Inf Syst Front
DOI 10.1007/s10796-015-9600-3

# Behavioral data mining to produce novel and serendipitous friend recommendations in a social bookmarking system

**Matteo Manca[1] · Ludovico Boratto[1] · Salvatore Carta[1]**



**Abstract** In the last few years, social media systems have experienced a fast growth. The amount of content shared in these systems increases fast, leading users to face the well known "interaction overload" problem, i.e., they are overwhelmed by content, so it becomes difficult to come across interesting items. To overcome this problem, *social recommender systems* have been recently designed and developed in order to filter content and recommend to users only interesting items. This type of filtering is usually affected by the "over-specialization" problem, which is related to recommendations that are too similar to the items already considered by the users. This paper proposes a friend recommender system that operates in the social bookmarking application domain and is based on *behavioral data mining*, i.e., on the exploitation of the users activity in a social bookmarking system. Experimental results show how this type of mining is able to produce accurate friend recommendations, allowing users to get to know bookmarked resources that are both novel and serendipitous. Using this approach, the impact of the "interaction overload" and the "over-specialization" problems is strongly reduced.

**Keywords** Social bookmarking · Friend recommendation · Behavioral data mining · Novelty · Serendipity

This work is partially funded by Regione Sardegna under project SocialGlue, through PIA - Pacchetti Integrati di Agevolazione "Industria Artigianato e Servizi" (annualità 2010), and by MIUR PRIN 2010-11 under project "Security Horizons".

✉ Ludovico Boratto
ludovico.boratto@unica.it

Matteo Manca
matteo.manca@unica.it

Salvatore Carta
salvatore@unica.it

1 Dipartimento di Matematica e Informatica, Università di Cagliari, Via Ospedale 72, 09124 Cagliari, Italy

## 1 Introduction

Social media systems are web-based services that allow individuals to build a public or semi-public profile within a bounded system, articulate a list of other users with whom they share a connection, and view and traverse their list of connections and those made by others within the system (Boyd and Ellison 2007).

A form of social media, known as *social bookmarking systems*, allows users to use keywords (*tags*) to describe resources that are of interest for them, helping to organize and share these resources with other users in the network (Farooq et al. 2007). The most widely-known examples of social bookmarking systems are Delicious,[1] where the bookmarked resources are web pages, CiteULike,[2] where users bookmark academic papers, and Flickr,[3] where each picture can be annotated with tags.

Due to the ever-growing amount of content and users that populate these systems, two problems arise in cascade.

**Social interaction overload problem** Social interaction overload (Guy et al. 2013; Simon 1971) is a problem related to the excessive amount of users and items that each user can

[1] http://www.delicious.com
[2] http://www.citeulike.org/
[3] http://www.flickr.com/

Published online: 26 October 2015

interact with. This leads to the scarcity of attention, which does not allow a user to focus on users or items that might be interesting for her/him.

In order to filter information in the social media domain, in the last few years the research on recommendation has brought to the development of a new class of systems, named *social recommender systems* (Ricci et al. 2011). These systems face the social interaction overload problem, by suggesting users or items that the users might be interested in. The filtering of the content in the social domain can be done by means of different metrics. For example, Chang et al. (2014) recently highlighted that the social similarity between users in a social network can be successfully adopted.

**Serendipity/Over-specialization problem** Recommender systems usually suggest items that have a strong match with the user profile, consequently the user always receives recommendations for items very similar to those that she/he already considered and never receives suggestions for unexpected, surprising, and novel items. This limit of recommender systems, known in the literature as "serendipity problem" or "over-specialization problem", worsens the user experience and does not give the users the opportunity to explore new items and to improve their knowledge (Shani and Gunawardana 2011). It is known that the serendipity problem affects both the content-based recommender systems (Lops et al. 2011) and the collaborative filtering approaches (Ziegler et al. 2005).

In Shani and Gunawardana (2011), the authors highlight that:

- the accuracy of a recommender system is important but it is not enough to evaluate it. Therefore, other metrics have to be considered in the evaluation of the system;
- users should be able to discover new items, i.e., diversity among the recommendations is a desired property (this aspect is also highlighted in Lops et al. (2011));
- when a recommender system is being designed, the aspects that have to be evaluated should have strong influence when deciding how to implement it.

Novelty and serendipity are two metrics that are gaining ever more attention in the performance evaluation of a recommender system. *Novelty* measures how many recommended items the user did not know about, while *serendipity* measures how surprising the successful recommendations are; serendipity can be seen as a way to diversify recommendations and to allow users to discover new items that they did not know they wanted. Lops et al. (2011) highlight that the main difference between a novel recommendation and a serendipitous recommendation is that a recommendation is novel when the user might have autonomously discovered the recommended item, while a recommendation is serendipitous when the user receives a recommendation that she/he might not have discovered. The authors also highlighted that the definition of new metrics to evaluate these aspects constitutes an interesting and important research topic.

**Our contributions** In this paper we present a friend recommender system that operates in the social bookmarking domain, designed and developed to face the social interaction and serendipity problems.

When a user adds another user as a "friend", she/he receives updates anytime a new bookmark is shared by the friend. Therefore, it is important that these bookmarks are related to resources that the target user has not already bookmarked. Moreover these bookmarks should generate interest in the user and should be diverse from those that already appear in her/his user profile. At the same time, the accuracy of the friend recommendations is a fundamental property. Indeed, becoming friend with someone with whom a strong relationship exists, also as an impact on the privacy and on the information the users share (Fogués et al. 2014), so predicting accurate friend recommendations is important.

Therefore, we aimed at developing a system able to:

- recommend friends with a high *accuracy*, i.e., users that are proved to be interesting for each other;
- recommend friends whose bookmarks are *novel* and *serendipitous*, i.e., bookmarks related to resources that the target user has not already considered and that are diverse enough from those available in her/his user profile.

More specifically, the problem statement is the following:

**Problem 1** We are given a social bookmarking system, defined as a tuple $Q = \{U, R, T, A, C\}$, where:

- $U$, $R$, and $T$ are sets of *users*, *resources*, and *tags*;
- $A$ is a ternary relation between the sets of *users*, *resources*, and *tags*, i.e., $A \subseteq U \times R \times T$, whose elements are the tag *assignments* of a user for a resource;
- $C$ is a binary relation between the users, i.e., $C \subseteq U \times U$, whose elements express the *connection* among two users. If we represent the user social relations by means of a graph, in which each node represents a user $u \in U$ and each edge $c \in C$ represents a connection among two users, we will have an undirected edge if the users are connected as *friends* and a directed edge if one user *follows* the other.

Our objective is to define a function $f : U \times U \to C$, which allows the system to derive if, given two users

$u \in U$ and $m \in U$, there is an undirected connection $c \in C$ among them.

The proposed solution is based on *behavioral data mining*, i.e., an analysis and exploitation of the user interaction with the content, in order to select and recommend only the users with the same interests. The effectiveness of behavioral data mining has been validated in various areas, such as the detection of tag clusters (Boratto et al. 2009, 2013), the creation of web personalization services (Mobasher et al. 2000), and the improvement of web search ranking (Agichtein et al. 2006). Moreover, Shih and Huang (2014) highlighted that by analyzing the interaction among the users in a social network, it is possible to predict future relationships among them, and Chen and Sharma (2013) showed that self-disclosure (i.e., the information shared by the users in a social website) enables the creation of connections among users. The choice to mine the implicit user behavior is also strengthened by the study by Centeno et al. (2015), which confirms the inaccuracy of the preferences explicitly expressed through numerical ratings to evaluate the reputation of the users. Therefore, we propose a form of mining that takes into account only the tags and the resources shared by the users, in order to be able to accurately recommend friends whose bookmarks can be novel and serendipitous for the target user.

