Information Processing and Management 53 (2017) 1269–1286

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Overlaying social information: The effects on users' search and information-selection behavior

Valeria Orso[a,*], Tuukka Ruotsalo[b], Jukka Leino[b], Luciano Gamberini[a,c], Giulio Jacucci[b]

[a] Department of General Psychology, University of Padova, via Venezia 8, Padova, Italy
[b] Helsinki Institute for Information Technology HIIT, Department of Computer Science University of Helsinki, Gustaf Hällströmin katu 2B, Helsinki, Finland
[c] Human Inspired Technologies Research Centre, University of Padova, via Venezia 8, Padova, Italy

ARTICLE INFO

Article history:
Received 10 November 2016
Revised 16 June 2017
Accepted 16 June 2017
Available online 28 July 2017

Keywords:
Social search
Information retrieval
Personalization

ABSTRACT

Previous research investigated how to leverage the new type of social data available on the web, e.g., tags, ratings and reviews, in recommending and personalizing information. However, previous works mainly focused on predicting ratings using collaborative filtering or quantifying personalized ranking quality in simulations. As a consequence, the effect of social information in user's information search and information-selection behavior remains elusive. The objective of our research is to investigate the effects of social information on users' interactive search and information-selection behavior. We present a computational method and a system implementation combining different graph overlays: social, personal and search-time user input that are visualized for the user to support interactive information search. We report on a controlled laboratory experiment, in which 24 users performed search tasks using three system variants with different graphs as overlays composed from the largest publicly available social content and review data from Yelp: personal preferences, tags combined with personal preferences, and tags and social ratings combined with personal preferences. Data comprising search logs, questionnaires, simulations, and eye-tracking recordings show that: 1) the search effectiveness is improved by using and visualizing the social rating information and the personal preference information as compared to content-based ranking. 2) The need to consult external information before selecting information is reduced by the presentation of the effects of different overlays on the search results. Search effectiveness improvements can be attributed to the use of social rating and personal preference overlays, which was also confirmed in a follow-up simulation study. With the proposed method we demonstrate that social information can be incorporated to the interactive search process by overlaying graphs representing different information sources. We show that the combination of social rating information and personal preference information improves search effectiveness and reduce the need to consult external information. Our method and findings can inform the design of interactive search systems that leverage the information available on the social web.

© 2017 The Authors. Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license.
(http://creativecommons.org/licenses/by-nc-nd/4.0/)

* Corresponding author.
E-mail address: valeria.orso@unipd.it (V. Orso).

http://dx.doi.org/10.1016/j.ipm.2017.06.001
0306-4573/© 2017 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license.
(http://creativecommons.org/licenses/by-nc-nd/4.0/)

1. Introduction

The era of social computing has kindled massive amounts of new types of data on the web, including tags, ratings, and reviews of a variety of content. A wealth of work before the rise of social review and rating sites has investigated ways to utilize social data in recommending and personalizing information. The early approaches successfully used social data in recommender systems by predicting ratings via collaborative filtering (Breese et al., 1998; Goldberg et al., 1992; Konstas et al., 2009; Resnick et al., 1994) and utilizing social information to enhance Web search ranking (Agrawal et al., 2015; Amitay et al., 2009; Carmel et al., 2009). Previous research has also demonstrated that personalization using user's behavioral data can lead to improved search effectiveness (Pitkow et al., 2002; Teevan et al., 2005) and that social information can be successfully incorporated in retrieval methods (Bao et al., 2007; Guy et al., 2010).

The motivation for further investigating approaches in this direction is twofold. Firstly, *blurring of search and recommendation* harnessing the variety of social data, such as ratings, tags, reviews, search time queries, and personal preferences, calls for new approaches to investigate how different social and content information can be simultaneously incorporated in an online information-seeking process to empower information access.

Secondly, *user benefits in information selection* quantifying the effect of social data for user behavior and utility in selecting information to enhance decision-making requires empirical approaches that extend beyond simulating rating predictions or rankings to reveal the behavioral and experimental benefits for the users.

Conversely, recommender systems or personalization research focuses on predicting static ratings or a ranking for content items (Konstas et al., 2009; Teevan et al., 2005). Instead, we propose studying how social information embedded in ranking and result presentation affects users' behavior when making decisions to select useful information.

For example, consider the case of a user looking for a place to have dinner in an unfamiliar city and seeking advice from the social web. The user may have search content preferences that he or she expresses as a query, e.g., "Asian restaurants". The user would also be likely to choose a restaurant matching her long-term personal preferences, say "Thai food," rather than relying solely on the query "Asian restaurant". At the same time, the user would be likely to rely on the common opinion of other like-minded users in the social web on the quality of a particular restaurant offering Asian food.

The scenario exemplifies different layers of social search and personalization that could be potentially relevant for a typical search task: search time content preferences (Asian restaurant), personal content preferences (Thai food), and social preferences (opinions of other users who are likeminded and have preferences regarding Asian restaurants, Thai food, or similar restaurants).

To make the decision about where to have dinner, the user could benefit from a ranking that takes into account all of this information. Moreover, the user could need information on why and how the restaurants suggested to him/her were ranked: based on their match to the query or her long-term preferences, the ratings that other like-minded people had given for the restaurants, or a combination of these.

To this end, we investigate user search behavior using socially created information layers in presence of various data: content tags, social ratings, personal preferences, and search time queries.

1.1. Research objective

The main objective of the present study is to investigate whether the combined deployment of different layers of social information affects the users' search and information selection behavior. More specifically, we aim at identifying which of the information layers can improve the search effectiveness, thus benefiting the user's information-selection behavior. In addition, we aim at testing the effectiveness of the proposed computational method in an interactive online setting.

1.2. Contributions

The contributions of this article are as follows:

- *Empirical evidence of the benefits of social overlay information for search effectiveness and information-selection behavior*. The effects are demonstrated by data comprising logs, eye-tracking recordings, simulation data, and questionnaires from a task-based user study in which 24 participants planned their activities in a point-of-interest search and recommendation scenario on the largest publicly available social web dataset from Yelp (Blomo et al., 2013) consisting of hundreds of thousands of ratings, tags, and items.
- *Computational method utilizing overlaid social graphs*. This method combines social, personal, and search-time user input. Such pieces of information are modeled as overlaid graphs, and relevance is computed by computing random walks with restarts on the overlaid graph. User's search-time preferences, such as queries or other feedback, are incorporated as prior information to the restart computation and arbitrary amount of overlays can be processed in the same graph computation framework.

2. Background

Our work builds on several areas of related work reviewed below, including search personalization, social data usage for information retrieval, recommender systems, and social personalization.

2.1. Personalized search

Personalized search traditionally tailors search results to an individual user's interests by incorporating information about the user beyond the query the user provides (Pitkow et al., 2002; Ruotsalo et al., 2014a; Teevan et al., 2005; Teevan et al., 2010). Information used for personalization varies from users' personal content preferences (Pitkow et al., 2002; Ruotsalo et al., 2014a) to social computing, which often relies on ratings, or social annotation, such as tags, ratings, or links to other users' profiles in the social web (Chi, 2009; Guy et al., 2010; Muralidharan et al., 2012). Services utilizing social annotation are becoming popular and are commonly available on many review and e-commerce services, such as Yelp,[1] TripAdvisor,[2] or Amazon.[3]

Other methods include interactive approaches leveraging adaptive navigation and user communities (Smyth, 2007) to build a relevance model that guides the promotion of community-relevant results and using task features, topic knowledge, and task products (Liu, 2009; Luxenburger et al., 2008).

2.2. Social data in information retrieval

We build on the concept of social search, which describes the search process over social data gathered from the social web applications, such as social bookmarking systems, review sites, and services that store users' preferences toward different entity types (items, other individuals, tags) and their interrelations, and allows personalizing search for the general Web as well as local search (Carmel et al., 2009; Gasparetti, 2016).

Existing research on using social data in information retrieval includes approaches purely based on tags, sometimes called folksonomies (Bouadjenek et al., 2013a; Hotho et al., 2006; Maniu & Cautis, 2013; Peltonen et al., 2017); more structured semantic annotation (Blanco et al., 2011); a combination of tags and people (Guy et al., 2010); and more general social media data across services (Carmel et al., 2009; Kahveci et al., 2016; Kawase et al., 2014). Early approaches of using social data to enhance information retrieval are the Adapted PageRank methods, in particular SocialSimRank, SocialPageRank (Bao et al., 2007), and Topic-Driven SocialRank (Kim & Park, 2013) which are used to compute similarities between users and their social networks to enhance the web searches. Similar method called ScenticRank that makes use of sentiment information, in addition to social graphs, have also been proposed (Xie et al., 2016).

Even simple social signals have been shown to improve Web search. For example, user interactions on digital content, such as likes or shares of a document, have been shown to improve the utility of searchable content (Pantel et al., 2012).

Empirical research also shows that information mined from social media services can lead to quite distinct results from those produced by the major Web search engines (Agrawal et al., 2015) and that personalization of search results can be enhanced using social networks (Bouadjenek et al., 2016; Bouadjenek et al., 2013b; Shafiq et al., 2015).

Recent research has also proposed to model social data as a composition of two networks, one consisting of information in web pages and the other of personal data shared on social media web sites. These are then used to analyze how social media tunnels the flow of information from person to person and how to use the structure of the social network to rank, deliver, and organize information specifically for each individual user.

This line of research comes closest to our computational approach, but we extend this by computing relevance in several graph overlays simultaneously (content, social, personal), as opposite to using social signals simply as a prior for document relevance. Thus, we are able to use several types of social data for relevance estimation and our model allows this computation to be performed online without the need to pre-compute similarities, thus supporting interactive search.

A major application area for social search has been Mobile local search (Gasparetti, 2016), which we also use as a scenario in our work. It is inherently different from general Web search as it focuses on local businesses and points of interest instead of general web pages, and finds relevant search results by evaluating different ranking features. Social signals have been found useful in local search to complement often short and categorical queries to learn improved ranking models.

