Interest Similarity of Group and the Members: the Case study of *Citeulike*

Danielle H. Lee & Peter Brusilovsky
School of Information Sciences, University of Pittsburgh
135 N. Bellefield Ave., Pittsburgh, PA 15260, USA
(hyl12, peterb)@pitt.edu

ABSTRACT

In this Web 2.0 era, many social Web systems support group activities. Groups are centered on the utility of information usefulness. Users join the same group on the Web because they are interested in the same topic in terms of a community of interest or practice. Herein, we examine the information similarity in self-defined group networks and specifically address not only the similarities between the same group members, but also the similarities between a group and the members. Our study found that a pair of users who are the members of the same group share significantly higher similarity in their personal collection than other pairs who are not members of any of the same groups on all explored levels (items, metadata, and tags). Especially, the degrees of similarity on the metadata and tag levels are much larger than the item similarity. The degree of the similarities between a group and the members, however, is much higher than the similarities of the same group members. More than 40% of all users have collections which are at least 50% overlapped with their group's collections. These results show that group is good source of information, but each member has his own specific information needs and it is rarely similar to other members. Another interesting property of information-sharing in group-based networks is that the number of groups that a user joined has significantly positive correlation with the size of their personal collection. Lastly, some members play an active role in introducing interesting information to their groups and further, some other members were perfectly influenced by the group collection (100% matches with the group collection). Overall, our findings support that groups could be feasible for guiding users to useful information.

1. INTRODUCTION

The Web 2.0 era, where users play not only a role of information consumers, but information creators, has produced very complicated online landscape, consisting of information items and users collected by various ways explicitly and implicitly. A number of researchers are now focusing on understanding this landscape, discovering its connections with real life, and building practical application on the basis of these discoveries. Among these topics, one that has attracted considerable attention is the correlation between user connections in the Web world and their similarities in real life. Recent research has demonstrated, for example, that users engaged in active forum discussion have more similar interests than non-connected users and users who exchanged instant messages frequently have more similar search queries than random pairs. Unlike other studies focused on friendships, this study considers users' group activities as social networks. Group activities are centered on the utility of information usefulness. Users join the same group on the Web because they are interested in the same topic in terms of a community of interest or practice. The relationships in group networks are known to be self-organized by the members and aim to distribute topic-relevant information or contribute related activities. Surprisingly enough, however, the information similarity of self-defined user's group was not yet explored. Most studies about the group dynamics and the information sharing patterns in groups have focused on derived communities which are discovered systematically by pattern mining approaches. This study aims to explore users' self-defined groups.

In the following study, specifically we explore the information sharing patterns on a social tagging system, *Citeulike*. The patterns are considered from two view points; community members and community per se. First, from community member's point view, we focus on how the information similarities are different between a pair of users who are the members of a same group and another pair of users who are not. This difference is examined according to several information levels – information items, authorship-based metadata, and tags. Secondly, from community's point view, we explore how much the information of a group's collection is similar to the information in the members' personal collections.

2. RELATED WORK

Backstrom, et al. (2006) tried to answer these questions – which social factors influence a user to join a group and what makes the group thrive – using decision tree technology. LiveJournal and DBLP were the data sets and the authors counted the number of a user's friends who are already members of a group and examined whether the large number of friends in the group proportionally increases the chance for the user to join the group. As the results, the number of friends who already participated in a group has little correlation with the user's possibility to join the same group. When a user's friends who already were in the group befriended each other, however, the user was significantly liable to join the group. They interpreted this result as the users with their friends in a group make the group trustworthy and information-advantageous. Additionally, group having less triad networks among the members tended to grow better than the groups consisting of more triad networks. They suggested that the triads may be the equal of cliqueness, which prevented active community growth. Groups in this study were explicitly defined. However, the information sharing patterns in group members or the patterns between groups and the members were not explored (Backstrom, Huttenlocher et al. 2006).