This work is based on a friend recommender system previously presented by us (Manca et al. 2014), and extended by introducing new metrics and perspectives related to its performance evaluation.

The scientific contributions coming from this paper are the following:

- since in the literature it is known that there is no universal definition of novelty and serendipity (Lops et al. 2011) and there is no other work previous to ours that recommends friends in the social bookmarking domain, we propose definitions of novelty and serendipity in this context; moreover, we evaluate the proposed friend recommender system by analyzing how novel and serendipitous the bookmarks of the recommended friends are;
- we present a set of best practices and a critical discussion of the proposed system and of the obtained results, in order to support researchers in the development of a friend recommender system in the considered domain.

Moreover, we extend the approach presented in Manca et al. (2014), as follows:

- we propose a novel analysis of the user behavior in a social bookmarking system, in order to present the motivation to design a friend recommender system that mines user behavior;
- we extend the design of the system, in order to be able to produce recommendations that can be evaluated through novelty and serendipity;
- more details are given in the presentation of our approach, by presenting the pseudocode of the algorithms used by the friend recommender system.

The rest of the paper is organized as follows: in Section 2 we present the related work on user recommendation in social environments, in order to study the aspects that characterize this research area. This analysis of the literature allows us to design our recommender system, by considering the strengths and weaknesses of the presented works; Section 3 describes the details of the recommender system presented in this paper; in Section 4 we analyze novelty and serendipity in their classic definitions and propose definitions of novel and serendipitous recommendation in the social bookmarking domain; Section 5 illustrates the performed experiments and outlines main results; in Section 6 we present a critical discussion of the proposed approach and present a set of best practices to develop a friend recommender system in the social bookmarking domain; Section 7 contains comments, conclusions, and future work.

## 2 Related work

This section presents related work on user recommendation in the social domain. These systems can be classified into three categories, based on the source of data used to build the recommendations:

1. Systems based on the analysis of social graphs, which explore the set of people connected to the target user, in order to produce recommendations. These systems recommend either the closest users in the graph, like friends of friends and followees of followees (the “People you may know” feature offered by Facebook (Ratiu 2008) is the most widely known example of this approach), or recommend the users that have the highest probability to be crossed in a random walk of the social graph (the main reference for this type of systems is the “Who to follow” recommendation in Twitter (Gupta et al. 2013)).
2. Systems that analyze the interactions of the users with the content of the system (tags, likes, posts, etc.). In order to exploit the user interests, these systems usually build a user profile by giving a structured form to content, thanks to the use of metrics like TF-IDF (Term Frequency - Inverse Document Frequency).

An example of this class of systems is presented in Chen et al. (2009).

3. Hybrid systems, which consider both the social graph and the interactions of the users with the content (an example is represented by Hannon et al. (2010)).

The rest of the section presents the main approaches developed for each class of systems.

### 2.1 Systems based on the analysis of social graphs

Barbieri et al. (2014) recently presented an approach to predict links between users with a stochastic topic model. The model also represents whether a connection is "topical" or "social" and produces an explanation of the type of recommendation produced.

Gupta et al. (2013) presented Twitter's user recommendation service, which is based on shared interests, common connections, and other related factors. The proposed system builds a graph in which the vertices represent the users and the directed edges represent the "follow" relationship; this graph is processed with an open source in-memory graph processing engine, called Cassovary. Finally, recommendations are built by means of a user recommendation algorithm for directed graphs, based on SALSA (Stochastic Approach for Link-Structure Analysis).

Liben-Nowell and Kleinberg (2003) studied the user recommendation problem as a link prediction problem. They develop several approaches, based on metrics that analyze the proximity of the nodes in a social network, to infer the probability of new connections among users. Experiments show that the network topology is a good tool to predict future interactions.

### 2.2 Systems based on the interactions with the content

In Quercia and Capra (2009), a recommender system based on collocation (i.e., the position of a user) is presented. It uses short-range technologies of mobile phones, to infer the collocation and other correlated information, which are the base for the recommendations.

In Brzozowski and Romero (2011), a study about what cues in a user's profile, behavior, and network are the most effective to recommend people to follow, is presented.

In Arru et al. (2013), Arru et al. propose a user recommender system for Twitter, based on signal processing techniques. A signal is determined by the number of times a given concept occurs in a predefined period for a user. The considered approach defines a pattern-based similarity function among users and makes use of a time dimension in the representation of a user profile.

### 2.3 Hybrid systems

In Chen et al. (2009), Chen et al. describe a people recommender system in an enterprise social network domain. They compare four algorithms, two based on social relationship information and two based on content similarity, and demonstrate that the algorithms that use social information are stronger at finding known contacts, while the algorithms based on content similarity are better to discover new friends.

Guy et al. (2009) describe a people recommender system for the IBM Fringe social network. The system uses enterprise information, like org chart relationships, paper and patent co-authorships, and project co-memberships, which are specific of this social network.

Hannon et al. (2010) describe a followee recommender system for Twitter, which is based on tweets and relationships of their social graphs. By using this information, they build user profiles and demonstrate how these profiles can be used to produce recommendations.

Zhou et al. (2010) propose a framework for users' interest modeling and interest-based user recommendation (meant as people to follow and not as a friend), tested on the Yahoo! Delicious dataset. Recommendations are produced by analyzing the network and fans properties.

## 3 Mining user behavior to produce friend recommendations

This section presents the friend recommender system developed in our proposal. We first present an analysis of the user behavior in a social bookmarking system (Section 3.1), which led to the design of our recommender system (Section 3.2). In conclusion (Section 3.3), the algorithms that compose our system are presented.

### 3.1 User behavior in a social bookmarking system

This section aims at analyzing user behavior in a social bookmarking system from a friend recommendation point of view. In particular, we study how the bookmarking activity of a user is related to that of the others.

This analysis has been conducted on a Delicious dataset, distributed for the HetRec 2011 workshop (Cantador et al. 2011), which contains:

- 1867 users;
- 69226 URLs;
- 53388 tags;
- 7668 bi-directional user relations;
- 437593 tag assignments (i.e., tuples [user, tag, URL]);
- 104799 bookmarks (i.e., distinct pairs [user, URL]).

In their profiles, users had an average of 123.697 tags used to bookmark the resources, and an average of 56.132 bookmarked resources.

In order to focus on the possible connections among users, which might lead to friend recommendations in this system, we calculated the number of common tags and resources between the users in the dataset. The average number of common tags among two users is 7.807, while the average number of common resources among two users is 0.042. In particular, if we take into account only the users who have at least a common tag, the average number of tags that are in common for a couple of users increases to 10.417; if we consider only the users who have at least a bookmarked resource in common, the average number of common resources for each couple of users increases to 1.673.

This analysis allows us to infer some properties related to the user behavior in a social bookmarking system, recapped below:

– the behavior of two users in a social bookmarking system is related, both in the way they use tags and in the resources they bookmark with those tags;
– both the number of tags and resources in common is much smaller than the number of tags and resources used by each user (more precisely, 10.4 out of 123.7 tags, and 1.7 out of 56.1 resources).

This behavioral analysis has been one of the aspects that characterized the design of the system, which is presented next.

