2009 Ninth International Conference on Intelligent Systems Design and Applications

Neighbor Selection and Recommendations in Social Bookmarking Tools

Antonina Dattolo, Felice Ferrara and Carlo Tasso
Dipartimento di Matematica ed Informatica
University of Udine
Via delle Scienze, Udine, Italy
E-mails: {antonina.dattolo,felice.ferrara,carlo.tasso}@dimi.uniud.it

***Abstract*—Web 2.0 applications innovate traditional informative services providing Web users with a set of tools for publishing and sharing information.**
Social bookmarking systems are an interesting example of this trend where users generate new contents. Unfortunately, the growing amount of available resources makes hard the task of accessing to relevant information in these environments.
Recommender systems face this problem filtering relevant resources connected to users' interests and preferences. In particular, collaborative filtering recommender systems produce suggestions using the opinions of similar users, called the neighbors. The task of finding neighbors is difficult in environment such as social bookmarking systems, since bookmarked resources belong to different domains.
In this paper we propose a methodology for partitioning users, tags and resources into domains of interest. Filtering tags and resources in accordance to the specific domains we can select a different set of neighbors for each domain, improving the accuracy of recommendations.

***Keywords*-Collaborative filtering recommender systems, social bookmarking systems, tags, zz-structures.**

I. Introduction

Web 2.0 applications allow users not only to consume services but also to produce new information. This process shifts the task of generating contents from a selected and restricted set of authors to a new, wide population of publishers.
An interesting example of this trend, which is going to innovate production and access to information, is represented from social bookmarking systems. These systems allow users to collect resources assigning them a set of tags. Often users employ the tags as indices for re-finding the resources which they have previously visited. The other users could also benefit from this process; in fact, tags offer a personal, meta description of the resources and can be used for finding peers with similar interests.
Unfortunately, the numeric explosion of generated resources makes this task difficult. Search engines, mainly Google™, are the most used tools for finding document on the Web, but they do not return personalized results; in particular they do not take in account users' preferences and goals.
Recommender systems [1] filter resources using information stored in a user model. In particular, collaborative filtering recommender systems compute similarity among users, identifying for each user the set of her/his *neighbors* (i.e. peers with similar preferences and features), and then suggest her/him new resources by considering the set of resources visited by neighbors.
So, collaborative filtering recommender systems apply the following two steps:

1) **Neighbor selection**. The behavior of a user A is compared with the behavior of other users in order to find a set of neighbors.
2) **Word-of-mouth simulation**. Resources, identified as interesting for neighbors of A, are suggested to her/him, simulating a word-of-mouth process. This step filters resources in accordance to the opinions of people similar to the user A for goals and background.

Focusing our attention on social bookmarking systems, two users can be considered "good neighbors" if they bookmarked a large common set of resources. Unfortunately, collaborative filtering recommender systems obtain appreciable results only when the resources belong to a same domain, since the process of neighbor selection is executed without taking in account the possibility that a user may have multiple interests. In order to clarify this limitation we consider Figure 1.

Using the traditional neighbor selection approach [2], the user B appears more similar to the user A than the user C, since B shares with A half of her/his bookmarks, while C shares with A only one-third of them. This neighbor selection does not take in account that items belong to different domains and each user may have variegated interests.

Our work faces the open issue of neighbor selection proposing for it a new approach. In previous works, we defined a reference model for a user concept space [3] and a publication sharing system [4]; here, extending these previous works, in order to avoid the problem described in Figure 1, we use tags for improving the neighbor selection phase: users' tags are grouped in clusters; each cluster identifies a user interest. During the neighbor selection phase, each cluster considers only neighbor's resources connected to the specific user interests. In this way, our technique recommends resources selecting a distinct set of neighbors for each different user interest.