User interface aspects have also been found to affect how users benefit from social information in search and information consumption context (Muralidharan et al., 2012; Ruotsalo et al., 2016). Improvements to the user interface design, such as exposing names and faces to end users within the search result listing, are shown to be required to make them more noticeable and useful (Fernquist & Chi, 2013). While previous research has concentrated in ad-hoc design features (Fernquist & Chi, 2013; Muralidharan et al., 2012), we study the effect of different graphs to an end-user evaluation in which the effect of different graph overlays are visualized for participants when they are performing realistic search tasks.

[1] http://www.yelp.com/.
[2] http://www.tripadvisor.com.
[3] http://www.amazon.com/.

2.3. Recommender systems

The target of recommender systems is closely related to our approach; to retrieve relevant information to the user based on historical user behavior. Recommender systems are based on either the content of the recommended items or the preferences of like-minded users. The most prominent approach utilized in recommender systems is collaborative filtering where the information provided by other users is used to predict the preferences of an individual user (Schafer et al., 2007; Ye et al., 2010; Ye et al., 2011; Yuan et al., 2013).

Recommender systems have also been built using data, such as explicit ratings or buying behavior of users. All of these approaches are based on batch processing of the data and they are not interactive. In recent years, there has been an increasing amount of research on interactive recommendation systems that are becoming more search-like by offering capabilities that enable users to direct their search instantly based on a recommendation functionality integrated with the search user interface (Glowacka et al., 2013; Klouche et al., 2017; Ruotsalo et al., 2014a; Ruotsalo et al., 2013). However, these approaches are typically based on data collected on a user's personal activities during the search session rather than social data.

Recent advances on tensor factorization have enabled multi-source data to be used in the collaborative filtering framework (Ng et al., 2011). Researchers have exploited different social signals, such as tags (Rendle & Schmidt-Thieme, 2010) and ratings (Karatzoglou et al., 2010). These approaches are based on estimating ratings for unseen items as in typical collaborative filtering scenarios. Our approach differs from these studies in two ways. First, we study the effect of the resulting recommendations with respect to user behavior as opposite to just estimating rating error. Second, our approach computes relevance estimates in real-time as opposite to the heavy low-rank tensor factorization methods. However, the lower-rank tensor factorization methods could be used as a basis for our ranking approach in scenarios where the data is sparse and missing information needs to be estimated using different data overlays, but the target is to provide the users with a real-time retrieval system that blurs search and recommendation functionalities.

2.4. Social personalization

Approaches for employing users' social-network data for personalization have also been proposed. A method presented in Noll and Meinel (2007) is based on users' tag profiles that are matched against the retrieved information. The problem of automatic query expansion using social data as part of the process is addressed in Biancalana et al. (2013).

Recent research has highlighted the interactive nature of information access behavior and promoted the potential value of harnessing user-activity patterns in social-information access tools. The proliferation of user-curated content in large volumes from the rapidly growing social networks has enabled new types of information-access services for users, including visual-search systems (Ahn & Brusilovsky, 2009; Ahn & Brusilovsky, 2013; Ahn et al., 2008; Freyne et al., 2007; Ruotsalo et al., 2014b) and social browsers (Church et al., 2010; Kammerer et al., 2009).

Researchers have also shown how social personalization can leverage the browsing behavior of past users to guide others to interesting and relevant information (Wexelblat & Maes, 1999) or use the search patterns (queries and selections) of users and their communities when responding to future searches to adapt a result list suited to the needs of a particular community (Smyth et al., 2005).

Despite the growing interest in personalization and social data, the main focus and corresponding evaluation criteria have been to maximize the accuracy of prediction for ratings of unseen users, as in collaborative filtering (Carmel et al., 2009; Guy et al., 2010), rather than studying user behavior, effectiveness, and information-selection behavior in completing real information-seeking tasks.

While these results are highly important and demonstrate the value of the techniques, there is a growing awareness that considering accuracy alone may be misleading (Knijnenburg et al., 2012; McNee et al., 2006). Studies report no direct link between the accuracy of the items suggested and the perceived quality or satisfaction from the user's side.

Our technique is based on three counterparts: data model that defines the representation of multiple data sources (content, social, and personal) as a set of overlaid graphs, the relevance-estimation model that performs random walks with restarts on the graph overlays and computes a relevance score for information items, and the user-interface design that transparently visualizes the effects of the different overlays for the user.

2.5. Data model

The data are described with four types of features: tags or other textual content, social ratings, personal ratings, and search-time textual queries. These data are modeled as graph overlays and compose a knowledge graph, as illustrated in Fig. 1.

Formally, the data model is an undirected and labeled graph $G_o = (N_o, P_o)$ composed of a set of subgraphs called overlays $o \in O$. The set of vertices N_o in each overlay o is a disjointed union of vertices of the particular type of data: 1) tags or users and 2) searchable items. P_o contains an undirected edge $\langle n, i \rangle$ if there is a link between node n and item i in overlay o. In other words, each overlay graph always consist of vertices between item i and node n representing the type of data encoded in the overlay, tags or users in our case.

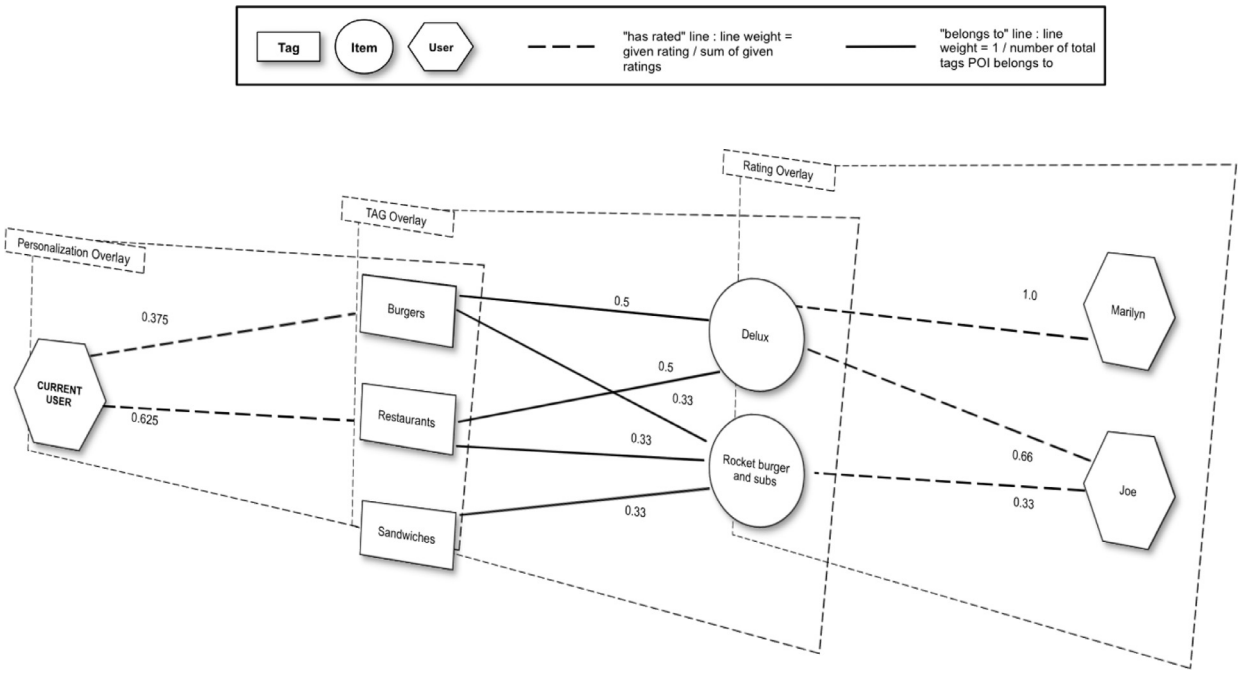

Fig. 1. Social web data are modeled as a set of overlaid graphs composed of social tags, social ratings, users' personal preferences, and search-time preferences observed from the interaction with the system. Each of these data are connected to the retrievable items on a graph overlay.

3. Overlaying social information

3.1. Transition probability estimation

We create transition probability matrices for each overlay separately. The tag overlay encodes the tags associated with each of the retrievable items. The transition probability between each tag is computed according to a share of the number of tags associated with an item.

The social-rating overlay encodes the rating an individual user provides for an item. The transition probability of the social layer is computed as a rating weight normalized by the share of the ratings that each user has given for the items.

In both cases, this procedure equates to estimation of probability that the overlay node (a user or a tag) would generate the item and can be estimated using a maximum likelihood estimation for text documents (Ponte & Croft, 1998). This provides transition probabilities $w \in [0, 1]$ between the nodes stored in a transition probability matrix A_o for each overlay o separately.

3.2. User preferences as prior probabilities

Preference can be observed either from query time as an input from the user or from a user's personal profile. In the case of a user profile, the weighting is conducted according to users' ratings. In the case of search-time preferences expressed as a query, the weight is uniformly assigned for the tag nodes that match the query. In our case, user profiles are initiated by the user with the interface shown in Fig. 4, and search-time preferences (queries) are mapped to the tags when the user types in a query (in our implementation via a query auto-completion mechanism).

These are modeled as a preference vector q, where $|q| = 1$ and $q(n)$ denotes the amount of preference for a node n. For example, a user who has given only query time preferences for tag Q, would have a uniformly distributed preference vector q where each query tag n would have an equal prior probability, such that $u(q) = \frac{1}{|Q|}$.

In case of preferences from both query and personal profile, we weight the input query and the preference vector equally. Formally, the input query entered by the user has the weight 0.5 and each q has the weight $0.5 \cdot \frac{1}{|Q|}$.

3.3. Relevance estimation

The relevance estimation is based on simultaneously estimating relevance from a set of overlaid graphs, i.e., tags, ratings, and personal preferences.

The estimation computes random walks with restarts on graph overlays using a modified version of the personalized PageRank method (Jeh & Widom, 2003) in which each overlay is computed at each iteration and affects the estimates of the next iteration. Formally, we consider the graph G to be undirected.