Zhou and the colleagues studied the information similarities in groups using semantic-rich contents. In the study based on Enron email corpus, as the first step, they ran the Bayesian network and chose the latent topic of emails. Then, the correlations between an email and the associated users (i.e. the author and the recipients) were taken into account. Particularly, they suggested two models to extract communities – one model was centered on each user's contacts and another was centered on topic. When they compared the resultant communities with the group formation from another study as a ground truth, they found that their approach succeeded to generate appropriate groups with high similarity in shared messages (Zhou, Manavoglu et al. 2006). The weakness of this study is that the groups were inferred by machine learning technology, which was based on content similarities.

The existing studies about group dynamics have largely concerned about the interactions only between/among group members or about the derived groups inferred by various machine learning technologies (O'Hara, Alani et al. 2002; Backstrom, Huttenlocher et al. 2006). In what follows, we focus on users' self-defined group activities and explore not only the information sharing dynamics among group members, but also interactions between a group and the group members.

3. THE DATA SET

3.1 The Data Source and the Relationship

As a source of data for our study we selected a collaborative tagging system, *Citeulike*. Along with Bibsonomy (Hotho, Jäschke et al. 2006) and Connotea (Lund, Hammond et al. 2005), *Citeulike* is one of the leading systems for managing and sharing bibliographic references. As many other collaborative tagging systems, *Citeulike* supports group activity. Users can create a group, join existing groups, or be invited to join the group. When group members find interesting references, through the *Citeulike* interface, they are able to add them not only in their personal repositories, but also in the group space with tags at the same time. The updated list of references is shown to all other group members. The group members are able to copy references on the group collection to their personal repositories, as well.

3.2 Data Collection

We collected the group data from *Citeulike*. As the first step, we visited the site in October and November of 2008. As of the time when we visited, there was a page showing the list of groups. We chose all groups that were displayed on the page at the time of the visit and collected the groups' collections, the group members and the members' personal collections. The information of each group' collection included the bibliography (article title, list of authors, journal/conference names, publication years, etc), the tags, and the posted date and time. We collected the same kind of information from individual group member's collection. Out of more than 700 groups, we filtered out single-member groups, groups having insufficient references (n < 5), and members who do not have any reference in their personal collection (n = 0). Then the total number of groups was 619 and these groups have 337,987 distinct items (i.e. research papers). We had 2643 users and they made 3528 memberships as total. Each user is a member of 1.34 groups and each group has 5.7 members on average. Table 1 and Figure 1 and 2 show the summary of data set and the data distribution in groups and group members' collections. The both figures display that the users and groups in our data set may have enough number of items to compare the information sharing patterns. Figure 3 shows the number of groups that each user participates in and displays that most of users are members of one group.

Table 1. Data Summary of *Citeulike*

Total No. of Groups	619
Total No. of Users	2643
Total No. of Group Memberships	3528
Average No. of Group per User	1.34
Average No. of Members per Group	5.70
Total No. of Unique Items	337987
Average No. of Items per Group	445.89
Average No. of Items per User	188.24
Average No. of Tags per Group	1039.31
Average No. of Tags per User	464.40