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.245

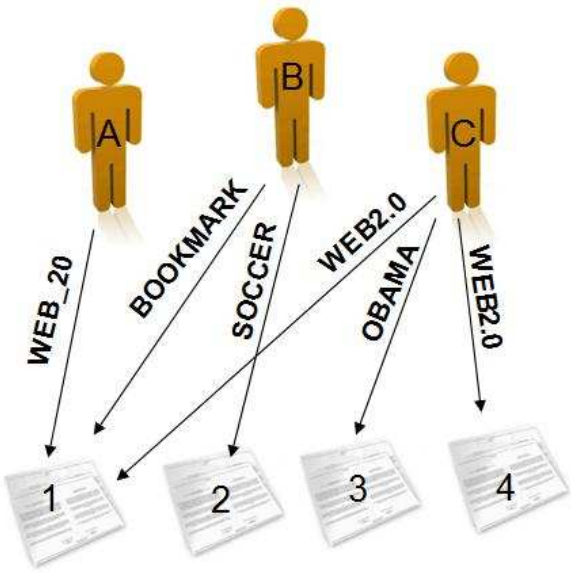

Figure 1. A social bookmark example.

This paper is organized as follow: Section II presents related work; Section III describes our application domain and the case study; Section IV provides a description of our methodology and in particular of the neighbor selection process; finally Section V concludes the paper, discussing future work.

II. Related work

Social bookmarking systems, such as BibSonomy[1] and delicious[2], collect resources, classified by users: these environments can be used in order to extract user profiles and generate personalized recommendations.
Actually, the use of tags for modeling users is still an open challenge; in fact, tags are often applied "just for a personal consumption" [5], making hard to infer users' preferences.
Different works, such as [6], [7] and [8], tackle limitations connected to the uncontrolled process of tagging inferring relationships between tags. In particular, in [6], the authors propose a recommender system based on tag clustering, which reduces the redundancy of folksonomies collapsing tags with a similar meaning in a single cluster: users' interests are modeled evaluating the number of accesses to each cluster, and then, the relevance of a resource is evaluated using clusters as a nexus between users and resources.
In [7], the authors assigns a ranking of Web pages taking in account factors such as tag popularity, tag representativeness and affinity between a user and a tag.
Tags' relationships are inferred also in [8], where the process of neighbor selection depends on tags used from users. This last work underlines that the accuracy of a collaborative filtering system depends on the precision of the neighbor selection process; tags can be used for finding similarities among users. This work does not recognize that neighbors should be selected in accordance to the current user's goal, but this feature can improve the accuracy of recommendations compared to traditional collaborative filtering techniques [9].
An attempt to apply this idea to social bookmarking tools is presented in [10] where the authors propose the idea of contextual collaborative filtering: tags detect a context in the neighbors' collections. However, this approach does not select neighbors in accordance to the user's goal since it only works on the word-of-mouth simulation.
Inferring relationships among tags, our work uses collective intelligence for partitioning tags and resources into domains of interests; in this way, the phase of neighbor selection is realized taking in account the current domain of interest: the similarity among users is evaluated considering only resources which appear connected to the user's domain of interest.

III. The case study

Our application domain are the social bookmarking systems; in particular we have realized a prototype, that works on the dump of the BibSonomy database, provided from organizers of the ECML PKDD Discovery Challenge 2009[3].

BibSonomy is both a social bookmarking and a publication sharing system; it can be used for sharing bookmarks and publication references. Using BibSonomy users can create and manage personal repositories organized as collections of tagged URLs and BibTeXs.
The starting dataset contain 1185 users, 14443 URLs, 7946 BibTeXs and 13276 tags. However, it presents some limitations which amplify the sparsity problem lowering the accuracy of results provided from a collaborative filtering recommender system.
Issue: *Different representations for the same Web content*.
The original dataset manages different URLs as different resources, without considering their content; but, if two or more URLs host a same content, it became important to identifies them as a unique resource, in order to correctly infer the neighbors.
There are two different situations in which different URLs may host the same content:

1) Two different URLs address the same page. For example, the URLs http://bibsonomy.org/ and http://www.bibsonomy.org/.
2) Two different URLs mirror a same site. By some estimates, 'as many as 40% of the Web pages are duplicates of other pages' [11]. Many of these are legitimate copies; for instance, certain information repositories are mirrored simply to provide redundancy

[1]http://www.bibsonomy.org/
[2]http://delicious.com/
[3]http://www.kde.cs.uni-kassel.de/ws/dc09/

and access reliability. Other duplicates are not legitimate copy, such as many blogs which re-propose information extracted from other Web sites.