Computing the relevance vector v for a given preference node q can then be formalized as follows. Let a vertex be denoted as r, and by $in(r)$ and $out(r)$ denote the set of in-neighbors and out-neighbors of r in G_o, respectively. Let A_o be the transition probability matrix corresponding to the graph G for an overlay o, where

$$A_{o_{ij}} = \frac{1}{\left|out_{ij} \cup in_{ij}\right|}, \tag{1}$$

where vertex i links to vertex j or vice versa in a specific graph overlay o, and $A_{oi\,j} = 0$ otherwise. For a given user's preferences q, the relevance estimation equation that uses several overlays $o \in O$ can be written as

$$v = \frac{|O|}{1 \le o \le |O|} \frac{1}{|O|}(1-c)A_o v + cq, \tag{2}$$

where c is the restart probability. We set $c = 0.5$. The solution v is a steady-state distribution of random surfers, in which a surfer teleports at each step to vertex r with probability $c \cdot q(r)$, or restarts from node q with probability 0.5.

3.4. Negative and positive ratings

The random walk model operates on the adjacency matrices that define the transition probabilities between the nodes in the graph overlays. We consider two principles to reflect the importance of the relations between the nodes. First, the low and high ratings have different direction of opinion. The users who rated the item low are expressing a negative preference towards the item and users who rated the item high are expressing a positive preference. Second, the popularity of an opinion is a signal of trustworthiness of the ratings. More ratings should convey stronger signal, but more low ratings should not lead to high estimated relevance value. For example, an item with large amount of low ratings should not be recommended over an item with less, but more positive ratings.

To this end, the pure random walk model considers all evidence positive. For example, a large amount of low ratings would accumulate to a high estimated relevance. In order to reflect our principles to cope with the positive and negative ratings, we construct two different graphs for each overlay: one encoding the negative opinions (ratings between the values 1 and 3) and another one encoding the positive opinions (ratings between the values 3 and 5). In this way, the model accounts for both the popularity and the direction of the opinion.

We compute the steady distribution by using the power-iteration method with 50 iterations for both graphs and rank the items by the mean of the positive and negative values in the corresponding vector v. The intuition is that if a user gives a rating between 1 and 2, it is considered a negative opinion and should indicate less interest than no rating at all or a missing rating.

3.5. User-interface design

The user interface of the full system is shown in Fig. 2. The left part (A) is a conventional search-entry field that uses auto-completion to suggest the tags available in the tag graph. Several tags can be selected to be part of a query simultaneously, and they are shown in a list above the query field. The middle of the screen (B upper part) displays the result listing of items that the system retrieves in response to the user's query. The result listing was designed to communicate how the different graph overlays affect the ranking of the items. Two bars are positioned under each result item. The green bar indicates the effect of the tag overlay on the ranking. The blue bar indicates the effect of the social-rating overlay. The bars were designed to be prominent enough so that users would benefit from the transparency of the effect of the graph overlays in real interactive relevance-assessment situations. This was important because it has been shown that due to specialized attention patterns that users exhibit while processing search result pages, users mainly pay attention to titles and then turn to snippets and annotations for further evidence of a good result to click on. Moreover, the reading of snippets and annotations appears to follow a traditional top-to-bottom reading order, and any additional information that is not clearly presented simply blends into other information associated with the search result and is not recognizable for rapid relevance judgement (Muralidharan et al., 2012). If the user clicks on a particular item, additional information will be displayed in the right part of the interface (C). Here, the user is provided with the name of the item, the total number of reviews received by the place, the average rating, the tags associated with the item, and the reviews. The user can store the items in an itinerary (B, lower area on the screen) used for logging and collecting users' selections during the experiments.

3.6. An example

To illustrate the functionality of the relevance estimation, transparent visualization, and other features of the user interface, we go through a concrete example scenario in which a user is attending a conference in Phoenix and seeks to find a good restaurant for dinner.

Fig. 2. The screen capture of the user interface of the system. The interface is divided into three sections. Section A is the search section in which the user performs autocomplete-aided text queries to convey his or her interests. Section B shows the result listing in which a ranked list of results is displayed. The influence of the tag overlay is visualized in section B as a green bar (upper) under each result item, and the influence of the social-rating overlay is visualized as a blue bar (lower). The bookmark list to which the user can add preferred items is shown on the bottom part of section B. Information section C shows additional information about an item and is activated by clicking an item in the result listing.

The user has initialized her profile, as partially shown in Fig. 4. The profile consists of preferences toward Cajun/Creole, barbeque, Asian fusion, burgers, breakfast, and Brazilian.

The user then begins a search by typing "restaurants" in the search field (Fig. 2, section A) and receives a list of recommended restaurants as a response (Fig. 2, section B). The system ranks high restaurants that serve gourmet burgers, American-style breakfast, and barbeque, and particularly those that are appreciated by the likeminded users in the social-rating graph overlay. This is visualized with high contribution to the rank score by the content preferences (green bar) and, in particular, with the social preferences (blue bar) in section B of Fig. 2. The example illustrates the search method's capability to combine the user's personal preferences, social ratings, and search-time criteria, which does not have to be exact as the personalization effect will find suitable restaurants for the user by using different data overlays even when the query is very general, such as "restaurants".

After receiving the recommendation list, the user can consult detailed information about each result by clicking on the result list. The information is displayed on the right side of the screen in section C of Fig. 2. The user can further select the preferred item by clicking the "plus" button in the result list, and the item then is added to the bookmark list shown in section C of Fig. 2.

4. User study

The purpose of the user study was to investigate the effect of the combination of different data overlays, i.e., tag, social-rating, and personalization data for the effectiveness of search and information-selection behavior. To this end, the focus was to quantify benefits of social information via behavioral measures: frequency to consult additional information when making decisions, ranking of selected information, the effect of different social-information overlays in producing those rankings, and task-execution time.

4.1. Experimental design

The experiment followed a within-subjects design, in order to limit the potential confounding effects due to individual differences in background and expertise. The independent variable of the experiment was the system configuration corresponding to the different data-overlay combinations. Each participant performed a total of twelve search tasks, four with every different version of the system, i.e., three. The order of presentation of the data-overlay combinations and the order of the search tasks were counterbalanced.

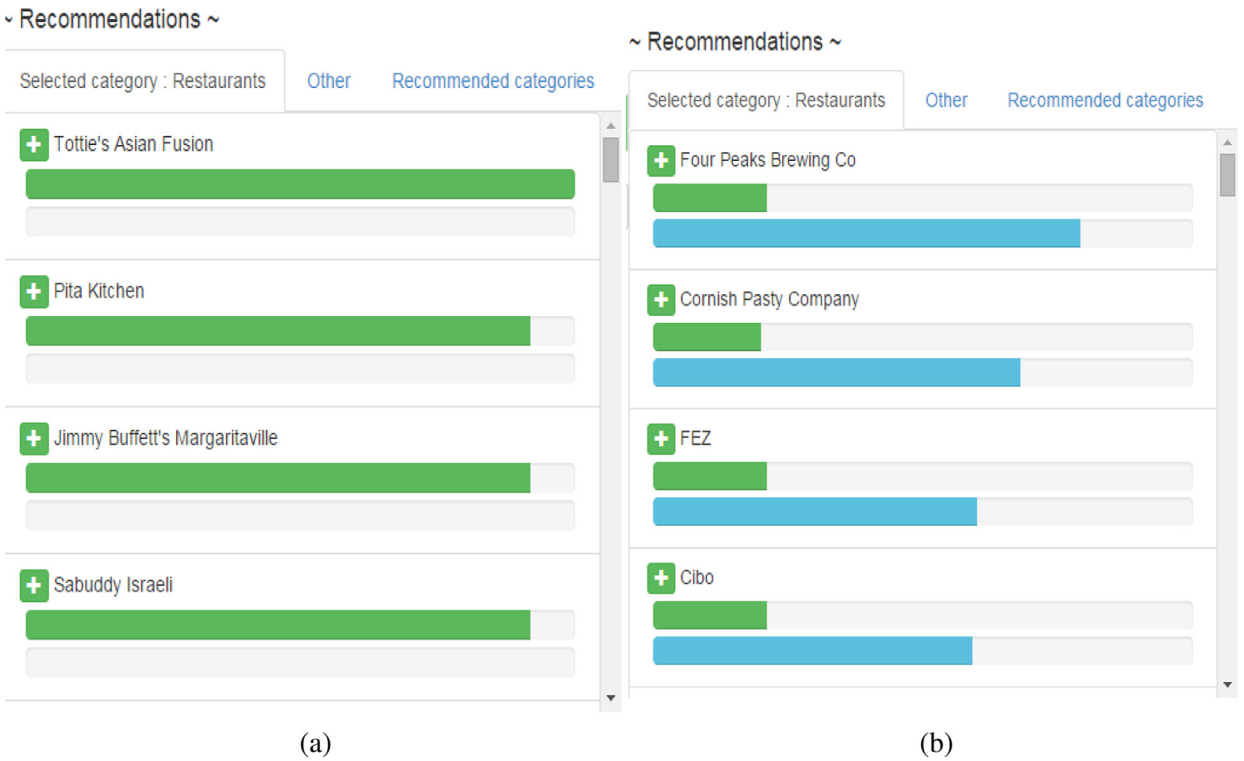

Fig. 3. Screen captures from the baseline system interfaces used in the experiments. The system with only the tag overlay (a) ranks and transparently visualizes the results only using the tag overlay as a green bar (upper) under every result item. The system with both tag and social-rating overlays (b) ranks and visualizes the results using both the tag overlay (green bar, upper) and the social-rating overlay (blue bar, lower) shown under every result item but not the user's personal preferences.

4.2. System configurations

We configured three systems with the following combinations of data overlays: (1) tags, (2) tags and social ratings, and (3) tags, social ratings, and user profile. The first configuration pertains to only the tag overlay (Fig. 1) and represents a conventional search system that matches the search query with the content description, tags in this case. The system includes an auto-completion support to find relevant tags that could be used as queries. The second baseline system combines the tag overlay and the social-ratings overlay (Fig. 1). These configurations were composed by adding the social rating layer on top of the tag layer. The full system used all data overlays. Essentially, for the full system, the user profile overlay was added over the tags and social-rating layers (Fig. 3). Because the focus of the experiment was to assess the effect of social data on end users' searching behavior, a combination of graph overlays that did not comprise social ratings was not included.