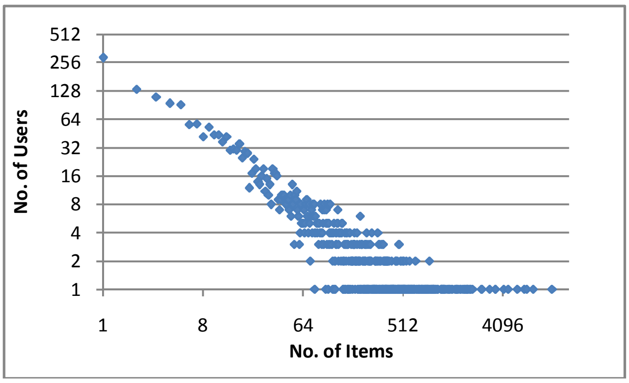

Figure 1. Distribution of Group Members' Information Collection

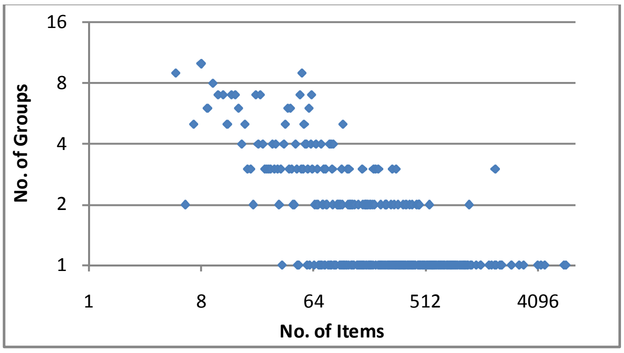

Figure 2. Distribution of Groups' Information Collection

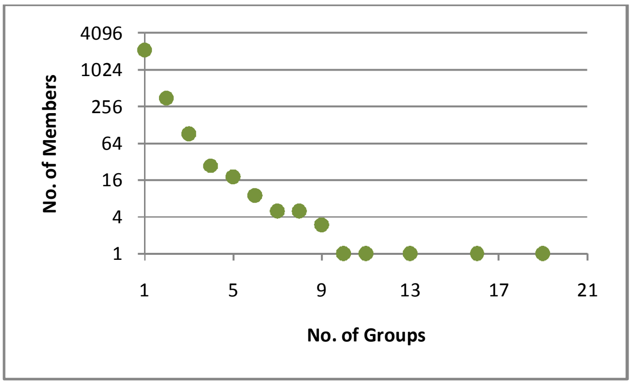

Figure 3. Distribution of Group Memberships per User

4. DATA ANALYSIS

The goal of this study is to explore the information sharing patterns between group members or the patterns between a group and the members. Especially, we are interested in the similarity of shared information on four levels - information item, metadata, and macro and micro tags level similarity.

First, item level similarity measures the number of common items (i.e. articles) between two group members' collection or between a group and one of the group members' collection. This item similarity is the most fundamental unit of measurement. Second, we take into account metadata as a way to measure the similarity beyond the item level. Due to the irregular opportunistic nature of the bookmarking process, users with similar interests may not necessarily end up with very similar collection. Therefore, we compare the users' interest similarities using metadata. Since the information items in *Citeulike* are bibliographic references, the authorship is taken into consideration as metadata. For instance, two members may have two different papers written by one author. This indicates that they are having similar interests even though they do not share exactly same item. Since the *Citeulike* users are able to navigate articles by clicking author's name, we considered that the authorship metadata may be an important way for the users to find interesting papers.

Tag similarity was assessed by counting the number of shared tags on two levels: *micro* level and *macro* level. On micro-level a tag was counted as shared if it was used by both users to tag the same common information item. The rationale behind this approach is that if two users annotate the same tags on the same item, they understand that item as a similar meaning because tags are cognitive expression showing how users comprehend one item with different viewpoints (Hung, Huang et al. 2008). Lastly, when two users do not share many identical information items but share many identical tags, they could be closely related. Therefore, we explored macro-level tag similarity, which counted common tags used by both users regardless of the tagged item.

4.1 Dependent Variables

Since the sizes of items and tag collections varied dramatically from member to member or from group to group, we examined not only absolute numbers (i.e. raw number of common items, metadata, or tags) but relative (normalized) measurements. Specifically, we used two different sets of dependent variables for the comparison between group members and the comparison between a group and the members.

For the calculation of information similarity between two group members, we used the Jaccard similarity coefficient – the portion of shared items in both members' union set (refer to eq. 1) – as an undirected relative measures (Guy, Zwerdling et al. 2009).

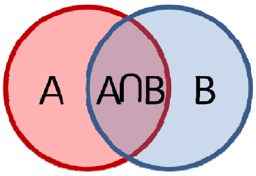

Figure 4. Information Overlap between Member A and B

$$\textbf{Jaccard Similarity Coefficient} = (A \cap B)/(A \cup B) \quad \text{eq. (1)}$$

For the information similarity between a group and each group member, we measured not only the Jaccard similarity coefficient, but the group and member fractions. The latter two variables measure the direction of influence. For example, user A is one of the members of group #1 and group #1 and user A have 450 items and 100 items, respectively. If there are 90 items in common, 90% of member A's collection is overlapped with the group #1's collection but only 20% of group #1's collection is covered by member A's collection. Depending on the way we counted the information overlap, the similarities are different. The *member fraction* (eq. 2) is the portion of shared information on the center of a group member. On the other hand, the *group fraction* (eq. 3) is the portion of shared information on the center of a group. For the above example, the *member fraction* of member A for the group #1 is 90% which is the portion of shared information in the user A's collection and it is the same value of *group fraction* of group #1 for member A. The *group fraction* of member A for group #1 is 20% and it is the same value with the *member fraction* of group #1 for member A. This relative similarity measures were counted for all levels, from item level and metadata level to micro and macro-level tags.