Solution. In our approach, a dedicated spider examines the URLs, highlighting existing re-directions. This operation identifies the set of URLs addressing the same page.
In order to identify mirror pages, a simple approach could calculate a digest for each page and then compare these values in order to find duplicate documents. However, this solution does not work when two resources differ only for few characters, such as the HTML code generated in a dynamic way for advertising aims. This problem is known in the information retrieval field as the near duplicate problem [11] and many approaches have been developed. In our approach, we apply the shingling technique [12]: a document is described by means of all consecutive sequences of k terms in the resource. We consider two resources as the same if they share at least 80% of sequences.

Issue: *Different representations for the same BibTeX item*. The information inserted by users is often not accurate and then typing errors have to be considered in order to lower the sparsity problem.
Solution. In our approach, we apply the Levenshtein distance between all pairs of BibTeX titles in order to evaluate the cost of changing a title in another. The Levenshtein distance calculates the number of edit and delete operations needed to change a string in another one; in this way, we use it to detect typing errors, collapsing BibTeXs which differ only for a limited number of characters.
The application of this pre-processing phase produced a reduction in the number of distinct resources of 15%.
Analyzing this new dataset, we extract the relation existing between the popularity of a resource and the number of distinct tags, assigned to it. In Figure 2, the x-axis identifies resource popularity, expressed in terms of the numbers of times a given resource has been included in the personal space of a user; for all resources, bookmarked x times, the y-axis value shows the average number of distinct tags associated to them.

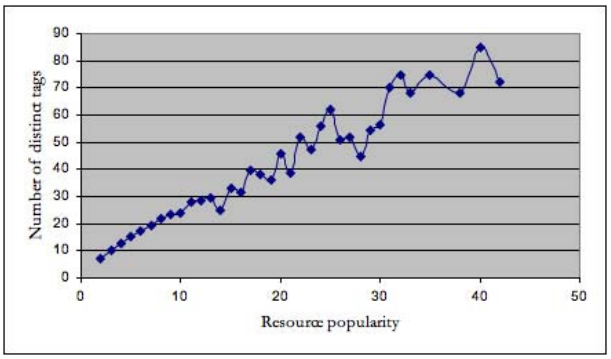

Figure 2. Resource popularity and number of distinct tags assigned to it.

Figure 2 highlights that there is a low level of agreement between users, since popular resources are typically tagged by means of an high number of distinct tags. For this reason, we believe that our approach, based on the analysis of similarities among tags, can improve the quality of generated recommendations overcoming limitations due to the redundancy of folksonomies.

IV. The Approach

This section offers details about the approach used for modeling users and the methodology proposed for analyzing the user model in order to provide personalized recommendations.

A. Organizing the knowledge base

In order to produce a set of recommendations we model a user u by means of his/her concept map, defined by means of a zz-structure [13]–[16], a graph-based structure S_u, used for organizing bookmarked resources by links indicating tags applied from u.
In particular, $S_u = (MG_u, T_u, t)$ is a zz-structure, an *edge-colored multigraph* where $MG_u = (V_u, E_u, f)$[4] is a multigraph, in which the set of vertices $V_u = \{p_1, \ldots, p_m\}$ is the collection of resources bookmarked from the user u, T_u is a set of colors (T refers to Tag) and $t : E_u \rightarrow T_u$ is an assignment of colors (tags) to edges of the multigraph; $\forall x \in V_u,\ \forall k = 1, 2, ..., |T_u|,\ deg^k(x) = 0, 1, 2$[5].

Figure 3 shows a simple example of a zz-structure used for modeling a generic user u.

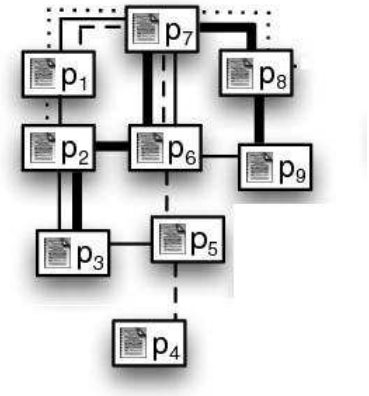

Figure 3. The zz-structure used for organizing bookmarks of a user u

Each tag identifies a link among vertices in order to represent the connection defined from the user. For istance, as shown in Figure 3, the user u tagged resources

[4]Multigraph definition: $MG_u = (V_u, E_u, f)$ is a multigraph composed of a set of *vertices* V_u, a set of *edges* E_u and a surjective function $f : E_u \rightarrow \{\{v, v'\} \mid v, v' \in V_u, v \neq v'\}$.