The interface transparently visualizes the influence of data overlays on the ranking. The effect of the tag overlay is represented by the length of the green bar, whereas the impact of the social-ratings overlay is represented by the length of the blue bar. Previous research has shown that the interface layout in a recommendation engine impacts users' behavior. Therefore, the appearance of the interface was maintained as the same for all of the conditions. Consequently, the blue bar was inactive in the system configuration in which only the tag layer was used, as no social rating information was available. In addition, the selection of the three combinations of graph overlays relevant to our purposes allowed us to keep the study feasible in the frame of a within-subjects design.

The system configurations are further referred to with the following acronyms. The system with only the tag overlay is referred to as T, the system with the tag and social-rating overlay is referred to as TR and the full system with tag, rating, and user profile overlays is referred to as TRP.

4.3. Data

We used the largest publicly available real-world dataset containing social ratings that was originally released by Yelp for the RecSys Yelp Rating Prediction Challenge in 2013.[4] The dataset contains information for a total of 11,537 local businesses in the Phoenix, AZ, metropolitan area. They are associated with 229,907 reviews and ratings from 43,873 users. The items

[4] https://www.kaggle.com/c/yelp-recsys-2013.

are also associated with 506 different socially created tags, such as restaurants, museums, airports, golf courses, and so on. The dataset gives a comprehensive representation of the local businesses and services in the Phoenix area.

4.4. Tasks

The participants were situated in a simulated work task with the following task scenario:

"You are attending a conference in Phoenix, and you are unfamiliar with the city. Because you don't know the city, you are looking for points of interest using a search engine."

We selected four search tasks corresponding to four different subcategories: (1) a place to have a meal, (2) a place for a cultural activity, (3) a place to get a haircut, and (4) a place to spend a night out. In other words, users searched within the subcategories of restaurants, arts, beauty and spa, and nightlife, respectively.

4.5. Participants

We recruited 24 participants to take part in the study. Seven of them were females. Their ages ranged from 20 to 47, with a mean age of 27.25 years ($SD = 5.7$). Because the text presented in the user interface was in English, only participants with a self-reported good knowledge of English were eligible to take part in the experiment. Participants were told that they could ask the experimenter for clarification at any time during the experiment. When asked whether they had previous experience with recommendation engines, all of the participants reported to have limited experience with interactive search engines, and none had experience with the system or data. All participants were university students or university employees and had a background in cognitive and psychological sciences. The lack of participants' familiarity with interactive search engine, may be accounted by their non-technical background. They were recruited by word of mouth and received no compensation for their participation in the experiment.

4.6. Apparatus and data logging

The experiment was run on a desktop PC connected to a 21-inch monitor. Participants' eye movements were recorded using a Tobii X120 eye-tracker. During the experiment, participants could use a mouse and a keyboard to operate the interface. Lighting conditions were kept constant for all participants. The search engine automatically logged the timestamp, the action performed by the user (i.e., the keyword typed in the search box and button pressed), the list of the recommended places in sequential order, and the item chosen by the user. Participants' subjective evaluations of the systems were recorded using a post-use questionnaire presented on the PC.

4.7. Procedure

The participant was first debriefed on the experimental procedure and the purpose of the study. Then, s/he gave informed consent to take part in the experiment. S/he was also told that s/he was free to withdraw from the experiment at any moment with no consequence. The participant then filled in a questionnaire collecting demographics and background information regarding his/her previous experience with interactive search engines.

Next, the participant was introduced to the interface and asked to build an initial personal profile by rating a list of tags (see Fig. 4). This approach was chosen to be a proxy for direct item rating, as participants in our study were not familiar with the items in the dataset and thus were unable to provide real ratings. This choice was further motivated by previous findings, according to which a very long profile-generating process may reduce the participant's willingness to create a profile. Targeting the profile construction with a selected set of representative tags instead of the actual items helps to avoid this effect (Cremonesi et al., 2012).

The experimenter then illustrated the search-interface functionalities and explained how to operate the interface. The participant was asked to make two searches to acclimatize to the system. If the participant agreed that the system functioning was clear, the actual experimental session began.

The participant was asked to use the system to search within a certain subcategory (the task topic) and then to select the item that best matched his/her interests from the result list.

After completing four search tasks, the participant was asked to fill in a questionnaire assessing his/her opinion regarding different aspects of the search engine. The eye-tracker was calibrated before the participant started a new task with a different combination of graph overlays. To make the participant feel at ease, the experimenter remained in the same room during the experiment but out of his/her sight. The participant was told that the experimenter was there in case s/he needed any clarification, but s/he was also advised not to turn his/her head to maintain the eye-tracker calibration. The entire experimental session lasted approximately 45 min.

4.8. Measures

To assess the effect of different data overlays on users' behaviors, we used both objective and subjective measures.

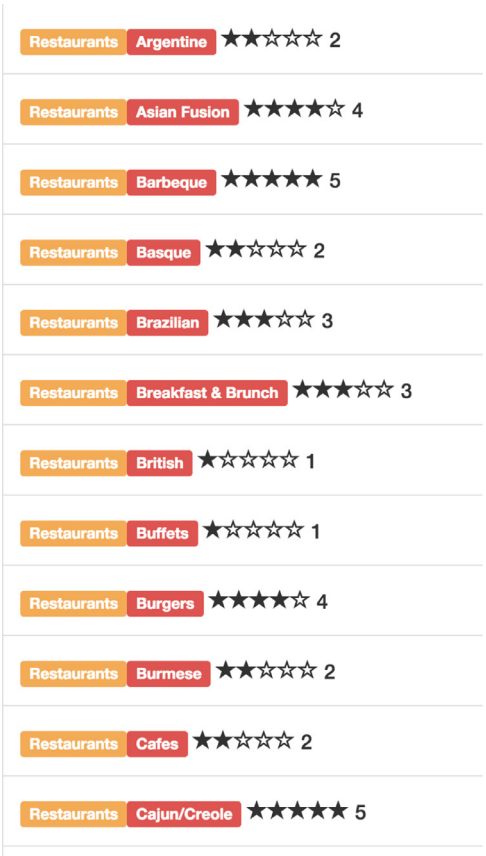

Fig. 4. A screenshot of the user interface that participants used to initialize their personal user profile. The participants rated a set of categories matching to the tasks used in the user experiment and a set of subcategories under each main category to create a user profile.

Consulting additional information. The clicked items in the list of results opened additional information describing the item. The clicks were recorded as an indicator of need to consult additional information for decision support in selecting information. The rationale was that if the participants were able to gain trust in the system by relating the items to their preferences, the click frequency would be lower, and if they needed to consult additional information to make their decision, the click frequency would be higher.

Task-execution time. The time required to complete a searching task was logged and constituted a general measure of search efficiency. However, participants were not asked to perform the task as quickly as possible, as we did not want to alter the natural information-selection behavior.

Mean reciprocal rank. The effectiveness of information selection behavior was measured via mean reciprocal rank (MRR), which is the most commonly used measure in single correct-item scenarios.

MRR was chosen because it directly measures information selection behavior when users only select the item that they find to best fit their task.

Unlike conventional effectiveness measures, such as precision, recall, or normalized cumulative gain that consider all relevant items up to a particular rank position, MRR considers only the position at which the user selects an item. This directly supports our target of studying what users are selecting as opposite to what users click or what could be generally considered relevant.

MRR is the multiplicative inverse of the rank of the first relevant result or the item that the user selected averaged across all tasks. The MRR is defined over a set of tasks T as $\frac{1}{|T|}\frac{T}{i=1}\frac{1}{rank_i}$.

Intuitively, the higher the position at which the user decided to select the item, the better the ranking method was able to estimate the relevance of the item for the participant and the better the MRR.

In addition to MRR in the tested condition, we computed the deviation between the average rank of the item selected by the user in the tested system variant and simulated the expected position that the same item would occupy in the result list of the other system variants with the same user input. In other words, given the interaction data observed in TRP, we computed the simulated MRR value in TR and T. In addition, we computed the value of the simulated MRR in T, given the interaction data observed in TR.

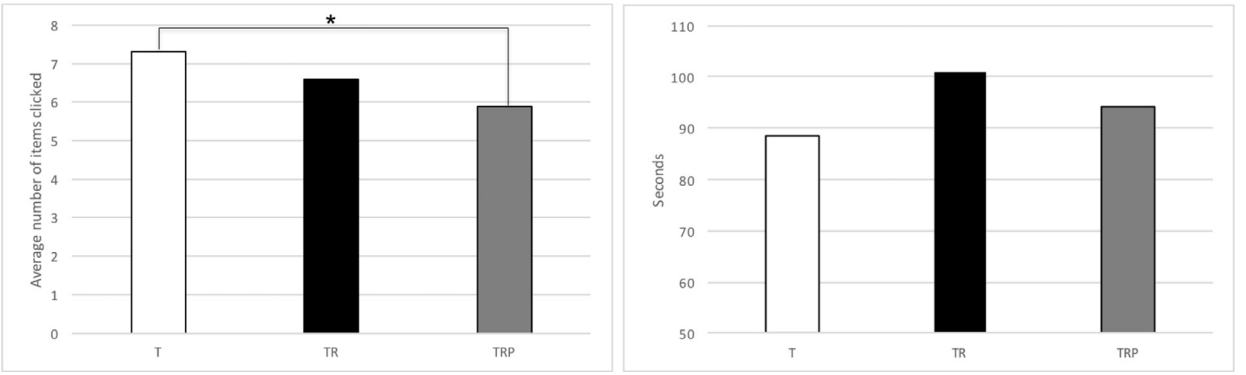

Fig. 5. Efficiency results. On the left, the average number of items clicked. On the right, the average task-execution time. Asterisks indicate a significant difference $^{*}p < .05$. Further details are provided in Table 1.

Social-rating overlay effect. The weight of the social-rating overlay for the items that the participants selected was recorded as an additional indicator of search effectiveness (only for conditions TR and TRP; condition T did not contain the rating overlay). Intuitively, the higher the social value associated with the item chosen, the more the social-rating layer influenced the ranking and the user's selection.

