ABSTRACT
This paper aims to examine whether users’ watching networks can improve collaborative filtering-based recommendations (CF). Watching networks are established by users upon their perceived usefulness or interests about other users’ information collections. The networks do not require mutual agreement between a watching party and a watched party. The typical example of this network is ‘following’ in Twitter, ‘watching’ on CiteULike, or ‘contacts’ on Flickr. Once a user declares that ‘I want to watch user A’, the user A’s information collection is displayed to the watching user, continuously. It can be interpreted to mean that a watching user found some shared interests in user A’s collection and want to refer to it in future. The approaches explored in this paper take advantage of this watching network as a part of user’s preferences for recommendations. To evaluate the potential of these approaches, we focus on a social tagging system, CiteULike. Our data shows that in this context, a hybrid recommendation approach that fuses CF and watching network-based recommendations outperforms both CF and network-based recommendations.

Categories and Subject Descriptors
H.5.3. Information Interfaces and presentation: Group and Organization Interfaces – Collaborative Computing

General Terms
Design, Experimentation, Human Factors

Keywords
Social Networks, Unilateral Relations, Watching Networks, Hybrid Recommendations, CiteULike

1. INTRODUCTION
Information recommendation or personalized information filtering has been getting attention with the increasing information glut problem. Collaborative filtering (CF) and content-based filtering are two popular technologies, which power modern recommender systems. The former technology is to utilize rating similarities among users or items, and the later is to utilize content similarities in items. Despite of the algorithmic differences, the common point is that, in order to acquire users’ preferences, these technologies utilize the information about ‘what users like’. After the advent of the successful Web 2.0, however, it has become equally important to know ‘whom users are interested in’. Compared with the era when the computer users stayed in isolation, users on the Web 2.0 have found it easier to find information through social networks in virtual space [30]. Following the lead of social networking sites (SNS), information systems based on collective intelligence such as social tagging systems, image or video sharing systems, etc. nowadays actively support the users to get connected with other users. However, the nature of connections in these two groups of systems is slightly different.

Unlike social networking systems mainly focusing on linking mutually agreed-up friendships, many social tagging systems, which help users to manage and share interesting information online, offer new kind of sociability. Once users find other users who have interesting or similar tastes, they are allowed to ‘watch’ the users’ information collections continuously. The ‘watch’ relations don’t require any offline interactions or emotional bonds to make connections or the mutual agreement for being connected. It is a unilateral relation which forms a special kind of social network, which we named a ‘watching network’. The most typical examples of watching network are “following” on Twitter, “watching” on CiteULike, “network” on Delicious or “contacts” on Flickr.

Unilateral relations gained attention along with the success of social tagging and micro-blogging applications [9-10, 12]. Wellman suggested that, in Web 2.0 era, various new relationships would emerge and the networks are “less bounded [30].” The unilateral relationship is one kind of the new relationships. Unlike befriending in SNS (e.g. facebook, MySpace or Friendster) where users increase the number of friends simply for fun or curiosity [11], the unilateral relationships aim to acquire information from the connected people’s collection. When a user unilaterally follows many users, it may cause rapid growth and potential dilution of the item list in the follower’s collections. Therefore, unilateral relationships require users to be careful to select people to follow, based on the utility of information.

Some researchers may argue that unilateral relations are not social connections, since they may be not based on social interactions or emotional bonds. However, in our previous works [15-16, 18], we found an important social property in the relationships; homophily. Specifically, the relations met the similarity attraction hypothesis [22] and held the transitivity power [23]. High degree of similarity was embedded in unilateral relations and the similarity decreased with the increase of distance. Even though two nodes (i.e. two users) tend not to interact personally or not to share any emotion, they are directed networks on the center of interesting information objects and it could be called as ‘object-centered sociality’ [11].

We interpreted that these connections mimic the process of bookmarking interesting items. In this context, users bookmark
other interesting users. Therefore, we utilize the watched users’ information as a part of the watching users’ preferences. In this paper, we show how the watching networks can be used as a component of user profiles and whether they improve the recommendation qualities or not, in a social tagging system; CiteULike.