### 3.2 System design

The objective of our work is to build a friend recommender system in the social bookmarking domain. In its design, we considered the following aspects:

(a) In Gupta et al. (2013), the authors highlight that Twitter is an “interest graph”, rather than a “social graph”. Moreover, the analysis of such a graph suffers from scalability issues and, in order to contain the complexity of a recommender system, no personal information could be used to produce the recommendations. The definition of interest graph can also be extended to social bookmarking systems, since a user can add as a friend or follow another user, in order to receive her/his newly added bookmarks.
(b) Social media systems grow rapidly. This means that the amount of content added to a social media system and the user population increase at a fast rate. A recommender system that operates in this context needs to build accurate profiles of the users, which have to be up-to-date with the constantly evolving preferences of the users.
(c) The previously conducted behavioral analysis highlighted that both the tags and the resources represent a form of connection among two users. In particular, the number of common tags and resources between users is a small subset of all the tags and resources bookmarked by the users.
(d) As Zhou et al. (2010) highlights, the tagging activity of the users reflects their interests. Therefore, the tags used by a user are an important source of information to infer the interests that characterize her/him.

Taking into account all these aspects, we designed a recommender system that operates in the following way.

Regarding point (a), we designed a system that only analyzes the content of the users (i.e., the tagged bookmarks). So, in order to avoid the limitations related to the graph analysis in this domain, our system belongs to the second class presented in Section 2, i.e., the one that analyzes the interactions of the users with the content of the system.

Regarding point (b), in order to efficiently and quickly update user profiles, the algorithms and metrics used by a system should be quickly computed, in order to keep the user profiles up-to-date. Therefore, we believe that a friend recommender system should mine *user behavior* (i.e., the interaction of the users with the content), more than the content itself. In fact, if metrics like TF-IDF were used in order to give a structured form to the resources, this would significantly increase the complexity of the system. Since social bookmarking systems grow at a fast rate, content mining would lead to have outdated profiles, so this alternative is discarded by our design.

Regarding point (c), we followed the intuition that, by analyzing users with a similar behavior (i.e., users who have a large amount of tags and resources in common), accurate recommendations can be produced. These recommendations, at the same time, can lead to novel and serendipitous bookmarks, since in a user profile there is a large amount of tags and resources that have not been used by the other users.

Regarding point (d), we embraced the theory that user interest is reflected by the tagging activity and we extended it, by following the intuition that users with similar interests make a similar use of tags and bookmark the same resources.

A detailed description of the system is presented next.

### 3.3 Algorithms

Given a target user $u_t \in U$, the system recommends the users with a high tag-based user similarity and a high percentage of common resources. The system works in five steps:

1. *Tag-based user profiling.* Given the tag assignments of each user, this step builds a user profile, based on the frequencies of the tags used by the user.
2. *Resource-based user profiling.* Given the tag assignments of each user, this step builds a user profile, based on the resources bookmarked by the user.
3. *Tag-based similarity computation.* The first metric, calculated among a target user $u_t$ and the other users, is based on the tag-based user profile. Pearson's correlation is used to derive the similarity.
4. *User interest computation.* The second computed metric is the interest of a user toward another user and it is represented by the percentage of common resources among them.
5. *Recommendations selection.* This step recommends to $u_t$ the users with both a tag-based similarity and a user interest higher than a threshold value.

The steps previously presented are recapped in Algorithm 1.

**Algorithm 1** Friend recommender system

Let $Q = \{U, R, T, A, C\}$ be a social bookmarking system;
Let $S$ be the candidate set of users to recommend;
**for** $i = 1 \ldots n$ **do**
    $u = U[i]$ ▷ User 1
    **for** $j = 1 \ldots n$ **do**
        **if** $U[i]! = U[j]$ **then**
            $m = U[j]$ ▷ User 2
            Let $\overrightarrow{v_u}$ be the Tag-based user profile for the user $u$;
            Let $\overrightarrow{v_m}$ be the Tag-based user profile for the user $m$;
            Let $\overrightarrow{p_u}$ be the Resource-based user profile for the user $u$;
            Let $\overrightarrow{p_m}$ be the Resource-based user profile for the user $m$;
            $sim = ts(\overrightarrow{v_u}, \overrightarrow{v_m})$ ▷ Eq. 3
            $user\_interest1 = ui(\overrightarrow{p_u}, \overrightarrow{p_m})$ ▷ Eq. 5
            $user\_interest2 = ui(\overrightarrow{p_m}, \overrightarrow{p_u})$ ▷ Eq. 6
            **if** $((sim > \alpha)$ && $((user\_interest1 > \beta)||(user\_interest2 > \beta)))$ **then**
                $S.add(u, m)$ ▷ Eq. 7
            **end if**
        **end if**
    **end for**
**end for**

In the following, we will give a detailed description of each step.

### 3.3.1 Tag-based User Profiling

This step builds a user profile, based on the tags available in the tag assignments of a user, considering the frequency of each used tag. Given the sets defined in Problem 1, we can first consider the tag assignments of a user $u$ as follows:

**Definition 1** Let $A(u) \subseteq A$ be the subset of A whose elements are the triples that contain a user $u \in U$, i.e., $\forall r \in R \wedge \forall t \in T, (u, r, t) \in A \Rightarrow (u, r, t) \in A(u)$. Let $A(u, t) \subseteq A(u)$ be the subset of A(u) whose elements are all the triples that contain a tag $t \in T$ used by a user $u \in U$, i.e., $\forall r \in R, (u, r, t) \in A(u) \Rightarrow (u, r, t) \in A(u, t)$.

A user can be profiled, according to her/his use of the tags, by considering the relative frequency of each tag, as follows:

$$v_{uj} = \frac{|A(u, t_j)|}{|A(u)|} \tag{1}$$

Equation 1 estimates the importance of a tag $t_j \in T$ in the profile of a user $u \in U$, by defining the relative frequency as the number of times the tag $t_j$ was used, divided by the number of tag assignments of $u$.

A tag-based user profile can be implemented by representing each user $u \in U$ as a vector $\overrightarrow{v_u} = \{v_{u1}, v_{u2}, ..., v_{uk}\}$, where each element $v_{uj}$ is the previously defined relative frequency and $k$ is the number of tags in the system.

### 3.3.2 Resource-based user profiling

This step builds another user profile, based on the resources bookmarked by each user. A user can be profiled, according to her/his bookmarked resources, by considering the fact that she/he bookmarked a resource (i.e., she/he expressed interest in it):

$$p_{uj} = \begin{cases} 1 \text{ if } \exists t \in T \mid (u, r_j, t) \in A(u) \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

Equation 2 estimates the interest of a user $u$ in a resource $r_j$ with a binary value, equal to 1 in case $r_j$ was bookmarked by $u$, and 0 otherwise.

A resource-based user profile can be implemented by representing each user $u \in U$ by means of a binary vector $\overrightarrow{p_u} = \{p_{u1}, p_{u2}, ..., p_{un}\}$, which represents the resources tagged by each user. Each element $p_{uj}$ is defined as previously illustrated and $n$ is the number of resources in the system.

### 3.3.3 Tag-based similarity computation

Since in Zhou et al. (2010) the authors highlight that the interests of the users are reflected in their tagging activities, our system computes the similarity among two tag-based user profiles with the Pearson's correlation coefficient (Pearson 1896). This metric was chosen because, as proved by Breese et al. (1998), it is the most effective for the similarity assessment among users. Moreover, an efficient algorithm that exploits a support-based upper bound exists (Xiong et al. 2004).