[5]$deg^k(x)$ denotes the degree (that is, the number of edges incident to x) of color t_k.

$p_3, p_2, p_6, p_7, p_8, p_9$ using the tag t_3, and then, the concept map stores this user interaction by means of a dimension graphically represented by means of a thick line.

Each color t_k can be used for selecting a specific sub-graph of M_u, constituted by the set of vertices V_u and edges $E_u^k \in E_u$, containing edges of the unique color t_k.
Each sub-graph of M_u is called *dimension* of color t_k and is denoted by D_u^k. Formally, a dimension $D_u^k = (V_u, E_u^k, f_u, \{t_k\}, t_u)$, with $k = 1, ..., |T_u|$, is a graph such that:

1) $E_u^k \neq \emptyset$ (at least an edge exists in each dimension);
2) $\forall x \in V_u$, $deg_u^k(x) = 0, 1, 2$ (the maximum degree of each vertex in each dimension is 2).

Looking at Figure 3, the concept space of the user u contains four dimensions identified with different types of line style.

B. Generating Recommendations

This section provides information about the approach used for detecting different domains of interest in the user profile. Then, it presents the methodology used for calculating recommendations connected to each discovered domain of interest.

1) Using tags for finding the user interest domains: For overcoming limitations presented in the Introduction, our approach infers information about similarities among tags. In particular, we consider two tags as similar if several users applied them on the same resources.
Formally, let $w^k(p)$ be the number of times that the tag t_k has been associated to the paper p from all the users of the system:

$$w^k(p) = \sum_{u' \in U} w_{u'}^k(p)$$

where

$$w_{u'}^k(p) = \begin{cases} 1 & \text{if } deg_{u'}^k(p) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

$w^k(p)$ is expressed in terms of $w_{u'}^k(p)$, which represents the number of times that t_k has been associated to the paper p *from each generic user* u'; in particular, $deg_{u'}^k(p) \neq 0$ indicates that the paper p has been tagged with t_k in the concept space of user u'.
Then, for each generic tag t_j, we can build the vector $\bar{w}^j = (w^j(p_1), \ldots, w^j(p_N))$, where $\{p_1, \ldots, p_N\}$ is the set of all the considered papers. This vector measures the similarity between a chosen tag t_k, and a generic tag t_j, applying the cosine similarity:

$$tag\_sim(t_k, t_j) = cos(\bar{w}^k, \bar{w}^j) = \frac{\bar{w}^k \cdot \bar{w}^j}{\|\bar{w}^k\| * \|\bar{w}^j\|}$$

Now, let T the set of tags used by all the users, $T_u \in T$ the tags used by user u, $T_u(f) \in T_u$ the set of tags applied from the user u at least f times.
In order to detect the domains of interest of a generic user u, we group tags contained in $T_u(f)$ in clusters, using the tag similarity metric described before; we include a tag in a cluster only if the similarity metric overcomes a given threshold, let it thr. In particular, starting from the most frequently tag, used by u, let it $t_i \in T_u(f)$, we define the cluster containing it, $T_u^i(f)$ in the following way:

$$T_u^i(f) = T_+^i \cap T_u$$

where $T_+^i = \{t \in T : tag\_sim(t_i, t) \geq thr\}$. Analogously, for each tag in $T_u(f)$, we generate another cluster, if it is not still included in other clusters.

Figure 4 shows an example of clusters generated for supporting a BibSonomy user u: in particular, starting from his/her most frequently used tag ($folksonomy$), on the left of the Figure 4 we show in the inner circle the set $T_+^{folksonomy}(0.4)$. Included in it and bordered using a polygon is $T_u^{folksonomy}(0.4)$: it contains the tags $\{bookmark, classification, tagging, social\}$.