2. RELATED WORK

2.1 Users’ Self-Defined Network-based Recommendations

In early CF-based research, users’ social networks were important parts of recommendation process. The first collaborative filtering system, Tapestry was based on explicit social connections and allowed users to retrieve personalized contents, using annotations added by their friends and colleagues [26]. Another pioneering CF project [19] combined explicit social connections with active “push” approach: users could directly send interesting research papers to other colleagues. However, these pioneer systems relied on exchange of information within a “small world” and found it difficult to retain users and to keep them actively contributing to the recommendation process. As the CF-based algorithms became mature, automatic recommendations by computation are now dominant. The technology proved its worth in recommending taste-based items such as movies, jokes, music, etc. where the preference is hard to be comprehended by the content. It became popular for its ability to recommend serendipitous and diverse recommendations.

More recently, however, some problems of CF technology caused some researchers re-assess the value of explicit social connections as a source of information for reliable recommendations. There are malicious ad-hoc users who play with recommenders to make profit or distort systems [14, 20-21], peculiar users with unusual tastes, cold-start problem [26] or expensive computational off-line process [20].

Sinha and Swearingen compared the quality of recommendations generated by an online recommender and the ones by friends. They concluded that friends’ recommendations were more useful and better than those of the recommender system. The recommendations of friends also are more trustworthy than those of the system [28]. In addition, Golbeck showed that users prefer recommendations from trusted people [6]. One of reasons why users prefer friends’ recommendations can be found in Singla & Richardson’s study [27]. They examined the logs of instant messenger and search engines, and determined how much information-seeking activities were similar among friends. They found that two people who had talked to each other on the messenger shared significantly similar interests than other pairs of users who did not. When they compared that with the common interests of random user pairs who were similar in demographics, friend pairs shared significantly similar interests. They found homophily exhibited on the Internet [27]. Groh and Ehmig also showed that users who were connected as friends had similar ratings in the taste-related domain, such as their preferences about bars. They also found that the rating similarity within a clique (a group of more than 2 friends) was stronger than that of two friends, because they feared isolation in their group [7].

Bonhard and Sasse explained another different reason for that. They suggested that recommendation is a part of decision-making process. Advice-seekers decide the value of suggested items according to the identity of the recommender. Therefore, the relationship between the receiver of recommended information and the source of that information is critical. Their study found that, along with rating overlap, profile similarities such as demographics, preferences and interests play an important role in trustworthy recommendations [3]. In current CF technology, the process to decide the value of an item is a black-box to users because it is based upon unknown users.

Massa and Avesani’s study showed that a user’s trust network can solve the ad-hoc user problem, improve recommendation prediction and attenuate the computational complexity. Using Epinions data set, a trust-based technology generated more precise recommendations than CF technology. In addition, for users with 4 ratings, trust-based technology count make recommendations for 66% of these users, while CF could only make recommendation for 14% of the users with a higher margin of error [20]. Another study indicated that a trust network decreases the recommendation error and increases the accuracy as well [24]. For users with a unique taste, their own trusted network could increase the satisfaction of recommendations, since they are able to know where the information came from [29].

2.2 Social Tag-based User Profiling

User profiles representing users’ preferences or interests are major basis for personalizing information. Traditional recommender systems including content-based and collaborative filtering based recommendations are utilizing numeric users’ ratings. In social tagging systems, however, the numeric rating schema is commonly unavailable. Hence, there have been several studies about how to build user profiles based only on users’ tags, which are known as expressions about users’ cognitive understanding. Au-Yeung, et al. (2008) started their study from an assumption that users are interested in multiple topics, and hence their personal collection of tags (i.e. personomy) may consist of several clusters. Using 1000 users’ data from Delicious, they produced clusters of Web resources tagged by each user’s personomy, and found that their assumption is true. In order to test the quality of clustered Web pages, they utilized the most often-annotated tags in a cluster (i.e. signature tags of the cluster) as a query and assessed how much the query can retrieve the relevant pages. As the result, neither too general nor too specific tags could not retrieve a good set of results in terms of precision and recall [2].