Let $(u, m)$ be a pair of users represented respectively by two vectors $\overrightarrow{v_u}$ and $\overrightarrow{v_m}$. Our algorithm computes the tag-based user similarity $ts$ as defined in Eq. 3:

$$ts(u, m) = \frac{\sum_{j \subset T_{um}} (v_{uj} - \overline{v}_u)(v_{mj} - \overline{v}_m)}{\sqrt{\sum_{j \subset T_{um}} (v_{uj} - \overline{v}_u)^2} \sqrt{\sum_{j \subset T_{um}} (v_{mj} - \overline{v}_m)^2}} \tag{3}$$

where $T_{um}$ represents the set of tags used by both users $u$ and $m$ and values $\overline{v}_u$ and $\overline{v}_m$ represent, respectively, the mean of the tag frequencies of user $u$ and user $m$. This metric compares the frequencies of all the tags used by the considered users. The similarity values range from 1.0, which indicates complete similarity, to $-1.0$, which indicates complete dissimilarity. Herlocker et al. (1999) demonstrated that negative similarities are not significant to evaluate the correlation among users, so in our algorithm we consider only positive values.

*3.3.4 User interest computation*

Given a pair of users $(u, m)$, in this step we compute two metrics based on the resources tagged by users. The former, $ui(u, m)$, represents the interest of a user $u$ toward a user $m$, while the latter, $ui(m, u)$, represents the interest of a user $m$ toward a user $u$.

We first consider the set of resources bookmarked by each user, and then consider the resources in common between two users.

**Definition 2** Let $R(u) \subseteq R$ be the subset of resources used by a user $u \in U$, i.e., $\forall r \in R, (u, r, t) \in A(u) \Rightarrow r \in R(u)$. Let $D(u, m) = R(u) \cap R(m)$ be the subset of resources bookmarked by both user $u$ and user $m$.

The *user interest* of a user $u$ in a user $m$ can be estimated as:

$$ui(u, m) = \frac{|D(u, m)|}{|R(u)|} \tag{4}$$

The level of interest of a user $u$ in a user $m$ is estimated as the number of resources bookmarked by both the users, divided by the number of resources bookmarked by user $u$. This means that the interest of user $m$ in user $u$ depends on the number of resources bookmarked by $m$ (i.e., when calculating $ui(m, u)$, the denominator would be $|R(m)|$).

The previously defined user interest $ui$, can be implemented by using the two resource-based user profiles $\overrightarrow{p_u}$ and $\overrightarrow{p_m}$ as follows:

$$ui(u, m) = \frac{\sum_{j=1}^{n} p_{uj} p_{mj}}{\sum_{j=1}^{n} p_{uj}} * 100 \tag{5}$$

$$ui(m, u) = \frac{\sum_{j=1}^{n} p_{uj} p_{mj}}{\sum_{j=1}^{n} p_{mj}} * 100 \tag{6}$$

where $n$ is the total number of resources of the system.

*3.3.5 Recommendations selection*

As Marlow et al. highlight (Marlow et al. 2006), the use of tags and resources in a social tagging system is associated to two different types of behavior in a tagging system. Therefore, the aggregation of the tag-based similarity and of the user interests into a single score would blur the information on how similar two users are for each type of behavior. This could lead to potentially inaccurate friend recommendations, like two users that use the same tags to describe completely different and unrelated types of resources. Therefore, our recommendation algorithm filters the users by considering both types of behavior. Once the tag-based similarities and the user interests have been computed for each pair of users, our system chooses a set of users to recommend to the target user by selecting:

- the ones that have a tag-based user similarity higher than a threshold value $\alpha$ (i.e., $ts > \alpha$);
- the ones that have a user interest (at least one of the two computed) higher than a threshold value $\beta$ (i.e., $ui > \beta$).

**Definition 3** Given a target user $u_t$, the candidate set of users to recommend $S(u_t)$ can be defined as

$$S(u_t) = \{u_i \in U \,|ts(u_t, u_i) > \alpha \,\&\&\, (ui(u_t, u_i) > \beta) \\ \| \,(ui(u_i, u_t) > \beta)\} \tag{7}$$

As Eq. 7 shows, the system creates a recommendation between two users if a similarity on both types of behavior exists. In particular, since in Section 3 we showed that the user interest (i.e., the amount of common resources between two users) represents a weaker form of connection, we relaxed the constraint on the reciprocity of the user interest and compared the similarities with an OR operator.

## 4 Novelty and serendipity in a friend recommender system

Novelty and serendipity are two metrics used in the evaluation of recommender systems. Precisely, *novelty* measures how many recommendations include items that the user did not know about (Konstan et al. 2006), while *serendipity* measures how surprising the successful recommendations are (Shani and Gunawardana 2011). Serendipity can be seen as a way to introduce diversification in the recommendation, in order to allow users to discover new items that they did not know they were interested in and to improve their knowledge. A novel recommendation might not be serendipitous, while a serendipitous recommendation, by definition, is always novel.

As highlighted in the Introduction, the development of new metrics to evaluate novelty and serendipity represents an interesting research topic (Lops et al. 2011). The recommender system presented in Section 3 is the first that operates in the social bookmarking domain, consequently approaches developed in order to evaluate the novelty and the serendipity of a recommender system in such domain do not exist. Here, we propose a definition of novelty and serendipity of the resources bookmarked by the recommended friends, and based on those definitions, the metrics employed to compute the novelty and the serendipity of a friend recommender system are proposed in Section 5.

**Definition 4** A resource $r \in R$ can be considered *novel* for a user $u \in U$ iff $r \notin R(u)$. We define as $N(u)$ the set of novel resources for the user $u$.

In the literature, several researches highlight that the serendipity of a resource can be computed by measuring its distance from the items previously considered by the target user (Lops et al. 2011; Shani and Gunawardana 2011; Iaquinta et al. 2008; Zhang and Hurley 2008). As already mentioned, our recommendations are produced by mining user behavior. When a user is recommended as a friend, we can determine if a resource she/he bookmarked is serendipitous for the target user, by computing how distant it is from the behavior of the user. So, the distance between a recommended resource and the resources already bookmarked by the target user is based on the tags used to classify the resources.

In order to define the concept of serendipitous resource for a given user, we first need to define the set $T(r)$ of tags used for a specific resource $r$:

**Definition 5** Let $T(r) = \{t \in T \mid \exists(u, r, t) \in A\}$ be the set of tags used for a given resource r.

Given the above definition, the similarity $sim(r_i, r_j)$ between two resources $r_i$ and $r_j$ can be defined by measuring the Jaccard index as follows:

$$sim(r_i, r_j) = \frac{|T(r_i) \bigcap T(r_j)|}{|T(r_i) \bigcup T(r_j)|} \tag{8}$$

where:

- $T(r_i) \bigcap T(r_j)$ represents the set of common tags used to bookmark the resources $r_i$ and $r_j$;
- $T(r_i) \bigcup T(r_j)$ represents the set that contains all tags used to bookmark the resources $r_i$ and $r_j$;

To better understand the computation of the resource similarity $sim(r_i, r_j)$, we can represent each resource $r$ as a k-dimensional binary vector $t = \{t_1, t_2, ..., t_k\}$, where $k$ is the number of tags used in the system and each value $t_i$ of the vector is computed as follows:

$$t_i = \begin{cases} 1 \text{ if } t_i \in T(r) \\ 0 \text{ otherwise} \end{cases} \tag{9}$$

Table 1 presents an example of how the resource similarity can be calculated. A resource $r_i$ was bookmarked with tags $t_1$ and $t_4$, while a resource $r_j$ was bookmarked with tags $t_2$ and $t_4$. So, $T(r_i) \cap T(r_j) = \{t_4\}$ (the cardinality of the set is 1), $T(r_i) \cup T(r_j) = \{t_1, t_2, t_4\}$ (the cardinality of the set is 3), and $sim(r_i, r_j) = \frac{1}{3}$.