Fixation duration. In addition to metrics mirroring information selection and decision-making support, task-execution time, and ranking effectiveness, the total fixation duration for specific areas for the visual exploration of the interface were recorded. In particular, two areas of the interface were selected: the one displaying the result listing and the one presenting the item description. Collecting data regarding fixation-duration had a twofold aim: first, we intended to ensure that participants were effectively attending to the stimuli they were presented and did not picked the points of interest randomly. Secondly, we expect that fixation time pertaining of item description area in TR and TRP conditions would decrease, as a consequence of the better ranking of the results.

Subjective evaluation. Subjective evaluations were recorded with a post-experience questionnaire. The items of the questionnaire were adapted from the ResQue questionnaires, which characterize the quality of user experience and users' subjective attitudes toward the acceptance of a recommender technology (Pu et al., 2011).

5. Results

The results of the experiment are presented with respect to the selected measures: consulting additional information, task-execution time, MRR, expected MRR, social overlay effect, fixation duration, and questionnaires. In addition, the results of the simulation are reported.

5.1. Consulting additional information

Data pertaining to the amount of information participants consulted were distributed asymmetrically; therefore, a logarithmic transformation was applied. On the transformed data, a 3 (graph overlay) $\times$ 4 (tasks) repeated-measures ANOVA was run to assess if the frequency with which participants consulted external information differed across the three system variants. The task was also included as a factor in the analysis to check whether the repetition of the search task affected the need to consult additional sources.

The analysis revealed a main effect of the graph overlay $F(2, 21) = 4.98$ $p = .011$ $\eta_p^2 = .78$, and a main effect of the task also emerged $F(2, 21) = 2.87$ $p = .043$ $\eta_p^2 = .66$. The interaction was non-significant. To further explore the effect of the combination of graph overlays, post-hoc comparisons with Bonferroni correction were run. These comparisons showed that, on average, participants tended to consult fewer reviews in TRP ($M = .56$; $SD = .06$) compared to T ($M = 71$ $SD = .05$). Analyses were also run to explore the effect of the task; however, post-hoc comparisons run with Bonferroni correction revealed no significant difference between the tasks.

In summary, *the combination of the user profile overlay and the social rating overlay was effective in lowering the amount of external information to consult compared to the baseline, thus suggesting that the information in the result list was more sufficient when integrated with the information from the social layer.*

5.2. Task-execution time

The average task-execution time is shown in Fig. 5 (right). The time required for the participant to make a selection was compared across the system variants; the task was also included in the analysis as a factor to assess if the repetition

Table 1
Efficiency and effectiveness results for the different combinations of data overlays. The first column shows the average task durations with standard deviation values in the three combinations of graph overlays. The second column shows the average number of additional sources consulted and the standard deviation per combination of graph overlays. The third column shows the MRR of the item users chose for each combination of graph overlays. The fourth column presents the average value of the social overlay for TR and TRP and standard deviations. $^{*}p = .05$, $^{**}p < .001$.

	Task-execution time	Number of information consulted	MRR of the item chosen	Social-overlay value
	M(SD)	*M(SD)*	*M(SD)*	*M(SD)*
T	88.46(59.34)	7.31(1.01)	.3(.2)	–
TR	100.92(63.04)	6.6(1.04)	.33(.22)	.43(.47)
TRP	93.99(63.38)	**5.89(1.01)***	**.40(.20)***	**1.29(.79)****
Results		TRP < T 19.42%	TRP > T 33.33%	TRP > TR 300%

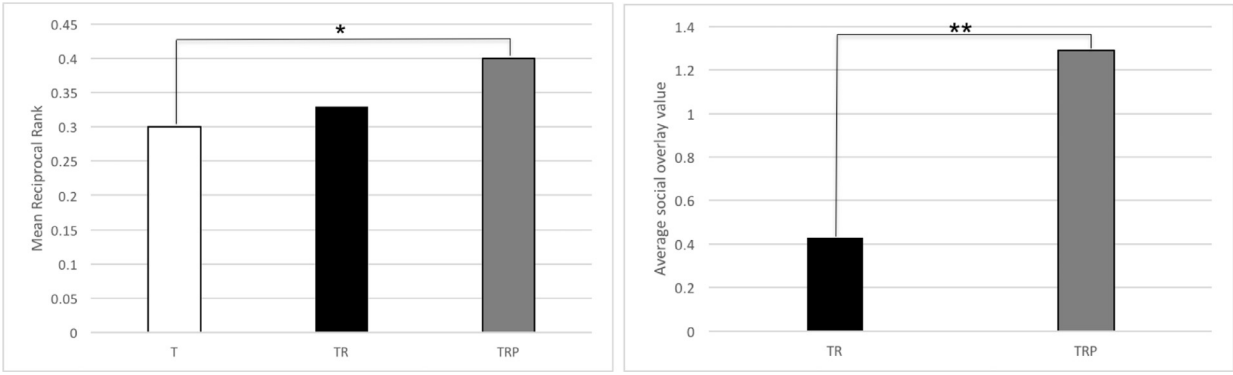

Fig. 6. Effectiveness results. On the left, the MRR for the observed items chosen by users. On the right, the average value associated with the social-information layer. Asterisks indicate a significant effect $^{*}p < .05$; $^{**}p < .001$. Further details are provided in Table 1.

of the search task affected the task duration. Thus, a 3(graph overlay) $\times$ 4(tasks) repeated-measures ANOVA was run. The analysis revealed that the combinations of the graph overlays had no effect on the time required to complete a searching task, $F(2,21) = 0.652$ $p = .52$. Additionally, no practice effect emerged $F(2,21) = 2.61$ $p = .11$, and the interaction was non-significant. The average time required to complete a searching task was 88.46 seconds ($S\ D = 59.34$) when using T; it took users on average 100.92 seconds ($S\ D = 63.04$) when using TR. Finally, when using the TRP, participants completed the searching task in 93.99 seconds ($S\ D = 63.38$) on average.

In summary, *the average task-execution time was not affected by the different combination of graph overlays.*

It is notable, however, that we did not limit the time available to complete the task, and participants were not asked to pay attention to the task-execution time. Therefore, it was assumed that no differences between the systems would emerge.

5.3. Mean reciprocal rank

The effectiveness of the different system variants was assessed by MRR, as shown in Table 1 and Fig. 6 (left).

The MRRs were compared across the three combinations of data overlays using Friedman's test. The analysis highlighted a significant effect of the combination of graph overlays, $\chi^2 = 6.038$ $p = .048$. A Wilcoxon post-hoc test showed that participants tended to choose, in TRP, items ranked higher in the list as compared to the choices they made using T, $Z = 2.11$, $M = .4$ ($S\ D = .2$) and $M = .3$ ($S\ D = .2$), respectively. No significant differences emerged comparing the mean reciprocal ranks of TRP and TR $Z = 1.14$ $p = .25$, with a MRR of .33 $S\ D = .22$ when using TR.

In summary, *the addition of the social layer showed an improvement in the ranking of the results, however only the combination of the social and personalization layers was effective in significantly improving the ranking.*

5.4. Effect of the social-data overlay

The effect of the graph overlay pertaining of the social ratings was assessed by comparing the average weights associated with the selected items in the TR and TRP. A paired sample *t*-test highlighted a significant difference between TR and TRP $t(22) = 4.41$, $p < .001$, with the mean social value in TRP ($M = 1.29$, $SD = .79$) being higher than the average social value in TR ($M = 0.43$, $SD = .47$), as shown in Fig. 6 (right). The effect size for this difference is $r = .67$. Participants tended therefore to select items more strongly affected by the evaluations previously made by other users.

In summary, *the ranking was improved in the TRP condition, to which the social-overlay computation in TRP affected significantly more than in the TR baseline condition.* Despite no differences being found in the ranking between the TR and TRP

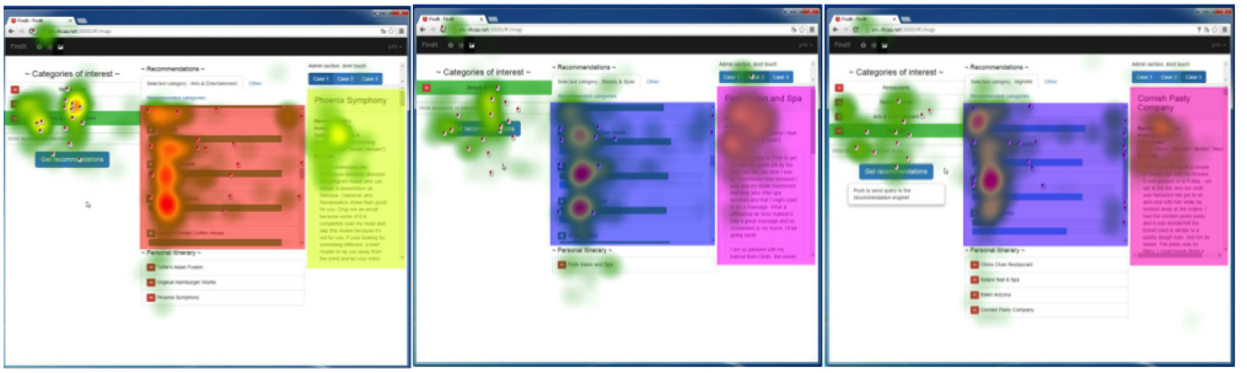

Fig. 7. Eye-tracking data from one experimental session represented as heat maps. The screenshot on the left refers to the use of T. The screenshot in the middle refers to the use of TR. The screenshot on the right refers to the use of TRP. The colored boxes represent the areas of interest (AOI) to which the analysis of the eye-tracking data was limited; the one in the middle of the screen refers to the result list, and the AOI on the left side of the screen refers to the area containing additional information. The small red and white dots represent participants' clicks. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

conditions, the results suggest that the improvements of MRR are explained by the personalization layer in combination with the social-rating overlay.