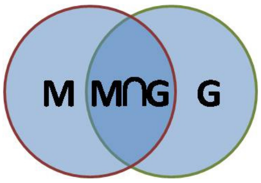

Figure 5. Information Overlap between Member and Group

$$\textbf{Member Fraction} = (\text{Member} \cap \text{Group})/\text{Member} \quad \text{eq. (2)}$$

$$\textbf{Group Fraction} = (\text{Member} \cap \text{Group})/\text{Group} \quad \text{eq. (3)}$$

5. THE RESULTS

5.1 Information Similarity between two Group Members

In the following section, we tested whether and how much two users who participated in the same group share common information. Since group activity is based on similar interests, we assumed that their personal collection may be similar enough to be a useful information source to each other.

First, we compared the absolute number of common items. Upper two rows of the Table 2 show the mean numbers of common information items between the same group members' pairs and between random pairs. The Mann-Whitney non-parametric test was used to assess the significance of the mean differences. The two users who are in the same group (M = 0.75) shared significantly larger number of common items than random pairs (M = 0.02). The same results were observed in the comparison of relative similarity measures.

In the comparison metadata (common authors), the absolute numbers of common metadata in group member pairs (M = 7.78) were almost 3 times larger than that of the random pairs (M = 2.77) and we also found the same results in relative similarity powers. These results are statistically significant (as described on the lower part of the Table 2).

Table 2. Difference of Shared Items and Metadata

		Members of the Same Group	Members of the Different Groups
Items	Absolute Numbers	**.75**	**.02**
		Mann-Whitney U = 123.0, *p* < .001	
	Relative Measures	**0.28%**	**0.01%**
		Mann-Whitney U = 139.5, *p* < .001	
Metadata	Absolute Numbers	**7.78**	**2.77**
		Mann-Whitney U = 9.3, *p* < .001	
	Relative Measures	**0.70%**	**0.16%**
		Mann-Whitney U = 15.5, *p* < .001	

As the next step, we compared the similarity in two kinds of tags – macro-level tags and micro-level tags. The results of these tags were comparable to the results of items and metadata. The members of the same group shared significantly larger macro-level and micro-level tags than the random pairs.

Table 3. Difference of Shared Macro-Tags and Micro-Tags

		Members of the Same Group	Members of the Different Groups
Micro-tags	Absolute Numbers	**.37**	**.00**
		Mann-Whitney U = -123.0, *p* < .001	
	Relative Measures	**0.07%**	**0.00%**
		Mann-Whitney U = -139.7, *p* < .001	
Macro-tags	Absolute Numbers	**3.98**	**.77**
		Mann-Whitney U = -9.3, *p* < .001	
	Relative Measures	**2.0%**	**0.32%**
		Mann-Whitney U = -15.5, *p* < .001	

Although the same group members shared significantly larger amount of information than the random pairs who were not in any common group in all explored levels, the amount is trivial (i.e. 0.29% of items, 0.83% of metadata and 0.86% of macro-tags). Said differently, even though each member of the groups have sufficient amount of information in their personal collection (M = 251.53 on Table 1), he/she shared a very little information with his/her group members. We considered that this result may be related to the *Citeulike* interface. When a user posted an article to his group collection, the poster information is rather invisible since the font is small and the items in group collection cannot be selectively retrieved or sorted by the poster information. Another possible reason of this little overlap is that each member desired very specific information and failed find the right one from other members' collection. Even though the poster information is not shown very clearly, users are able to see the poster's personal collection, when they clicked the poster name. It means that users had a chance to refer what the poster had and to copy interesting items to their own collection. In what follows, in order to check whether this little overlap was caused by the interface problem or members' very specific needs, we compared the groups' collection and the members' collection. The system displays the group collection in the same format of members' personal, hence it is intuitive for users to navigate and refer to it. If there is large overlap with the group collection, the small fraction of similarity between members may be due to the interface. Otherwise, since members are seeking too detailed information and develop their own strategy of finding information, being a member of a group may be just a fruitless attempt to find information. We will check out the overlap between groups and the members' collection and whether group members are sufficiently similar to their group.