Starting from Eq. 8 we can define a resource $r \in R$ as serendipitous for a user $u \in U$, as follows:

**Definition 6** A resource $r_i \in R$ can be considered serendipitous for a user $u \in U$ iff $r_i \notin R(u) \wedge \forall r_u \in R(u), sim(r_i, r_u) < 0.5$. We named $B(u)$ the set of serendipitous resources for the user $u$.

The choice to use 0.5 as a threshold was made because the Jaccard index takes values between 0 and 1 so, by considering similarities lower than 0.5, we filter only highly dissimilar resources.

## 5 Experimental framework

This section presents the framework used to perform the experiments. The employed dataset and the data preprocessing are first described (Section 5.1). Then, the metrics used for the evaluation are presented (Section 5.2). The last part of the section presents the experimental setup and the obtained results (Section 5.3).

**Table 1** Example of the vectors used to calculate the resource similarity

| | $t_1$ | $t_2$ | $t_3$ | $t_4$ |
|---|---|---|---|---|
| $T(r_i)$ | 1 | 0 | 0 | 1 |
| $T(r_j)$ | 0 | 1 | 0 | 1 |

### 5.1 Dataset and pre-processing

The experiments were performed on a Delicious dataset, distributed for the HetRec 2011 workshop (Cantador et al. 2011), which was presented in the analysis of the user behavior (Section 3). In particular, we can now associate the content of the dataset to the sets defined in this paper:

- the set of users is the set $U$ previously defined;
- the set of URLs is the set $R$ previously defined;
- the set of tags is the set $T$ previously defined;
- the set of user relations is the relation $C$ previously defined;
- the set of tag assignments is the relation $A$ previously defined;
- the set of bookmarks is the union of the subsets $R(u)$ previously defined.

We pre-processed the dataset, in order to remove the users that were considered as "inactive", i.e., the ones that used less than 5 tags or less then 5 URLs.

### 5.2 Metrics

As highlighted throughout all the paper, we aimed at developing a system whose accuracy was not the only objective that had to be pursued. This section presents the metrics used for the performance evaluation of the system.

#### *5.2.1 Precision*

In order to measure the accuracy of the system, we evaluate the effectiveness of the recommendations (i.e., which recommended friends are actually friends with the target user), by measuring its *precision*.

**Definition 7** Let $W$ be the total amount of recommendations produced by the system, i.e., $W = \cup S(u_t), \forall u_t \in U$. This set represents the positive outcomes, i.e., the sum of the *true positive* and the *false positive* recommendations. Let $Z$ be the amount of correct recommendations produced by the system, i.e., $Z \subseteq W = \{(u, m) \mid (u, m) \in W \wedge (u, m) \in C\}$. So, $Z$ represents the subset of recommendations for which there is a relation (i.e., a friend correlation) in the dataset. This subset represents the *true positive* recommendations.

Given the previously defined two sets, $W$ and $Z$, we can measure the *precision* of our recommender system as the number of correct recommendations, divided by the number of produced recommendations:

$$precision = \frac{true\ prositive}{true\ prositive\ +\ false\ positive} = \frac{|Z|}{|W|} \tag{10}$$

Even if the recall metric is usually computed along with precision, it captures a perspective that differs from the way our system operates. We propose a constraint-based approach that reduces the amount of selected users, while the recall measures completeness and quantity of recommendations (Buckland and Gey 1994). Because of the nature of the metric, it would be misleading to compute it in order to evaluate the accuracy of our system.

#### *5.2.2 Percentage of satisfied users*

The *percentage of satisfied users* metric evaluates the system from a similar (but different) perspective with respect to the precision. In fact, precision measures for how many pairs of users a correct recommendation was produced, while the percentage of satisfied users measures for how many individual users a correct recommendation was produced.

**Definition 8** Let $X \subseteq U$ be the subset of users for which a recommendation was produced, i.e., $X = \{u \in U | \exists (u, m) \in W\}$. Let $Y \subseteq U$ be the subset of users for which a correct recommendation was produced, i.e., $Y = \{u \in U | \exists (u, m) \in Z\}$.

The percentage of users satisfied by the recommendations can be computed by dividing the set of users for which a correct recommendations was produced by the set of users for which a recommendation was produced, as follows:

$$\%\ satisfied\ users = \frac{|Y|}{|X|} * 100 \tag{11}$$

#### *5.2.3 Novelty and serendipity*

The friend recommender system presented in this paper is based on a form of mining of the user interests. When a friend recommendation is built, at the same time also the content of the suggested users is recommended (i.e., their bookmarks).

The *novelty* for a set of recommendations can be computed as follows:

$$novelty = \frac{|\bigcup N(u)|}{|\bigcup R(u)|}, \forall u \in Y \tag{12}$$

So, the novelty is computed as the sum of novel resources recommended to each user, divided by the sum of resources recommended to each user. Novelty values range from 0 (all the recommended resources were already considered by target users) to 1 (all the recommended resources were novel).

The *serendipity* for a set of recommendations, instead, is computed as shown in Eq. 13.

$$Serendipity = \frac{|\bigcup B(u)|}{|\bigcup R(u)|}, \forall u \in Y \quad (13)$$

As Eq. 13 shows, serendipity is computed as the sum of serendipitous resources recommended to each user, divided by the sum of resources recommended to each user. Also serendipity values range from 0 to 1.

In the computation of the novelty and serendipity we considered only the bookmarks of the recommendations that belong to the set of true positives $Z$. In fact, if we measured the novelty of the bookmarks for all the users (no matter if the recommendation was correct of not), they might be new, but also worthless (Shani and Gunawardana 2011).

### 5.3 Strategy

We performed three different experiments. The first aims to make an *evaluation of the accuracy*, by measuring the precision of the system with different threshold values. The second experiment, performs an *evaluation of the satisfied users* in the produced recommendations, given a precision value. The third experiment is an *evaluation of novelty and serendipity* of the bookmarks shared by the friends.

In order to evaluate the accuracy, we implemented a state-of-the-art policy (Zhou et al. 2010), which we used as reference system. Zhou et al. (2010) developed a tag-based user recommendation framework and demonstrated that tags are the most effective source of information to produce recommendations. We compare the performance of our system with respect to that of the reference one (which uses only tags, i.e., $ui = 0$) in terms of precision. Supported by the thesis that the use of only one source of data leads to a better performance, we considered a second reference system, which considers only the user interest (i.e., $ts = 0$).

During the analysis of the performance, we evaluated all the values of the parameters $\alpha$ and $\beta$ between 0 and 1, using a 0.1 interval.

### 5.4 Experiments

The details of each performed experiment and its results are now presented. In Section 5.4.1, the evaluation of the accuracy is presented, while in Section 5.4.2 the satisfaction of the users is evaluated, and in Section 5.4.3 the details of the evaluation of novelty and serendipity are given.

#### *5.4.1 Evaluation of the accuracy*

Given a target user $u_t$, the system builds a set of candidate users to recommend, $S(u_t)$. For each recommended user $u_i \in S(u_t)$, we analyze the bi-directional user relations in the dataset (i.e., if $(u_t, u_i) \in C$), to check if there is a connection between the target user $u_t$ and the recommended user $u_i$ (i.e., if the users are friends). This set of experiments analyzes the performance of the system in terms of *precision*. Given different values of $\alpha$ and $\beta$, the precision of the system is calculated, in order to analyze how the performance of the system varies as the similarity between users grows. The obtained results are illustrated in Figs. 1 and 2.