Godoy and Amandi [5] suggested a way to define users’ profiles based on users’ visited Web pages and the tags. The profile is built upon an incremental, unsupervised conceptual learning algorithm; WebDCC (Web Document Conceptual Clustering). Web pages that users had visited are represented as word vectors, which later are clustered conceptually. The clustered vectors are summarized by each succinct description using automatic extraction of the descriptors. Moreover, unlike other studies based on social tags, the authors organized a hierarchy of vectors and tags [5]. Guan and the other [8] also generate user profile based on tags and the contents of documents (e.g. scientific articles or Web pages). Specifically their profiles are generated by a graph-based representation learning algorithm. In order to infer users’ interests, they took advantage of three bi-partite relations of social tags such as user-document, tag-document, and user-tag. In this situation, however, it is hard to locate semantically related tags to a document. For instance, the documents tagged by “automobile” are treated differently with the ones tagged by “car.” As the solution, they utilized an affinity graph of documents, which are built by word vectors. To recommend interesting documents, they
set three bi-partite relations and affinity graph in one space. Then they examine how three components (a.k.a. user, document, and tag) are close to each other and the closest documents to a target user is recommended \[8\].

3. Recommendation Description

Our paper aims to compare the quality of traditional CF-based recommendations with one approach based on watching networks (WN), and a hybrid approach using both technologies. While our goal is the use of WN for recommendations, we are especially interested in hybrid approaches since we do not want to neglect the effectiveness and success of CF-based recommendations. This technology has been successfully applied to a number of systems for years. At the same time, we want to use the power of WN. We suggest that if there is any user who a target user is interested in, we need to consider him as a peer user. Therefore, our strategy of WN-based recommendations focuses on how to include watched users as a part of peer users.

To build a CF-based recommender engine, which is used as a baseline of this study, we use a special version of CF approach designed for social tagging systems, where numeric ratings are typically not available and users express their interests by bookmarking and tags. As explained, we use a CiteULike dataset. In the system, it is possible to rate the items, but CiteULike rating schema is both non-mandatory and non-standard, and hence can’t be used reliably. Many articles do not have any rating or do have a dubious rating “I’ve already read it”, which did not show any preference. Therefore, we use the Jaccard similarity coefficient for co-bookmarked information instead of Pearson correlation to compute the similarity between two users (equation 1) \[13\]. The Jaccard similarity represents the fraction of shared information in the joint information space of both users. Here A and B are sets of items in the collection of two users being compared.

\[
\text{Jaccard Similarity} = \frac{(A \cap B)}{(A \cup B)} \quad \text{eq (1)}
\]

We select peer users regardless of whether a peer and a target user are in a watching network or not. Once we select peer users based on the Jaccard similarity, we produce CF-based item rankings according to the aggregated prediction scores of the items in the top 10 peers’ collections, as following equation 2.

\[
CF_{ia} = \sum_{j=1}^{10} \text{sim}_{ij} \times w_{ja} \quad \text{eq (2)}
\]

\(CF_{ia}\) is a CF recommendation prediction score for user i about item a. It is calculated for every item a, which belonged to the collection of at least one of the top 10 peers’ collections (peer j) a, \(a \in A\). Of course, A collection excludes the items user i has. \(\text{Sim}_{ij}\) is the Jaccard similarity between the target user i and the peer user j. As explained, we take into account only the top 10 users, and hence the maximum number of peers in recommendations is 10 for every target user. \(w_{ja}\) represents whether peer j have the item a in his collection A. If he has the item, the value is 1, otherwise the value is set as zero.

