards popular items.

This study corroborates previous findings about the superiority of the graph approach in integrating multiple sources of information. This work may be first to represent structured relational background knowledge together with user ratings using this framework. While we consider the museums domain, we believe that incorporating background knowledge using the graph-based approach can be beneficial also in other domains, especially those that operate in cold start settings.

There are several directions for future work that we would like to pursue. The limited textual content modeled in this work may be enriched using the Web (Grieser, Baldwin, Bohnert, & Sonenberg, 2011) or linguistic resources (Bohnert & Zukerman, 2014) to support better inter-item similarity assessment. An advantage of the graph framework is its flexibility, enabling multiple forms of recommendation. In addition to recommending *multimedia presentation* items, one can use the random walk algorithm for advising the visitor on the next *position* to visit. These predictions can be further organized into path recommendations.

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 Kknowledge and Data Engineering (TKDE), 17*(6).

Adomavicius, G., & Tuzhilin, A. (2011). Context-aware recommender system. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender systems handbook.* 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*(1-2).

Bagci, H., & Karagoz, P. (2015). Context-aware location recommendation by using a random walk-based approach. *Knowledge and Information Systems*. doi: DOI10.1007/s10115-015-0857-0

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*.

Berkovsky, S., Heckmann, D., & Kuflik, T. (2009). *Addressing challenges of ubiquitous user modeling: Between mediation and semantic integration*. Springer Berlin Heidelberg.

Biran, A., Poria, Y., & Oren, G. (2011). Sought experiences at (dark) heritage sites. *Annals of Tourism Research*, *38*(3).

Bohnert, F., & Zukerman, I. (2014). Personalised viewing-time prediction in museums. *User Modeling and User-Adapted Interaction*, *24*(4).

Bohnert, F., Zukerman, I., & Laures, J. (2012). *Geckommender: Personalised theme and tour recommendations for museums*. Berlin Heidelberg, Springer-Verlag.

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).*

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, *12*(4).

Craswell, N., & Szummer, M. (2007). Random walks on the click graph. In *Proceedings of the international acm sigir conference on research and development in information retrieval (sigir).*

Davey, G. (2005). What is museum fatigue? *Visitor Studies Today*, *8*(3).

Desrosiers, C., & Karypis, G. (2011). A comprehensive survey of neighborhood-based recommendation methods. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender systems handbook.* Springer.

Dim, E., & Kuflik, T. (2014). Automatic detection of social behavior of museum visitor pairs. *ACM Transactions on Interactive Intelligent Systems (TIIS)*, *4*(4).

Falk, J. H. (2009). *Identity and the museum visitor experience*. Left Coast Press.

Fogaras, D., Rácz, B., Csalogány, K., & Sarlós, T. (2005). Towards scaling fully personalized pagerank: Algorithms, lower bounds, and experiments. *Internet Mathematics*, *2*(3).

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*(3).

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*(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*(1), 10-18.

Haveliwala, T. H. (2002). Topic-sensitive PageRank. In *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 *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*(8).

Kuflik, T., Kay, J., & Kummerfeld, B. (2012). *Challenges and solutions of ubiquitous user modeling*. Berlin Heidelberg, Springer-Verlag.

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*(3).

Kuflik, T., Wecker, A. J., joel Lanir, & Stock, O. (2014). An integrative framework for extending the boundaries of the museum visit experience: linking the pre, during and post visit phases. *Information Tehnology and Tourism)*, *15*(1).

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*.

Lao, N., & Cohen, W. W. (2010). Relational retrieval using a combination of path-constrained random walks. *Machine Learning*, *81*(1).

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.* 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*(8).

Lucchese, C., Perego, R., Silvestri, F., Vahabi, H., & Venturini, R. (2012). *How random walks can help tourism*. Springer.

Minkov, E., & Cohen, W. W. (2011). Improving graph-walk based similarity with reranking: Case studies for personal information management. *ACM Transactions on Information Systems (TOIS)*, *29*(1).

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 trust (socialcom-passat).*

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.*

Panniello, U., Tuzhilin, A., & Gorgoglione, M. (2014). Comparing context-aware recommender systems in terms of accuracy and diversity. *User Modeling and User-Adapted Interaction*, *24*.

Pazzani, M., & Billsus, D. (1997). Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, *27*(4).

Pham, T.-A. 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 *NIPS.*

Salton, G. (Ed.). (1971). *Relevance feeback in information retrieval*. Englewood, Cliffs, New Jersey: 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.*

Snijders, M. (2014). *Buying art at the museum* (Unpublished master's thesis).

Erasmus School of History, Culture and Communications, Erasmus University, Amsterdam, Netherlands.

Stock, O., Zancanaro, M., Busetta, P., Callaway, C., Krüger, A., Kruppa, M., ... Rocchi, C. (2007). Adaptive, intelligent presentation of information for the museum visitor in peach. *User Modeling and User-Adapted Interaction, 17*(3).

Tiroshi, A., Berkovsky, S., Kaafar, M. A., Chen, T., & Kuflik, T. (2013). Cross social networks interests predictions based on graph features. In *Proceedings of acm conference on recommender systems (recsys).*

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 de?nition 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).*

Toutanova, K., Manning, C. D., & Ng, A. Y. (2004). Learning random walk models for inducing word dependency distributions. In *Proceedings of the international conference on machine learning (ICML).*

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*(5).

Yildirim, H., & Krishnamoorthy, M. S. (2008). A random walk method for alleviating the sparsity problem in collaborative filtering. In *Proceedings of the acm conference on recommender systems (RecSys).*