5.5. Eye-tracking

Eye-tracking recordings from three participants were excluded from the analysis due to poor calibration. A one-way ANOVA was run to assess if the combination of graph overlays affected the fixation duration in the two areas of interest considered, i.e., the search-result list and the additional-information area (Fig. 7). The analysis revealed that the combination of graph overlays had no effect on the total fixation duration on the results list $F(2, 18) = 2.034$ $p = .75$. The average total fixation time was 96.56 seconds ($SD = 49.9$) in T, 120.04 in TR ($SD = 59.24$), and 100.159 in TRP ($SD = 59.24$). Similarly, the graph-overlay combination seemed not to affect the total fixation duration on the additional-information area $F(2, 18) = 1.48$, $p = .24$. The total fixation duration was on average 130.58 seconds in T ($SD = 109.9$), 131.94 in TR ($SD = 77.59$), and 103.78 in TRP ($SD = 57.72$).

In summary, *the results suggest that participants focused on the results list and the additional-information area in a similar manner in all conditions. However, contrary to our expectations, the different conditions did not affect the fixation time on the item description area.*

5.6. Questionnaires

Concerning the subjective evaluations obtained via questionnaires, all of the three system variants were overall well received by the participants. Friedman's test was run to ascertain differences in the subjective evaluations for all the dimensions considered, but none emerged. Overall, regarding system quality, it seemed that participants perceived all results returned by the search engine as valuable (for T $M = 3.45$($SD = .52$), for TR $M = 3.61$ ($SD = .65$) and for TRP $M = 3.62$ ($SD = .66$)). The interface design was also appreciated (for T $M = 3.41$ ($SD = .65$), for TR $M = 3.54$ ($SD = .64$), and for TRP $M = 3.39$ ($SD = .83$)), and participants also rated positively the transparency of the graph overlay (T $M = 3.77$ ($SD = .57$), for TR $M = 3.66$ ($SD = .5$), and for TRP $M = 3.79$ $SD = .75$). Moreover, participants expressed the intention of using the search engine again (for T $M = 3.62$ ($SD = .75$), for TR $M = 3.64$ ($SD = .69$), and for TRP $M = 3.56$ ($SD = .83$)). Participants also expressed a positive attitude regarding the trustworthiness of the system (for T $M = 3.12$ ($SD = .48$), for TR $M = 3.26$ ($SD = .55$), and for TRP $M = 3.2$ ($SD = .47$)).

In summary, *participants received all system variants positively, but no differences in subjective evaluations were found.*

6. Simulations

The user study compared the MRRs of click and select position in each list for different systems. This is a good indicator of the average system performance. However, the selected and clicked items may have been different in different systems. To confirm that the results from the user study were not dependent on the differences in the subjectively selected items, we run simulations in which the positions of the selected item were compared to the expected positions of these items in other system variants. In other words, we compared where the items would have been positioned in the other systems if the same interactions were used as an input.

Table 2
Results of the simulations. The first column reports the MRR of the item chosen by the user in TRP and the expected MRR in T and TR. The second column shows the MRR for TR and the expected MRR for T. $^{*}p = .05$.

	Ranking improvement in TRP	Ranking improvement in TR
	M(SD)	*M(SD)*
T	.19(.24)	.16(.22)
TR	.28(.28)	**.31(.2)***
TRP	**.4(.20)***	–
Results	TRP < T 19.42%	TR > T 93.75%

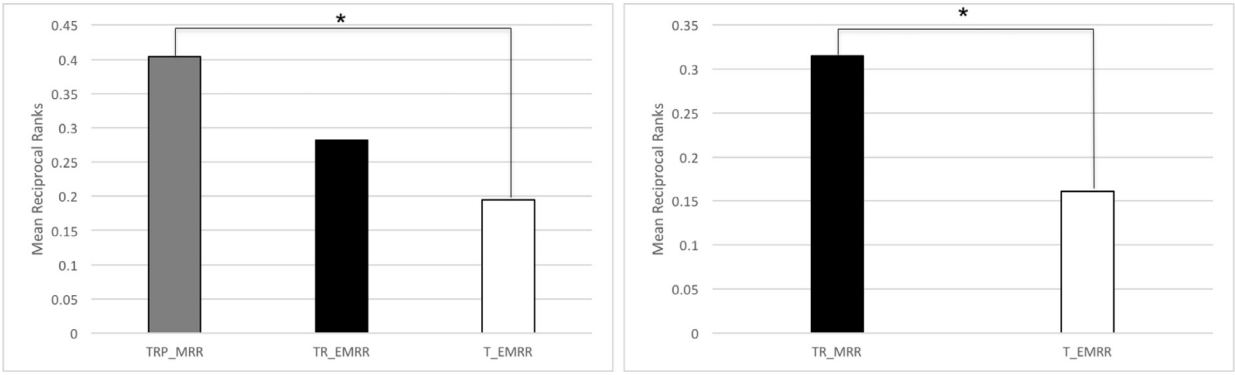

Fig. 8. Simulation results. On the left, the comparisons between the actual MRR observed in TRP and the expected MRR simulated in TR and in T. On the right, the comparison between the actual MRR observed in TR and the expected MRR simulated in T. Asterisks indicate a significant effect $^{*}p < .05$. Further details are provided in Table 2.

6.1. Simulation setup

The hypothesis was that the system variants employing multiple information overlays would yield to an improved ranking compared to those employing a fewer data layers.

The simulation compared the actual mean rank of the item participants selected when using the most complex system variant, i.e., TRP, against the mean rank that the user's choice would have had in the simpler system variants, that is, the expected mean rank in TR and T. A further simulation was run to contrast the mean rank of the item actually chosen by participants in TR against the mean rank that item would have had in the simplest system variant, that is, the expected mean rank in T. Consistently with the metrics chosen, the comparative analysis was run on the MRR. In particular, the MRR of the user's effective selection in TRP was compared against the expected MRR in TR and T using of two separate Wilcoxon tests.

6.2. Results

The results of the simulations are shown in Table 2 and Fig. 8. The results show a significant difference between the MRR in TRP and the expected MRR in T $Z = 2.914$ $p = .004$, with the MRR in TRP being higher ($M = .4$, $SD = .20$) as compared to the expected MRR in T ($M = .19$, $SD = .24$). Despite the MRR in TRP being higher compared to the expected MRR in TR ($M = .28$, $SD = .28$), the difference was not significant $Z = 1.6$ $p = .11$. A further Wilcoxon test was run to assess if there was an improvement in the ranking between TR and the most basic system variant, i.e., T. The analysis showed a significant difference between the two system variants, $Z = 2.286$, $p = .022$, with the MRR in TR ($M = .31$, $SD = .2$) being higher than the expected MRR in T ($M = .16$, $SD = .22$).

In summary, *the ranking was found to be improved in system variants that utilized the social-rating and personalization overlays.*

7. Discussion and conclusions

Popular services on the social web (e.g., booking or advising on trips, e-commerce, or local business services) increasingly use social data to provide better access to information. This motivates research on understanding how social data can be used and how it affects users' search performance and behavior.

Previous works provide evidence of the viability and benefit of personalized social search (Bao et al., 2007; Carmel et al., 2009; Guy et al., 2010); however, these are either not investigated in interactive online settings or do not include an actual user study.

We contributed by presenting a method that can utilize social-web data for interactive search and reported a user study that showed the effectiveness of personalization and social-data layers for social-web search.

7.1. Empirical findings

We conducted a user study that measured how social data affects users' search performance and behavior. The following effects on improved decision-making to select information and search effectiveness were found when using the social overlays:

- **Information-selection support.** The combination of the user profile overlay and the social-rating overlays was effective in reducing the need to consult external sources to select information, thus suggesting that the information provided in the result list is more sufficient for selecting information when complemented with the information from the social layer.
- **Search effectiveness.** The social-rating overlay was effective in improving the rankings, more specifically we observed an improvement when using the social-rating overlay in combination with the personalization overlay. This effect emerged in the user study and in the simulations.

Differences were not found in subjective evaluations, task-execution time, or user focus:

- The participants positively received the system variants, but no differences in subjective evaluations were found.
- The participants focused on the result list and the additional-information area in a similar manner in all conditions.
- The participants focused on user-interface elements in a similar manner.
- The average task-execution time was unaffected by the different combination of graph overlays. However, this was expected, as we did not limit the time available, and the participants were not asked to pay attention to the task-execution time.

The results indicate that personalized social ranking leads to more effective search outcomes compared to baseline variants of the system, as highlighted by both the user experiment and the simulations. The effect of the social layer was significant in the system variants encompassing the personal preference overlay.

The results also indicate that participants who used the system with social ratings and personalization overlays needed less external information to make their decisions on the information to select.

According to subjective evaluations, all of the system variants were well received by users; however, no significant difference emerged in any of the dimensions investigated. This is not surprising, as previous findings suggested that users cannot discriminate between recommendation systems with different degrees of accuracy (Knijnenburg et al., 2012) and that users recognize the benefits of personalization only for items that are meaningful to them (Tintarev & Masthoff, 2012). Additionally, behavioral evidence from the eye-tracking data did not highlight differences among the system variants. Behavioral evidence and subjective evaluations, being equal for all conditions, suggest that the effects that emerged in the user study are to be attributed solely to the different data-overlay combinations. Users looked for a comparable amount of time at the results list and the additional information before making a decision and judged their interactions similarly. This supports the conclusion that external information was not a factor and that the improvements can be solely attributed to the use of the social-information overlays. This data also suggest two different types of consulting behaviors: when they received more general results, i.e., T condition, they tended to consult a higher number of items' descriptions, each for a short time. Whereas, when the search results were more pointed, i.e., TRP condition, they still read item descriptions, but fewer of them and for a longer time. The former pattern of actions described resembles a quick check of a high number of elements, i.e., a skimming behavior, while the latter, suggests that participants were searching for a confirmation regarding the choice they were about to make.

7.2. Methodological contribution

The proposed method is based on a graph composed of data overlays referring to tags, ratings, and personal preferences. The method computes the relevance of different graph overlays and quantifies the contribution of each overlay to the ranking of search results.