A WN-based recommender engine developed for our study is solely based on watched users’ collections. We also use Jaccard similarity for the calculation and the same ranking approach as used in CF step, equation (2). The only difference is that we use watched users’ information instead of the top 10 peer users. In equation 3, \(WN_{ib}\) is WN-based recommendation for user i about item b. The variable k represents the watched users of user i and b is one of the items that user i’s watching networks have but the user i does not (\(b \in B\)). \(W_{ib}\) shows whether a watched user k has the item b or not. The variable \(n\) represents the number of the watched users.

\[
WN_{ib} = \sum_{k=1}^{n} \text{sim}_{ik} \times w_{kb} \quad \text{eq (3)}
\]

Lastly, our hybrid engine fuses both CF-based recommendations and WN-based recommendations. Specifically, we use a simple version of mixed hybridization approach like the following.

\[
H_{ic} = \sum_{p=1}^{n} \text{sim}_{ip} \times w_{pc} \quad \text{eq (4)}
\]

4. Experimental Evaluation

To assess the prospects of WN-based recommendations, we run a traditional \(n\)-fold cross-validation study with a data set.

4.1 The Data Source

As the data source, our study used a social tagging system, CiteULike. CiteULike is one of the leading systems for managing and sharing bibliographic references. CiteULike users can add interesting articles to their CiteULike library in two ways. As one way, they are able to add an article by themselves using booklet, uploading a citation file or entering bibliography manually. As another way, they are able to copy interesting items from other users’ libraries. CiteULike has a well-developed ‘watchlist’ function. Once they perceive that other users have interesting or useful set of items, they can watch the users. Then all items (i.e., references) assembled by the ‘watched’ users are displayed to the ‘watching’ users as a ‘watchlist’. In the watchlist, the watching users can see not only the bibliography information but also the annotated tags.

The data set for this study was collected using snowball sampling [1]. To choose initial set of users, we randomly visited CiteULike in September and October, 2008 and July and August of 2009. We chose the users who posted new resources to the site at the time of visit. The information collected for each user included the bibliography (article title, list of authors, journal name, publication year, etc.), the tags, the posted date and time and the
watchlist. After collecting a group of initial users, we collected data of their watching users through breadth-first search. To diversify data collection, all of the other users who have the same articles were chosen, additionally. Watching connections of these users were crawled as well. Table 1 shows the descriptive statistics of the data set. For the detailed characteristics of this data set, refer to [18].

<table>
<thead>
<tr>
<th>Total no. of users</th>
<th>19958</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total no. of distinct items (papers)</td>
<td>1070389</td>
</tr>
<tr>
<td>Average no. of items per user</td>
<td>64.75</td>
</tr>
<tr>
<td>Total no. of distinct tags</td>
<td>233816</td>
</tr>
<tr>
<td>Average no. of tags per user</td>
<td>190.75</td>
</tr>
<tr>
<td>No. of Users who watch other users</td>
<td>1826</td>
</tr>
<tr>
<td>Total no. of unilateral relations</td>
<td>11295</td>
</tr>
</tbody>
</table>

### 4.2 The Formal Evaluation

We used our CiteULike data set to perform a traditional formal \(N\) fold cross validation (\(N=5\)). The idea of the cross-validation approach is to assess the quality of a recommendation approach by its ability to predict bookmarks already made by the users. As the test or target users, we randomly chose 410 users who have watching networks. Table 2 shows the description of the test users.

| Average no. of items per test user | 294.8 |
| Average no. of tags per item assigned by test users | 745.1 |
| Average no. of unique tags assigned by test users | 124.4 |
| Average no. of unilateral relations that a test user has | 5.6 |

In our 5-fold cross validation, the information collection of each test user was randomly partitioned into 5 equally sized subsets. Each one subset was then used as the test set and the other four subsets were used as the training set to predict recommendations in the test set. This procedure was run 5 times, and the results of 5 runs were averaged to produce the final accuracy. The final results were assessed through evaluation metrics for information retrieval: precision and recall. Precision aims to measure how precise the prediction is and recall aims to measure how complete the prediction is. More specifically, precision at point \(N\) (precision@\(N\)) is the ratio of the number of correctly predicted items in the top-\(N\) list to \(N\) (eq. 6). Recall at point \(N\) (recall@\(N\)) is the ratio of the number of correctly predicted items in the top-\(N\) list to the total number of relevant items (eq. 7) [25, 31].