Then, in the broader circle we show $T_+^{folksonomy}(0.3)$, the set of clustered tags generated by using a lower threshold (0.3); while, $T_u^{folksonomy}(0.3)$ (in the broader polygon) extends $T_u^{folksonomy}(0.4)$ by means of the tags $\{bookmarking, taxonomy, web20, tag\}$.

A similar process has been applied on the right of Figure 4 on the second most frequently used tag ($programming$) for the user u.

2) Generating recommendations for a domain of interest: This subsection presents our methodology for recommending to a user u the resources connected to a specific domain of interest, identified by the cluster $T_u^i(f)$. In particular, the recommendation process is based on the following steps:

1) **Finding neighbors.** Using the tags in $T_+^i(f)$, we extract from the concept maps of each user u' the set of dimensions, named $D_{u'}^i(f)$, which are identified by these tags. On these dimensions we look for neighbors and recommendable resources. The overlapping set between $D_u^i(f)$ and each other $D_{u'}^i(f)$ (where $u' \in U - \{u\}$) defines the user similarity on the topic t_i, as stated also from traditional collaborative filtering techniques [2]. The Jaccard similarity coefficient is applied as similarity metric.
Formally, $\forall\ u' \in U$:

$$user\_sim(D_u^i(f), D_{u'}^i(f)) = \frac{\left|V_u^i(f) \cap V_{u'}^i(f)\right|}{\left|V_u^i(f) \cup V_{u'}^i(f)\right|}$$

where $V_u^i(f)$ is the set of vertices in $D_u^i(f)$.
This metric evaluates the similarity between the user u and the others relatively to a specific topic t_i.

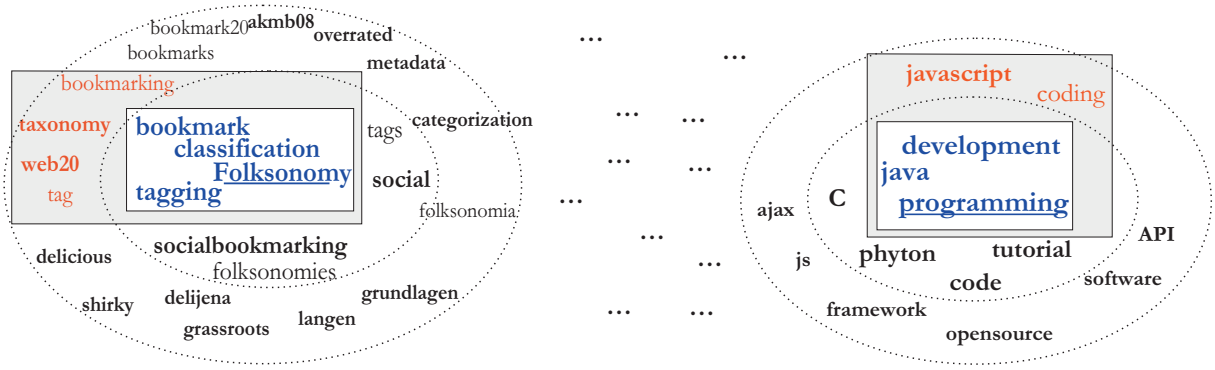

Figure 4. Four clusters for 'Folksonomy'

2) **Providing recommendations.** For ordering the resources and producing recommendations , we need to assign a score to each resource.
So, $\forall p \in \bigcup_{u' \in U - \{u\}} V(f)^i_{u'}$, we assign the following score:

$$score^i_u(p) = \sum_{u' : p \in V(f)^i_{u'}} user\_sim(D^i_u(f), D^i_{u'}(f))$$

This score measures the relevance of the resource p for a given user u in a given domain of interest t_i. Top scored resources are suggested to the user u. The same reasoning can be applied for each domain of interest.

V. Conclusion and Future Work

This paper proposed a collaborative filtering recommender system for suggesting resources in social bookmarking tools. The approach considers the user interests and selects her/his neighbors relatively to specific domains of interest. In order to evaluate the accuracy of our predictions we are going to execute:

- an *off-line analysis*. Using temporal information about bookmarks, we splitted the starting data set into two sets: a data set which covers bookmarks from 1995 to 2007, and a test set with all the other bookmarks (from January to December 2008).
  We are going to use the data set both for inferring relationships among tags and for modeling users in accordance to the Section IV-A, while the test set for a first measure of accuracy and for quantifying the relevance of user interest evolution over the time.
- a *live user experiment*. We are going to apply specific evaluation metrics, such as accuracy, novelty, coverage and user satisfaction [17].