Figure 1 shows how the precision values change with respect to the user interest $ui$. The figure contains a line for each possible value $\alpha$ of the threshold for the tag-based user similarity $ts$. We can observe that the precision values grow proportionally to the $ui$ values. This means that the more similar the users are (both in terms

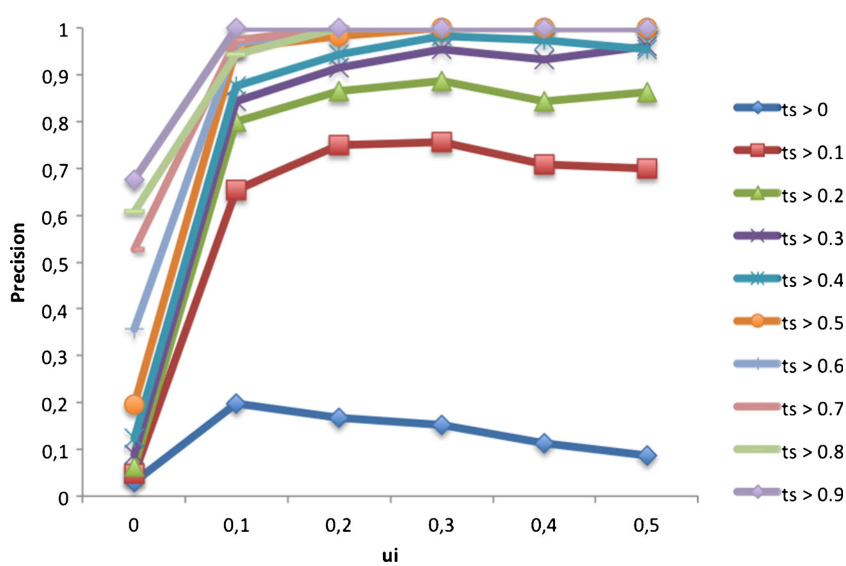


**Fig. 1** Precision of the system with respect to user interest $ui$, for each value of the $ts$ user similarity

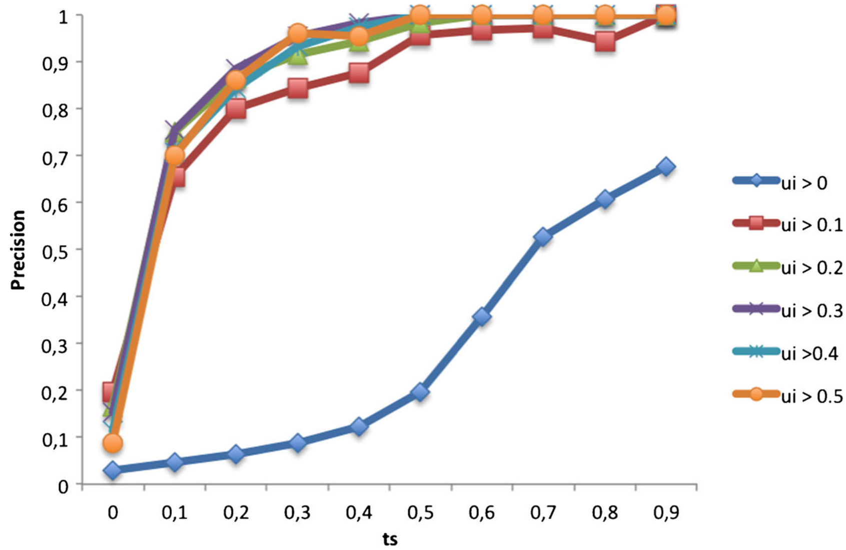


**Fig. 2** Precision of the system with respect to tag-based user similarity $ts$, for each value of the $ui$ user similarity

of tag-based similarity and of user interest), the better the system performs. However, for $ui$ values higher than 0.5 no user respects the constraints, so the system cannot produce any recommendation. This characteristic reflects our analysis of the user behavior, which highlighted that the amount of common resources among two users is low.

Figure 2 shows the same results from the tag-based user similarity point of view. The figure presents the precision values, with respect to the tag-based user similarity $ts$; here, each line shows the results for a given value $\beta$ of the threshold for the user interest $ui$. Also from this perspective, the precision grows proportionally to $ts$.

The blue lines in Figs. 1 and 2 show the results of the reference systems, where $ts = 0$ and $ui = 0$. In both cases, the two metrics combined improve the quality of the recommendations with respect to the cases where only one is used. These results show that, even if in the analysis of the user behavior previously presented we highlighted that the average number of resources in common among two users is very small, considering them through the metric $ui$ is important in order to compute accurate friend recommendations.

### 5.4.2 Evaluation of the satisfied users

The second set of experiments aims at analyzing the trend of the satisfied users, with respect to the precision values. So, for each precision value obtained in the previous experiment, we computed the percentage of satisfied users as shown in Eq. 11.

In order to present the results, Fig. 3 reports just a subset of precision values. These values have been selected by dividing the range $[0 - 1]$ of possible precision values into intervals of 0.1 (i.e, $[0 - 0.1)$, $[0.1 - 0.2)$, ..., $[0.9 - 1]$) and assigning each previously computed value of precision to the right interval. From each interval, we selected the record that corresponds to the precision value that maximized the percentage of satisfied users. The reason why there are no values for the intervals $[0.2 - 0.3)$ and $[0.4 - 0.5)$, is that in the previous experiments there are no values of $\alpha$ and $\beta$ that led to precision values inside these intervals.

In Fig. 3 we can observe that the percentage of satisfied users grows as the precision grows. Given that also in the previous experiments we obtained that the more similar the users were, the higher the precision was, we can conclude that the more similar the users are (both in terms of tag-based similarity and of user interest), the higher is the likelihood that users are satisfied by the recommendations.

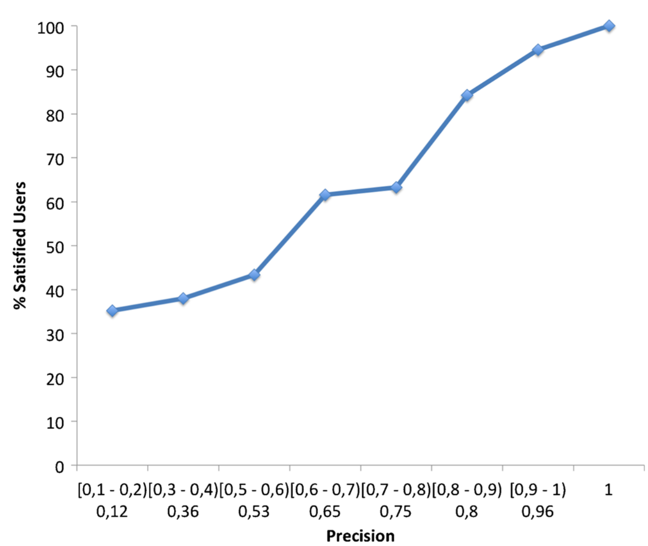


**Fig. 3** Percentage of satisfied users for different values of precision

These results show an interesting property of our recommender system. In fact, even if the precision values are split into intervals that cover the same range (i.e., 0.1), there are two of them (i.e., [0.6 − 0.7) and [0.8 − 0.9)) in which the percentage of individual users satisfied by the recommendations significantly increases. So this experiment, by showing the impact of precision on individual users, is very useful when tuning the parameters of the system.