\[
\text{precision@}N = \frac{\text{No. of correct prediction}}{N} \quad \text{eq. (6)}
\]

\[
\text{recall@}N = \frac{\text{No. of correct prediction}}{\text{size of test set}} \quad \text{eq. (7)}
\]

The 5-fold cross validation was used to compare the quality of two approaches (i.e. watching network-based and hybrid recommendations) discussed in the previous section including CF-based recommendations as a baseline. We compare precision and recall at five points taking into account the top 1, 3, 5, 10 and 20 recommendations.

### 5. RESULTS

In this experimental study, we studied whether users’ unilateral relations help to improve CF-based recommendation quality. First, Table 3 shows the results of precisions. Even though our recommendation algorithm was simple, we were able to predict more than 20% correct recommendations for top 1 and top 3 assessment points. For all 5 levels of examinations, hybrid recommendations (\(H_{\text{all}}\)) always delivered higher precisions than CF (\(C_{\text{all}}\)) or WN-based recommendations (\(W_{\text{n}}\)). Based on One-way ANOVA test, we found the significance of the mean difference (refer to the bottom row of Table 3) for precision of all top \(N\). We found the same results in recalls (refer to Table 4). Because the results of hybrid approach were the largest for the all cases, we did not count the harmonic score of precision and recall, F-measure.

As explained in Section 3, we produced recommendations according to the prediction probability values. We tested the mean difference of these probability values between the correct predictions and the wrong predictions using Kruskal-Wallis H T-test. In the test using the precision of top 20, there was the significant difference (\(p < .001\)) between the correctly predicted cases (\(M = 0.42\)) and the wrongly predicted ones (\(M = 0.06\)). That is to say, the higher the precision probability was, the more likely the recommendation was correct, and vice versa.

<table>
<thead>
<tr>
<th>Table 3. Precision Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 1</td>
</tr>
<tr>
<td>(C_{\text{all}})</td>
</tr>
<tr>
<td>(S_{\text{all}})</td>
</tr>
<tr>
<td>(H_{\text{all}})</td>
</tr>
<tr>
<td>Sig.</td>
</tr>
</tbody>
</table>

* indicates \(p < .005\)

<table>
<thead>
<tr>
<th>Table 4. Recall Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 1</td>
</tr>
<tr>
<td>(C_{\text{all}})</td>
</tr>
<tr>
<td>(S_{\text{all}})</td>
</tr>
<tr>
<td>(H_{\text{all}})</td>
</tr>
<tr>
<td>Sig.</td>
</tr>
</tbody>
</table>

* indicates \(p < .005\)
There is any distinct characteristic the group of users whose results point, we wondered what makes these cases produced the ones generated equivalently. Except 77 users whose best cases are CF, for most of users F predicted the exactly same number of correct recommendations. There were some cases more than one approaches produced the same results. F there were limited recommendations for all the test users. We used the precision results of the top 5. The Error! Reference source not found. shows the results.

Table 5. The Number of Test Users for Whom Each Approach Produces the Best Precision Results (the precision of the Top 5)

<table>
<thead>
<tr>
<th>The Best Case</th>
<th>No. of Test Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>CF(\text{ui})</td>
<td>77</td>
</tr>
<tr>
<td>WN(\text{ui})</td>
<td>20</td>
</tr>
<tr>
<td>H(\text{ui})</td>
<td>43</td>
</tr>
<tr>
<td>CF(\text{ui}) + WN(\text{ui})</td>
<td>168</td>
</tr>
<tr>
<td>WN(\text{ui}) + H(\text{ui})</td>
<td>33</td>
</tr>
<tr>
<td>Totally Incorrect</td>
<td>69</td>
</tr>
</tbody>
</table>

There were some cases more than one approaches produced the same results. For 168 users, CF and hybrid approach produced the same results. Said differently, for these users, CF and hybrid predicted the exactly same number of correct recommendations. For 33 users, WN and hybrid approaches produced the same best results. Except 77 users whose best cases are CF, for most of users (81.22% of test users), WN-based or hybrid recommendations generated equivalent or better recommendations. However, not only the recommendations utilizing watched networks, but also the ones partially or wholly powered by CF recommendations produced good results for sufficient number of people. At this point, we wondered what makes these cases different and whether there is any distinct characteristic the group of users whose results belong to each case have.