5.2 Information Similarity between a Group and the Members

In above section, we examined the information similarity among group members and found that each member had a tiny little portion of common information with other members. Will they share enough information with group? In this section, we investigate whether the information sharing pattern of the group and the members is different with the one of group members. Before the computation of similarity, we compared the collections of groups and the collections of the members. As you can reckon the difference, groups' collection with 445.89 items on average is significantly larger than the members' personal collections with 188.24 items on average (*Wilcoxon Z* = -31.43, p < .001). When several group members contribute to organize information in the group collection, this asymmetric proportion is natural. The similar suggestion can be made in the result of a correlation test. When we calculated the correlation between the number of group members and the size of group collection, there was significantly positive correlation (*r* = 0.22, *p* < .001) meaning that the more members a group has, the more items the group's collection contains. The group collection may be constituted evenly by the group members, not by only one or two leading members. We examined this even contribution of the members to the group collection later.

We also tested whether the number of group of which each user is part correlates with the size of their personal collections. There are significantly positive correlations (*r* = 0.18, *p* < .001) even though the correlation is relatively small. That is to say, the more groups a user participated in, the more information he collected in his personal collection. It seems that being a member of a group may be helpful to gather useful information. So as to investigate this idea further, we examined the similarities between a group and the individual member.

As the similarity test, we computed four different similarity values – the absolute numbers of common information, the member fractions, group fractions and Jaccard coefficients – for item, metadata, micro-tag and macro-tag levels. As shown in Table 4, nearly half of group members' personal collection (42.16% of items and 45.73% metadata) was overlapped with the collection of their groups. Out of 3,528 group memberships, 997 users did not have any common information item with their group. On the other hand, users whose item collection was 100% matched with their group's collection were 873 users. The members whose personal collection was at least 50% overlapped with their group's collection were more than 40% of all the users in data set (using item similarity, 40.70% of users and using metadata similarity, 44.20% of users). This is the interesting finding. People are much more similar to their groups but not their group members even though they participate in the same group. Specifically, rather than information items per se, the similarity of semantic level information such as metadata and macro tags is higher. We can interpret these results as groups or communities are good source to get interesting information but due to the inappropriate interface, users may be unable to see other members' collection. Figure 6 and Figure 7 display the distribution of absolute numbers of overlapped information. Both figures show that many users have large information overlap with their groups regardless whether it is about information items, metadata or tags.

Table 4. Portion of Shared Information between a Group and the Each Member in Each Member's Collection

	Absolute	Group Fraction	Member Fraction	Jaccard Similarity
Items	**48.87**	**42.16%**	**16.35%**	**11.51%**
Metadata	**167.05**	**45.73%**	**18.98%**	**13.13%**
Macro Tags	**42.22**	**51.10%**	**18.89%**	**9.80%**
Micro Tags	**160.86**	**38.06%**	**14.31%**	**7.03%**

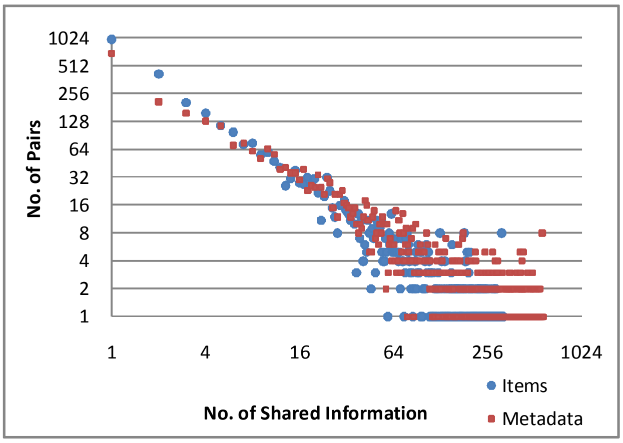

Figure 6. Absolute numbers of Shared Items and Metadata between a group and the members