Also, we are analyzing and calibrating the parameters involved in the generation of recommendations. In particular, we are working on:

- defining a tag and a user similarity thresholds;
- investigating the impact of the number of users which apply a tag on the tag similarity.
- analyzing how the user behavior (number of tagged resources, number of tag used) influences results.

References

[1] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[2] J. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative filtering recommender systems," in *The Adaptive Web*, ser. LNCS (4321), P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Berlin, Germany: Springer-Verlag, May 2007, ch. 9, pp. 291–324.

[3] A. Dattolo, F. Ferrara, and C. Tasso, "Supporting personalized user concept spaces and recommendations for a publication sharing system," in *Proc. of the first and Seventeenth International Conference User Modeling, Adaptation, and Personalization*, ser. LNCS (5525), H. G. and alt., Eds. Trento, Italy: Springer-Verlag, June 2009, pp. 325–330.

[4] ——, "Modeling a publication sharing system 2.0," in *Proc. of the Spesial Session on Accessing, Structuring, Analyzing and Adapting Information in Web 2.0, in connection with the 2nd International Conference on Human System Interaction*, May 2009, pp. 495–501.

[5] J. Riedl, "Altruism, selfishness, and destructiveness on the social web," in *AH '08: Proceedings of the 5th international conference on Adaptive Hypermedia and Adaptive Web-Based Systems*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 9–11.

[6] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke, "Personalized recommendation in collaborative tagging systems using hierarchical clustering," in *Proc. of the 2nd ACM Int. Conf. on Recommender Systems*, Lausanne, Switzerland, October 2008, pp. 259–266.

[7] F. Dolog and P. Durao, "A personalized tag-based recommendation in social web systems," in *Proc. of the Workshop on Adaptation and Personalization for Web 2.0, in connection with the first and Seventeenth International Conference User Modeling, Adaptation, and Personalization*, 2009, pp. 40–49.

[8] V. Zanardi and L. Capra, "Social ranking: Finding relevant content in web 2.0," in *Proc. of the 2nd ACM Int. Conf. on Recommender Systems*, Lausanne, Switzerland, October 2008, pp. 51–58.

[9] L. Baltrunas and F. Ricci, "Locally adaptive neighborhood selection for collaborative filtering recommendations," in *AH '08: Proceedings of the 5th international conference on Adaptive Hypermedia and Adaptive Web-Based Systems*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 22–31.

[10] R. Nakamoto, S. Nakajima, J. Miyazaki, and S. Uemura, "Tag-based contextual collaborative filtering," *International Journal of Computer Science*, vol. 34, no. 2, pp. 214–219, 2007.

[11] C. Manning, P. Raghavan, and H. Schutze, *An Introduction to Information Retrieval*. Cambridge, ENG: Cambridge University Press, 2008.

[12] A. Broder and K. Bharat, "Mirror, mirror on the web: a study of host pairs with replicated content," in *In Proceedings of the Eighth International World Wide Web Conference*, Toronto, Canada, May 1999, pp. 501–512.

[13] T. Nelson, "A cosmology for a different computer universe," *Jodi*, vol. 5, no. 1, p. 298, 2004.

[14] A. Dattolo and F. Luccio, "Visualizing personalized views in virtual museum tours," in *Proc. of the Int. Conf. on Human System Interaction*, Krakow, Poland, May 2008, pp. 339–346.

[15] ——, "A formal description of zigzag structures," in *Proc. of the Workshop on New Forms of Xanalogical Storage and Function, in connection with the 20th ACM Conference on Hypertext and Hypermedia*, Torino, Italy, June 2009, pp. 7–11.

[16] ——, "A state-of-art survey on zigzag structures," in *Proc. of the Workshop on New Forms of Xanalogical Storage and Function, in connection with the 20th ACM Conference on Hypertext and Hypermedia*, Torino, Italy, June 2009, pp. 1–6.

[17] J. Herlocker, J. Konstan, L. Terveen, and J. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 5–53, 2004.