### 5.4.3 Evaluation of novelty and serendipity

In this set of experiments we are going to evaluate the novelty and serendipity of our recommender system, by using the metrics previously proposed. Also this experiment is performed on a subset of cases and the evaluation has been done on the intervals previously considered.

In Table 2 we reported the values computed for $novelty$ and $serendipity$. Results highlight that both $novelty$ and $serendipity$ are inversely proportional to the precision. This means that the number of novel recommended bookmarks and the number of serendipitous recommended bookmarks decrease as the precision of the recommendations grows. However, we can notice that both novelty and serendipity decrease at a much lower rate, with respect to the increase of the accuracy. These results show that our system is capable to produce novel and serendipitous recommendations even when its precision is high.

## 6 Discussion

In this section, a summary of the main results of our proposal is given, in the form of a set of best practices aimed at a researcher or a software designer involved in real world scenarios where friend recommendations have to be produced in the social bookmarking domain.

**Table 2** Novelty and Serendipity

| Interval | Precision | Novelty | Serendipity |
|---|---|---|---|
| [0.0 − 0.1) | 0.03 | 0.96 | 0.92 |
| [0.1 − 0.2) | 0.12 | 0.93 | 0.81 |
| [0.2 − 0.3) | – | – | – |
| [0.3 − 0.4) | 0.36 | 0.90 | 0.65 |
| [0.4 − 0.5) | – | – | – |
| [0.5 − 0.6) | 0.53 | 0,89 | 0,54 |
| [0.6 − 0.7) | 0.65 | 0.83 | 0,69 |
| [0.7 − 0.8) | 0.75 | 0.74 | 0.59 |
| [0.8 − 0.9) | 0.88 | 0.79 | 0.61 |
| [0.9 − 1.0) | 0.97 | 0.79 | 0.53 |
| [1.0] | 1.00 | 0.67 | 0.47 |

The main questions that arise when approaching the design of the system are the following:

- given a social bookmarking system, composed by tagged resources and a social network/interest graph that connects the users, which source(s) of information should be exploited when producing the recommendations?
- if content is exploited, what type of mining should be done on it?
- how can a system be designed to produce recommendations that are accurate but also novel and serendipitous?
- is there a source of information that is not useful when building the recommendations?

Some general answers, coming from the results of our experiments, are the following:

- An analysis of the state of the art highlighted that mining the interest graph leads to scalability issues (see Sections 2 and 3.2). Therefore, friend recommendations in the social bookmarking domain should be built only by analyzing which resources each user bookmarked and with which tags. The experiment presented in Section 5.4.1 confirms that the mining of the resources and the tags leads to accurate friend recommendations;
- The mining done on the social bookmarking system should be *behavioral*. Social bookmarking systems grow at a very fast rate. So, in order to quickly update user profiles and follow the interests of the users, a friend recommender system that operates in this content has to analyze *how* the users bookmark the resources (i.e., with which tags and with which frequency), instead of analyzing its content, which would strongly increase the complexity of the recommender system. The results reported in Sections 5.4.1 and 5.4.2 report that this type of mining is strongly effective when producing friend recommendations in this context.
- In order to produce friend recommendations that lead to novel and serendipitous resources for the user, a system should be able to capture similarities among user behaviors, but also endorse, in the filtering, the users who have a subset of bookmarked resources and tags that are *diverse* from those considered by the user. Since in the behavioral analysis conducted in Section 3.1 we highlighted that the amount of common tags and resources is relatively small, but the experiments confirm that accurate recommendation can be produced, the tags and resources not in common with the other users leave space for the recommendation of novel and serendipitous resources.

– A behavioral data mining of the use of both tags and resources is necessary, in order to build accurate recommendations. Even though in the behavioral analysis of the users we showed that the amount of resources in common is very small (see Section 3.1) and in the system design we highlighted that in the literature only tags are considered (see Section 3.2), our experiments confirm that both tags and resources represent important sources of information when producing friend recommendations in this domain (see Sections 5.4.1 and 5.4.2).

More considerations can be done when particular conditions exist:

– Even if graph mining might lead to complexity and scalability issues in this context, it might be useful when a user has a little amount of resources and tags in her/his profile. Since our system works with common resources and tags, it presents limitations in this case.
– If the resources and the tags used by a user are different from those used by the others, she/he might not receive recommendations. In other words, the diversity exploited by our system to produce novel and serendipitous recommendations might become a limitation, if it is not also associated to resources in common with the other users. Also in this context, graph mining might be useful to produce recommendations.

## 7 Conclusions and future work

This paper presented a friend recommender system based on a form of behavioral data mining of the users in a social bookmarking system. By considering the frequency of the tags and which resources each user bookmarked, we selected only the users with similar profiles.

An analysis of the user behavior in this domain highlighted that the amount of tags and resources in common between two users is limited with respect to the amount of tags and resources bookmarked by each user. The fact that, given a user, a large amount of resources was not considered by the others, allowed us to design and implement a friend recommender system whose intent was to suggest friends with a high accuracy and that allowed users to come across novel and serendipitous bookmarks.

Since in the literature it is known that the definition of metrics to evaluate novelty and serendipity in a recommender system is an open research problem, we proposed new metrics that could be applied to our application domain and to the behavioral data mining used to build the recommendations.

Experiments evaluated the accuracy in terms of *precision* and results highlighted the capability of our system to build recommendations with an increasing accuracy as the similarity among users grows. Moreover, we evaluated the capability of our system to suggest friends whose bookmarks are *novel* and *serendipitous* and results show that even when a system achieves a high accuracy, it is still capable of producing novel and serendipitous recommendations.

Future work will be focused on adding a graph mining component to our system, in order to be able to produce recommendations also in the previously highlighted cases, in which users cannot receive recommendations.

## References

Agichtein, E., Brill, E., & Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06, pp. 19–26. ACM, New York, NY, USA*. doi:10.1145/1148170.1148177.

Arru, G., Gurini, D.F., Gasparetti, F., Micarelli, A., & Sansonetti, G. (2013). Signal-based user recommendation on twitter. In Carr, L., Laender, A.H.F., Lóscio, B.F., King, I., Fontoura, M., Vrandecic, D., Aroyo, L., de Oliveira, J.P.M., Lima, F., & Wilde, E. (Eds.), *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume, pp. 941–944. International World Wide Web Conferences Steering Committee / ACM*.

Barbieri, N., Bonchi, F., & Manco, G. (2014). Who to follow and why: Link prediction with explanations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, pp. 1266–1275. ACM, New York, NY, USA*. doi:10.1145/2623330.2623733.

Boratto, L., Carta, S., Manca, M., Mulas, F., Pilloni, P., Pinna, G., & Vargiu, E. (2013). A clustering approach for tag recommendation in social environments. *International Journal of E-Business Development*, *3*, 126–136.

Boratto, L., Carta, S., & Vargiu, E. (2009). Ratc: A robust automated tag clustering technique. In *E-Commerce and Web Technologies, 10th International Conference, EC-Web 2009. Proceedings, Lecture Notes in Computer Science, vol. 5692, pp. 324–335. Springer*.

Boyd, D.M., & Ellison, N.B. (2007). Social network sites: Definition, history, and scholarship. *Journal Computer-Mediated Communication*, *13*(1), 210–230.

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, 43–52. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA*. http://dl.acm.org/citation.cfm?id=2074094.2074100.

Brzozowski, M.J., & Romero, D.M. (2011). Who should i follow? recommending people in directed social networks. In Adamic, L.A., Baeza-Yates, R.A., & Counts, S. (Eds.), *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*: The AAAI Press.