Our approach consequently allows a user-interface design which can transparently visualize the contribution of each overlay in the user interface. The implementation of three versions of the user interface employing combinations of the different data overlays exemplify the generality of the method, as additional data overlays and different combinations can be considered within the same framework. By combining graph overlays in different ways, one can not only determine how the overlays contribute to retrieving and ranking items but also the information that is transparently provided in the user interface. As a consequence, the contribution of different overlays can be traced and communicated to the user. We believe such methods are needed to empirically study search and information-selection behavior of users and to quantify users' responses and consequently the benefits of visualizations.

7.3. Limitations

Our study was a controlled within-subjects user study in which all participants employed all of the different system variants. This experimental design allowed a direct comparison of the contribution of the different system variants and

to controlling for individual differences and behavioral patterns, which have been found to play a role in the perceived usefulness of recommender systems (Ekstrand et al., 2014; Hu & Pu, 2010).

However, the advantages of a controlled study come with limitations. Firstly, we used simulated work tasks, which may impair the naturalness of the tasks and possibly the associated user behaviors.

Secondly, our findings are based on a single dataset, which may limit the generalizability of our findings for scenarios involving data from different domains. However, the dataset that we used is the largest publicly available dataset of social-web data and contains real-world content, which guarantees that the data is also associated with typical problems and biases of real web data, such as sparsity and unbalance of observations.

Thirdly, in the experiment, the participants were asked to complete their personal profiles, rating a list of items, and then try the recommendation engine after a short time. The creation of the user profile and the usage of the system within the same session, may have affected what users found relevant, thus influencing their choices. In fact, they may have rated categories that they would not consider by themselves in the first place. Even though participants were asked to rate and select different elements (in the rating process they are asked to evaluate categories, whereas in the experimental sessions, they were asked to choose an actual point of interest), the arrangement of the data collection in two experimental sessions would possibly limit this effect.

In addition, we have to acknowledge that the system we employed did not consider the geographical location of the items recommended. However, for the purpose of the present study providing such information could have been misleading: participants did not know the city and such information would be trivial to them. Nevertheless, it could have affected their choices, thus confounding the effects of the information layers that were the subject of the test. In addition, it should be considered that typically geographical distance is used to filter out information, thus not being a core component of the computational method.

We found that social data generally improves search effectiveness and efficiency. However, we have to acknowledge that the contribution of the personal data integrated with the tag overlay was not investigated independently in the present study. Nevertheless, the effect of personal information requires further investigation.

In summary, our results suggest that the layer pertaining to social information is helpful to improve search quality, but the layer pertaining to personal data seems to be necessary to determine a significant improvement with respect to the baseline condition, i.e., the tags layer. This finding seems in line with a simulation and survey studies (Carmel et al., 2009), which showed that both in the off-line study and user survey, social network-based personalization significantly outperforms non-personalized social search and topic or content-based strategies.

Finally, the questionnaire chosen to investigate users' subjective impressions of the system variants was possibly not the best instrument for highlighting possible differences. From the participants' point of view, all the system variants may have looked very similar either in their appearance and responses, as the information in the dataset was new for them. Therefore, the recall of their interactions with each system variant may have produced similar scores. The addition of an online evaluation, e.g., a thinking-aloud protocol, could have been helpful to highlight the subtle differences that users may have notice while using each variant.

7.4. Summary and future work

In the present study we investigated the effect of combined deployment of different layers of social information on the users' search and information selection behavior. We further aimed at identifying which of the information layers could improve the search effectiveness, thus benefiting the user. In addition, we tested the effectiveness of the proposed computational method in an interactive online setting. Our findings suggest that the user's search and information-selection behavior was actually affected by the different overlays of information. In particular, we showed that the personalization of the user's searches using the information available from the social web can be an advantage to users, as demonstrated by the improved ranking and selection positions and the reduced need to consult external sources when selecting information. Additionally, the effectiveness of the social and personal layers was found also in the simulation. In line with previous work, we found that even if there was a significant change in users' search performance, they did not seem aware of differences in self-reports (Knijnenburg et al., 2012).

Whereas present findings show a significant advantage for the users, we see further research directions to be addressed. First, our study was a laboratory experiment, and users were primed with search tasks and were unable to try their own areas of interest and determine if they found that the suggestion effectively met their preferences and expectations. Our results suggest that larger scale in-the-wild deployments and experimentation could reveal additional insights into user behavior. Second, although we did not see improvement in task-completion time, this might have been due to the unconstrained time to perform the task. The computational and user-interface support for decision-making and information selection in time-constrained tasks could reveal additional benefits of our approach. These data can be used to inform the design of interactive search systems that leverage the social information available on the social web.

Third, the contributed method is aimed at providing more resources for the research community to investigate in a replicable manner how to combine diverse social and content overlays to study search and selection behavior. Similar alternative models have been proposed in multilayer-network literature (Barrat et al., 2004); discovery of community structures, connected components, and the use of tensor decompositions and various types of dynamical processes on multilayer networks (Kivelä et al., 2014). Such models could complement our model by being able to compute rankings from multi-relational

data (Ng et al., 2011), and leverage the rich semantic meaning of structural types of objects and links in the networks, and develop a structural analysis approach on mining semi-structured, multi-typed heterogeneous information networks (Sun & Han, 2013). Also approaches utilizing tensors as a lower-dimensional representation to combining the benefits of tensor factorization (Karatzoglou et al., 2010) with random walks would allow more local computation that can be feasible also with dynamic user input as in our search scenario.

Furthermore, in the current version of the system, the additional information that the participants used when making their information selection decisions is not personalized for the query. Nevertheless, this may be interesting enhancement to explore in future investigation.

Finally, the primary focus in our experiments was studying the effects of social information overlays and we did not extensively compare different computational methods. Future work could investigate alternative computational approaches including collaborative filtering using tensor factorization approaches (Rendle & Schmidt-Thieme, 2010), combining collaborative filtering with tag prediction (Wetzker et al., 2010), and alternative popularity based approaches (Trattner et al., 2016).

Acknowledgments

This work was partially funded by the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement No 601139 (CultAR) and Academy of Finland (278090, 305739).

References

Agrawal, R., Golshan, B., & Papalexakis, E. (2015). Whither social networks for web search? In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, *KDD '15, New York, NY, USA* (pp. 1661–1670). ACM.

Ahn, J.-W., & Brusilovsky, P. (2009). Adaptive visualization of search results: Bringing user models to visual analytics. *Information Visualization, 8*(3), 167–179.

Ahn, J.-W., & Brusilovsky, P. (2013). Adaptive visualization for exploratory information retrieval. *Information Processing & Management, 49*(5), 1139–1164.

Ahn, J.-W., Brusilovsky, P., He, D., Grady, J., & Li, Q. (2008). Personalized web exploration with task models. In *Proceedings of the 17th international conference on World Wide Web* (pp. 1–10). ACM.

Amitay, E., David, Har'El,N., Ofek-Koifman, S., Soffer, A., Yogev, S., & Golbandi, N. (2009). Social search and discovery using a unified approach. In Proceedings of the 20th ACM conference on hypertext and hypermedia, *HT '09, New York, NY, US* (pp. 199–208).

Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., & Su, Z. (2007). Optimizing web search using social annotations. In Proceedings of the 16th international conference on world wide web, *WWW '07* (pp. 501–510). New York, NY: ACM.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America, 101*(11), 3747–3752.

Biancalana, C., Gasparetti, F., Micarelli, A., & Sansonetti, G. (2013). Social semantic query expansion. *ACM Transactions on Intelligent Systems and Technology, 4*(4) 60:1–60:43.

Blanco, R., Mika, P., & Vigna, S. (2011). Effective and efficient entity search in rdf data. In Proceedings of the 10th international conference on the semantic web – Volume Part I, *ISWC'11, Berlin, Heidelberg* (pp. 83–97). Springer-Verlag.

Blomo, J., Ester, M., & Field, M. (2013). RecSys challenge. In *Proceedings of the 7th ACM Conference on Recommender Systems, New York, NY, US* (pp. 489–490). ACM. [online] Available at: https://www.kaggle.com/c/yelp-recsys-2013. Accessed 13.07.17.

Bouadjenek, M. R., Hacid, H., & Bouzeghoub, M. (2013a). Sopra: A new social personalized ranking function for improving web search. In Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval, *SIGIR '13, New York, NY, USA* (pp. 861–864). ACM.

Bouadjenek, M. R., Hacid, H., Bouzeghoub, M., & Vakali, A. (2013b). Using social annotations to enhance document representation for personalized search. In Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval, *SIGIR '13, New York, NY, USA* (pp. 1049–1052). ACM.

Bouadjenek, M. R., Hacid, H., Bouzeghoub, M., & Vakali, A. (2016). Persador: Personalized social document representation for improving web search. *Information Sciences, 369*, 614–633.

Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the fourteenth conference on uncertainty in artificial intelligence, *UAI'98, San Francisco, CA, USA* (pp. 43–52). Morgan Kaufmann Publishers Inc.

Carmel, D., Zwerdling, N., Guy, I., Ofek-Koifman, S., Har'el, N., & Ronen, I. (2009). Personalized social search based on the user's social network. In Proceedings of the 18th ACM conference on information and knowledge management, *CIKM '09, New York, NY, USA* (pp. 1227–1236). ACM.

Chi, E. H. (2009). Information seeking can be social. *Computer, 42*(3), 42–46.

Church, K., Neumann, J., Cherubini, M., & Oliver, N. (2010). Socialsearch-browser: A novel mobile search and information discovery tool. In Proceedings of the 15th international conference on intelligent user interfaces, *IUI '10, New York, NY, USA* (pp. 101–110). ACM.

Cremonesi, P., Epifania, F., & Garzotto, F. (2012). User profiling vs. accuracy in recommender system user experience. In Proceedings of the international working conference on advanced visual interfaces, *AVI '12, New York, NY, USA* (pp. 717–720). ACM.

Ekstrand, M. D., Harper, F. M., Willemsen, M. C., & Konstan, J. A. (2014). User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM conference on recommender systems* (pp. 161–168). ACM.