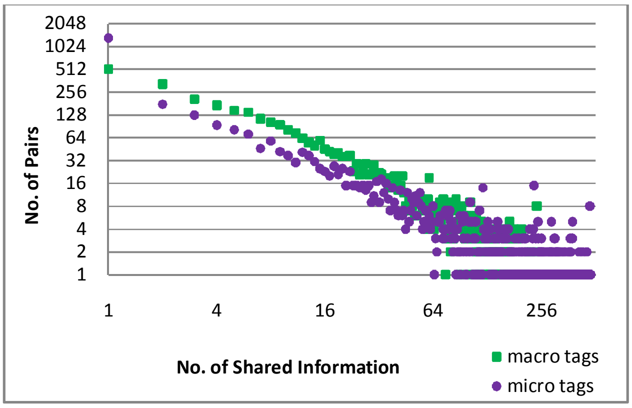

Figure 7. Absolute Numbers of Shared Micro Tags and Macro Tags between a group and the members

Error! Reference source not found. and Figure 9 show the distribution of the relative powers, especially member fractions from members' point of view. In these two graphs, we found interesting points. There are two distinctive peaks on the both extreme sides. A subset of members had the personal collection that was barely overlapped with their group's collection and another subset of members had the collection that was perfectly overlapped with their group collections. We investigated what make this difference by tracing the differences in the posting times of the common information. Many members whose collections were perfect match signed up the groups on earlier time and did post items to the group collections. Using the *Citeulike* interface, users are able to add interesting items not only to their personal collection but also to the group collection simultaneously. That is to say, they are active contributors or aggregators who are leading the information dissemination. There was also another interesting observation about perfectly matched users. Some users were just highly influenced by their group collection. For instance, the 'group #2' has 567 items in group collections and 28 members. Out of 10 members whose personal collections were 100% matched with the group collection, member A has 159 items and member B has 69 items in personal collection respectively and these items were all in the group's collection. However, we couldn't find any evidence showing that they had posted any item to the group. For the users who were active to disseminate information in there group collections, they tend to participate in the group in the early days and contribute to forming the collection. For another kind of users who were highly influenced by group are prone to join the group in the later time when the group collected abundant amount of information. Unfortunately, we failed to find constant patterns of behaviors in the members who had no common items with the groups. However, the number of items that zero-overlapped users had (M = 104.58) was significantly larger than the members who had perfect overlap with group (M = 30.70, t = 4.18, p < .001). We interpreted this result as the rich users have their own strategies to find useful information and inclined not to rely on somebody else or since they have amassed abundant information, they did not look for another information source. In addition, these zero-overlapped users could have very specific information needs, since they did share little information either with the group or with the members.

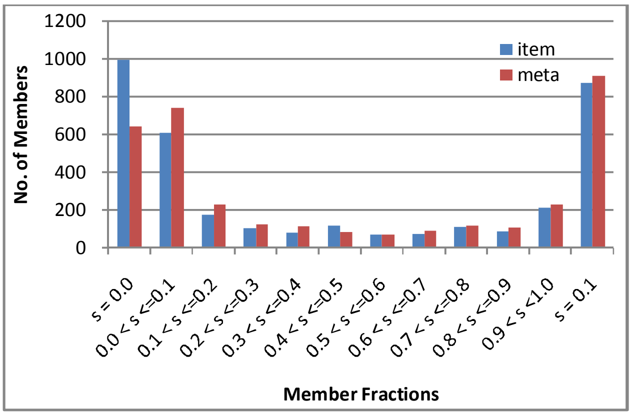

Figure 8. Member Fractions of Shared Items and Metadata (from Members' Point of View)