Table 6. The Summary Info of Each Best Case Group

<table>
<thead>
<tr>
<th>The Best Case</th>
<th>Avg. No. of Items</th>
<th>Avg. No. of Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>CF(\text{ui})</td>
<td>366.14</td>
<td>7.40</td>
</tr>
<tr>
<td>WN(\text{ui})</td>
<td>732.70</td>
<td>3.60</td>
</tr>
<tr>
<td>H(\text{ui})</td>
<td>409.93</td>
<td>7.44</td>
</tr>
<tr>
<td>CF(\text{ui}) + WN(\text{ui})</td>
<td>260.54</td>
<td>5.41</td>
</tr>
<tr>
<td>WN(\text{ui}) + H(\text{ui})</td>
<td>317.67</td>
<td>4.85</td>
</tr>
<tr>
<td>Totally Incorrect</td>
<td>89.25</td>
<td>3.54</td>
</tr>
</tbody>
</table>

Table 6 is the summary information about the people whose results belong to each case. Obviously, the users who did not receive any correct recommendation (users in ‘totally incorrect’) have the fewest items and the fewest networks. The users whose best cases were CF or hybrid seem to have similar number of items and similar number of networks. On the other hand, the users who got the best results in WN-based recommendations tend to have much larger number of items but smaller networks. However, this interpretation is just an approximation. So as to examine what make these cases different, we divided the test users into four groups according the number of their bookmarked items and the size of the watching networks like the following Figure 4. The division was made by the mean number of items (\(M \approx 295\)) and watching networks (\(M \approx 6\)). That is to say, when a test user has 210 items (< 295 items) and 5 networks (< 6 networks), his group is ‘A’. If another test user has 295 items (>= 295 items) and 20 networks (>= 6 networks), he fits into group ‘D’. There were 226, 61, 81 and 42 users for each group A, B, C and D. In order to test how these four groups correlate with their best cases, we counted the number of users who fit into one of four groups for every best case.

6. CONCLUSION AND DISCUSSION

In the context of social tagging systems, we investigated whether users’ self-defined networks can be a part of recommendations and improve the traditional collaborative filtering approach. We specifically compared the quality of the traditional CF recommendation with an approach based on watching network, which employed the information about watched users’ collections. We also compared the quality of a hybrid approach, which combined the traditional CF and the watching network-based approaches. The recommendations were compared using 5-fold
Figure 5. Division of User Groups according their Best Cases

cross validation on a crawled CiteULike data set. Measured by precision and recall, the hybrid recommendations generated the best results. When we counted the number of cases yielding the best result for individual users, CF recommendations produced good results for a large number of users. However, conclusively the results also showed that the recommendations taking advantage of watching networks help CF-based recommendations improve the quality of recommendations for many users or generate better recommendations than CF approach for some users.

We interpreted this result to mean that users’ self-defined watching connections, which were mainly based on similar interests and the users’ perceived usefulness, were important source to acquire good information and should be a part of personalizing their information. They can also reinforce the existing CF-based recommendation quality.

In the future, we will evaluate the WN-based recommendations through other viewpoints, for example, novelty, serendipity, or diversity. We also plan to develop a more elaborated recommendation algorithm. According to our previous study [17], we found that semantically rich information, for instance, tags or metadata, is a better measure of similarity than rather than standard item-unit based similarity comparison. Therefore, metadata and tag-based algorithm will be considered.

7. REFERENCES