Buckland, M., & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, *45*(1), 12–19. doi:10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L.

Cantador, I., Brusilovsky, P., & Kuflik, T. (2011). Second workshop on information heterogeneity and fusion in recommender systems (hetrec2011). In Mobasher, B., Burke, R.D., Jannach, D., & Adomavicius, G. (Eds.), *Proceedings of the 2011 ACM Conference on*

*Recommender Systems, RecSys 2011, Chicago, IL, USA, October 23-27, 2011, pp. 387–388*: ACM.

Centeno, R., Hermoso, R., & Fasli, M. (2015). On the inaccuracy of numerical ratings: Dealing with biased opinions in social networks. *Information Systems Frontiers*, *17*(4), 809–825. doi:10.1007/s10796-014-9526-1.

Chang, W.L., Diaz, A., & Hung, P. (2014). Estimating trust value: A social network perspective. *Information Systems Frontiers*, 1–20. doi:10.1007/s10796-014-9519-0.

Chen, J., Geyer, W., Dugan, C., Muller, M.J., & Guy, I. (2009). Make new friends, but keep the old: recommending people on social networking sites. In D.R.O. Jr., Arthur, R.B., Hinckley, K., Morris, M.R., Hudson, S.E., & Greenberg, S. (Eds.), *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, Boston, MA, USA, April 4-9, 2009, pp. 201–210*: ACM.

Chen, R., & Sharma, S.K. (2013). Self-disclosure at social networking sites: An exploration through relational capitals. *Information Systems Frontiers*, *15*(2), 269–278. doi:10.1007/s10796-011-9335-8.

Farooq, U., Kannampallil, T.G., Song, Y., Ganoe, C.H., Carroll, J.M., & Giles, C.L. (2007). Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. In Gross, T., & Inkpen, K. (Eds.), *Proceedings of the 2007 International ACM SIGGROUP Conference on Supporting Group Work, GROUP 2007, Sanibel Island, Florida, USA, November 4-7, 2007* (pp. 351–360): ACM.

Fogués, R.L., Such, J.M., Espinosa, A., & Garcia-Fornes, A. (2014). Bff: A tool for eliciting tie strength and user communities in social networking services. *Information Systems Frontiers*, *16*(2), 225–237. doi:10.1007/s10796-013-9453-6.

Gupta, P., Goel, A., Lin, J., Sharma, A., Wang, D., & Zadeh, R. (2013). Wtf: the who to follow service at twitter. In Schwabe, D., Almeida, V.A.F., Glaser, H., Baeza-Yates, R.A., & Moon, S.B. (Eds.), *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, pp. 505–514. International World Wide Web Conferences Steering Committee / ACM*.

Guy, I., Chen, L., & Zhou, M.X. (2013). Introduction to the special section on social recommender systems. *ACM TIST*, *4*(1), 7.

Guy, I., Ronen, I., & Wilcox, E. (2009). Do you know?: recommending people to invite into your social network. In Conati, C., Bauer, M., Oliver, N., & Weld, D.S. (Eds.), *Proceedings of the 2009 International Conference on Intelligent User Interfaces, February 8-11, 2009, Sanibel Island, Florida, USA* (pp. 77–86): ACM.

Hannon, J., Bennett, M., & Smyth, B. (2010). Recommending twitter users to follow using content and collaborative filtering approaches. In Amatriain, X., Torrens, M., Resnick, P., & Zanker, M. (Eds.), *Proceedings of the 2010 ACM Conference on Recommender Systems, RecSys 2010, Barcelona, Spain, September 26-30, 2010* (pp. 199–206): ACM.

Herlocker, J.L., Konstan, J.A., Borchers, A., & Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA, pp. 230–237. ACM*.

Iaquinta, L., de Gemmis, M., Lops, P., Semeraro, G., Filannino, M., & Molino, P. (2008). Introducing serendipity in a content-based recommender system. In *HIS, pp. 168–173. IEEE Computer Society*.

Konstan, J.A., McNee, S.M., Ziegler, C.N., Torres, R., Kapoor, N., & Riedl, J. (2006). Lessons on applying automated recommender systems to information-seeking tasks. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA, pp. 1630–1633. AAAI Press*.

Liben-Nowell, D., & Kleinberg, J.M. (2003). The link prediction problem for social networks. In *Proceedings of the 2003 ACM CIKM International Conference on Information and Knowledge Management, New Orleans, Louisiana, USA, November 2-8, 2003, pp. 556–559. ACM*.

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

Manca, M., Boratto, L., & Carta, S. (2014). Mining user behavior in a social bookmarking system - A delicious friend recommender system. In Helfert, M., Holzinger, A., Belo, O., & Francalanci, C. (Eds.), *DATA 2014 - Proceedings of 3rd International Conference on Data Management Technologies and Applications, Vienna, Austria, 29-31 August, 2014* (pp. 331–338): SciTePress.

Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia, HYPERTEXT '06, pp. 31–40. ACM, New York, NY, USA*. doi:10.1145/1149941.1149949.

Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of the ACM*, *43*(8), 142–151. doi:10.1145/345124.345169.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. iii. regression, heredity and panmixia. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Math. or Phys. Character (1896–1934), *187*, 253–318.

Quercia, D., & Capra, L. (2009). Friendsensing: recommending friends using mobile phones. In Bergman, L.D., Tuzhilin, A., Burke, R.D., Felfernig, A., & Schmidt-Thieme, L. (Eds.), *Proceedings of the 2009 ACM Conference on Recommender Systems, RecSys 2009, New York, NY, USA, October 23-25, 2009* (pp. 273–276): ACM.

Ratiu, F. (2008). Facebook: People you may know. https://blog.facebook.com/blog.php?post=15610312130.

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In Ricci, F., Rokach, L., Shapira, B., & Kantor, P.B. (Eds.), *Recommender Systems Handbook* (pp. 1–35): Springer.

Shani, G., & Gunawardana, A. (2011). Evaluating recommendation systems. In Ricci, F., Rokach, L., Shapira, B., & Kantor, P.B. (Eds.), *Recommender systems handbook* (pp. 257–297). Springer.

Shih, H.P., & Huang, E. (2014). Influences of web interactivity and social identity and bonds on the quality of online discussion in a virtual community. *Information Systems Frontiers*, *16*(4), 627–641. doi:10.1007/s10796-012-9376-7.

Simon, H.A. (1971). Designing organizations for an information rich world. In Greenberger, M. (Ed.), *Computers, communications, and the public interest* (pp. 37–72). Baltimore: Johns Hopkins Press.

Xiong, H., Shekhar, S., Tan, P.N., & Kumar, V. (2004). Exploiting a support-based upper bound of pearson's correlation coefficient for efficiently identifying strongly correlated pairs. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, pp. 334–343. ACM, New York, NY, USA*. doi:10.1145/1014052.1014090.

Zhang, M., & Hurley, N. (2008). Avoiding monotony: improving the diversity of recommendation lists. In *RecSys , 123–130. ACM*.

Zhou, T.C., Ma, H., Lyu, M.R., & King, I. (2010). Userrec: A user recommendation framework in social tagging systems. In Fox, M., & Poole, D. (Eds.), *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*: AAAI Press.

Ziegler, C.N., McNee, S.M., Konstan, J.A., & Lausen, G. (2005). Improving recommendation lists through topic diversification. In Ellis, A., & Hagino, T. (Eds.), *Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005, pp. 22–32*: ACM.