Fernquist, J., & Chi, E. H. (2013). Perception and understanding of social annotations in web search. In Proceedings of the 22nd international conference on world wide web, *WWW '13, New York, NY, USA* (pp. 403–412). ACM.

Freyne, J., Farzan, R., Brusilovsky, P., Smyth, B., & Coyle, M. (2007). Collecting community wisdom: Integrating social search & social navigation. In Proceedings of the 12th international conference on intelligent user interfaces, *IUI '07, New York, NY, USA* (pp. 52–61). ACM.

Gasparetti, F. (2016). Personalization and context-awareness in social local search: State-of-the-art and future research challenges. *Pervasive and Mobile Computing*.

Glowacka, D., Ruotsalo, T., Konuyshkova, K., Athukorala, k., Kaski, S., & Jacucci, G. (2013). Directing exploratory search: Reinforcement learning from user interactions with keywords. In Proceedings of the 2013 international conference on intelligent user interfaces, *IUI '13, New York, NY, USA* (pp. 117–128). ACM.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM, 35*(12), 61–70.

Guy, I., Zwerdling, N., Ronen, I., Carmel, D., & Uziel, E. (2010). Social media recommendation based on people and tags. In Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval, *SIGIR '10, New York, NY, USA* (pp. 194–201). ACM.

Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In Proceedings of the 3rd European conference on the semantic web: Research and applications, *ESWC'06, Berlin, Heidelberg* (pp. 411–426). Springer-Verlag.

Hu, R., & Pu, P. (2010). A study on user perception of personality-based recom-mender systems. *User Modeling, Adaptation, and Personalization*, 291–302.

Jeh, G., & Widom, J. (2003). Scaling personalized web search. In Proceedings of the 12th international conference on world wide web, *WWW '03, New York, NY, USA* (pp. 271–279). ACM.

Kahveci, B., Ismail Sengör, Altingövde, & Özgür, Ulusoy (2016). Integrating social features into mobile local search. *Journal of Systems and Software, 122*, 155–164.
Kammerer, Y., Nairn, R., Pirolli, P., & Chi, E. H. (2009). Signpost from the masses: Learning effects in an exploratory social tag search browser. In Proceedings of the SIGCHI conference on human factors in computing systems, *CHI '09, New York, NY, USA* (pp. 625–634). ACM.
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 '10, New York, NY, USA* (pp. 79–86). ACM.
Kawase, R., Siehndel, P., Pereira Nunes, B., Herder, E., & Nejdl, W. (2014). Exploiting the wisdom of the crowds for characterizing and connecting heterogeneous resources. In Proceedings of the 25th ACM conference on hypertext and social media, *HT '14, New York, NY, USA* (pp. 56–65). ACM.
Kim, Y. A., & Park, G. W. (2013). Topic-driven socialrank: Personalized search result ranking by identifying similar, credible users in a social network. *Knowledge-Based Systems, 54*, 230–242.
Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., & Porter, M. A. (2014). Multilayer networks. *Journal of Complex Networks, 2*(3), 203.
Klouche, K., Ruotsalo, T., Micallef, L., Andolina, S., & Jacucci, G. (2017). Visual re-ranking for multi-aspect information retrieval. In Proceedings of the 2017 conference on conference human information interaction and retrieval, *CHIIR '17, New York, NY, USA* (pp. 57–66). ACM.
Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction, 22*(4-5), 441–504.
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 '09, New York, NY, USA* (pp. 195–202). ACM.
Liu, J. (2009). Personalizing information retrieval using task features, topic knowledge, and task product. In Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, *SIGIR '09, New York, NY, USA*. ACM 855–855.
Luxenburger, J., Elbassuoni, S., & Weikum, G. (2008). Task-aware search personalization. In Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, *SIGIR '08, New York, NY, USA* (pp. 721–722). ACM.
Maniu, S., & Cautis, B. (2013). Network-aware search in social tagging applications: Instance optimality versus efficiency. In Proceedings of the 22nd ACM international conference on information & knowledge management, *CIKM '13, New York, NY, USA* (pp. 939–948). ACM.
McNee, S., Riedl, J., & Konstan, J. (2006). Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on human factors in computing systems* (pp. 1097–1101).
Muralidharan, A., Gyongyi, Z., & Chi, E. (2012). Social annotations in web search. In Proceedings of the SIGCHI conference on human factors in computing systems, *CHI '12, New York, NY, USA* (pp. 1085–1094). ACM.
Ng, M. K.-P., Li, X., & Ye, Y. (2011). Multirank: Co-ranking for objects and relations in multi-relational data. In Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, *KDD '11, New York, NY, USA* (pp. 1217–1225). ACM.
Noll, M. G., & Meinel, C. (2007). Web search personalization via social book-marking and tagging. In Proceedings of the 6th international the semantic web and 2nd asian conference on asian semantic web conference, *ISWC'07/ASWC'07, Berlin, Heidelberg* (pp. 367–380). Springer-Verlag.
Pantel, P., Gamon, M., Alonso, O., & Haas, K. (2012). Social annotations: Utility and prediction modeling. In Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval, *SIGIR '12, New York, NY, USA* (pp. 285–294). ACM.
Peltonen, J., Belorustceva, K., & Ruotsalo, T. (2017). Topic-relevance map: Visualization for improving search result comprehension. In Proceedings of the 22nd international conference on intelligent user interfaces, *IUI '17, New York, NY, USA* (pp. 611–622). ACM.
Pitkow, J., Schütze, H., Cass, T., Cooley, R., Turnbull, D., & Edmonds, A. (2002). Personalized search. *Communications of the ACM, 45*(9), 50–55.
Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, *SIGIR '98, New York, NY, USA* (pp. 275–281). ACM.
Pu, P., Chen, L., & Hu, R. (2011). A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on recommender systems* (pp. 14–21). i.
Rendle, S., & Schmidt-Thieme, L. (2010). Pairwise interaction tensor factorization for personalized tag recommendation. In Proceedings of the third ACM international conference on web search and data mining, *WSDM '10, New York, NY, USA* (pp. 81–90). ACM.
Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). Grouplens: An open architecture for collaborative filtering of netnews. In Proceedings of the 1994 ACM conference on computer supported cooperative work, *CSCW '94, New York, NY, USA* (pp. 175–186). ACM.
Ruotsalo, T., Jacucci, G., Myllymäki, P., & Kaski, S. (2014a). Interactive intent modeling: Information discovery beyond search. *Communications of the ACM, 58*(1), 86–92.
Ruotsalo, T., Klouche, K., Cabral, D., Andolina, S., & Jacucci, G. (2016). Flexible entity search on surfaces. In Proceedings of the 15th international conference on mobile and ubiquitous multimedia, *MUM '16, New York, NY, USA* (pp. 175–179). ACM.
Ruotsalo, T., Peltonen, J., Eugster, M., Glowacka, D., Konyushkova, K., & Athukorala, K. (2013). Directing exploratory search with interactive intent modeling. In Proceedings of the 22nd ACM international conference on information and knowledge management, *CIKM '13, New York, NY, US* (pp. 1759–1764).
Ruotsalo, T., Peltonen, J., Eugster, M. J., Glowacka, D., Reijonen, A., & Jacucci, G. (2014b). Intentradar: Search user interface that anticipates user's search intents. In CHI '14 extended abstracts on human factors in computing systems, *CHI EA '14* (pp. 455–458). New York, NY: ACM.
Schafer, J., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), The adaptive web, *volume 4321 of* lecture notes in computer science (pp. 291–324). BerlinHeidelberg: Springer.
Shafiq, O., Alhajj, R., & Rokne, J. G. (2015). On personalizing web search using social network analysis. *Information Sciences, 314*, 55–76.
Smyth, B. (2007). A community-based approach to personalizing web search. *Computer, 40*(8), 42–50.
Smyth, B., Balfe, E., Freyne, J., Briggs, P., Coyle, M., & Boydell, O. (2005). Exploiting query repetition and regularity in an adaptive community-based web search engine. *User Modeling and User-Adapted Interaction, 14*(5), 383–423.
Sun, Y., & Han, J. (2013). Mining heterogeneous information networks: A structural analysis approach. *SIGKDD Explorations Newsletter, 14*(2), 20–28.
Teevan, J., Dumais, S. T., & Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. In Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, *SIGIR '05, New York, NY, USA* (pp. 449–456). ACM.
Teevan, J., Dumais, S. T., & Horvitz, E. (2010). Potential for personalization. *ACM Transactions on Computer-Human Interaction, 17*(1) 4:1–4:31.
Tintarev, N., & Masthoff, J. (2012). Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction, 22*(4-5), 399–439.
Trattner, C., Kowald, D., Seitlinger, P., Ley, T., & Kopeinik, S. (2016). Modeling activation processes in human memory to predict the use of tags in social bookmarking systems. *The Journal of Web Science, 2*(1), 1–16.
Wetzker, R., Zimmermann, C., Bauckhage, C., & Albayrak, S. (2010). I tag, you tag: Translating tags for advanced user models. In Proceedings of the third ACM international conference on web search and data mining, *WSDM '10, New York, NY, USA* (pp. 71–80). ACM.
Wexelblat, A., & Maes, P. (1999). Footprints: History-rich tools for information foraging. In Proceedings of the SIGCHI conference on human factors in computing systems, *CHI '99, New York, NY, USA* (pp. 270–277). ACM.
Xie, H., Li, X., Wang, T., Lau, R. Y., Wong, T.-L., & Chen, L. (2016). Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy. *Information Processing & Management, 52*(1), 61–72.
Ye, M., Yin, P., & Lee, W.-C. (2010). Location recommendation for location-based social networks. In Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems, *GIS '10, New York, NY, USA* (pp. 458–461). ACM.
Ye, M., Yin, P., Lee, W.-C., & Lee, D.-L. (2011). Exploiting geographical influence for collaborative point-of-interest recommendation. In Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval, *SIGIR '11, New York, NY, USA* (pp. 325–334). ACM.
Yuan, Q., Cong, G., Ma, Z., Sun, A., & Thalmann, N. M. (2013). Time-aware point-of-interest recommendation. In Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval, *SIGIR '13, New York, NY, USA* (pp. 363–372). ACM.