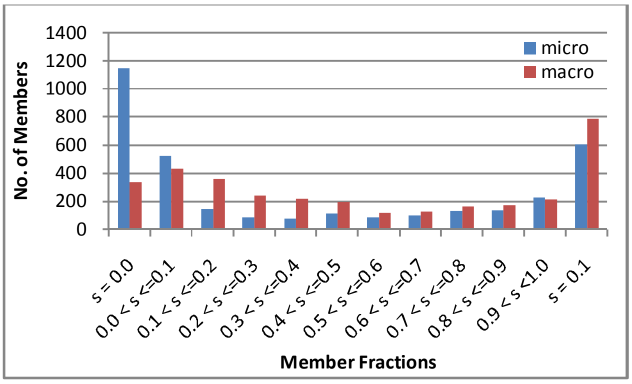

Figure 9. Member Fractions of Shared Micro Tags and Macro Tags (from Members' Point of View)

We also considered that two kinds of users on the both extreme sides could be made by the nature of the groups, and classified the groups into three categories – one having large portion of perfect matched users, another one having large portion of non-matched users, and the last one having relatively equal portion of these two extremes. We compared the number of members and the number of items in group collection for these three categories of groups and failed to find any significant results.

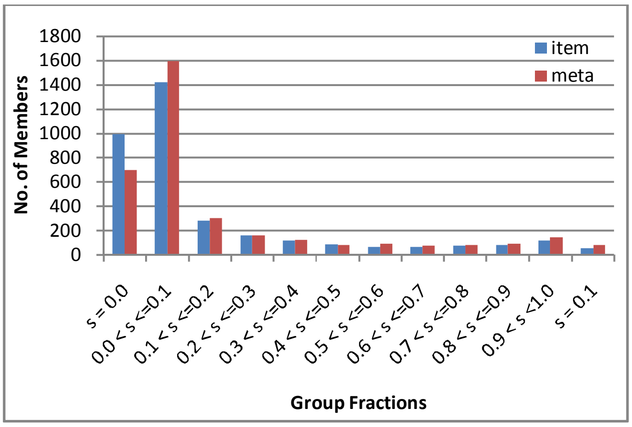

Figure 10. Group Fractions of Shared Items and Metadata (from Groups' Point of View)

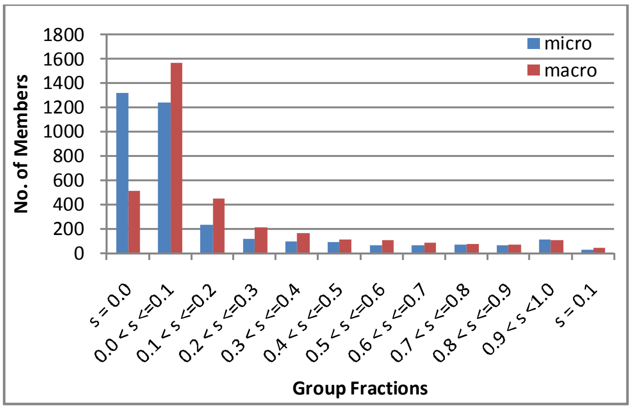

Figure 11. Group Fractions of Shared Items and Metadata (from Groups' Point of View)

As the last analysis, the view point of groups' side was taken into account. 16.35% of items and 18.98% of metadata in group's collection are overlapped with the members' personal collection on average. In addition, the number of group members whose personal collection contains all the items of the group (100% group fraction) is just 64. Put differently, this 100% overlap of group collection means, for example, if a group has 50 items in the group collection and one of the members, user 'A' has the all 50 items in his collection. 469.22 members have more than 50% of the group collection. As aforementioned, the groups' information space is larger than the members' personal spaces; hence the portion of overlapped information in groups' spaces is much smaller than the portion of the overlap in members' spaces.

6. CONCLUSION AND DISCUSSION

In this paper, we explored how much the information of my group and the information of my group members are similar with mine. We found that the information overlap between group members was significantly larger than the overlap between random pairs, but the amount of overlap was small. However, the information similarity between the group and the members were quite large. We saw that the little overlap between group members compared with the large overlap between the groups and the members may be caused by the interface problem of *Citeulike*. According to the result about the group and group member's information sharing patterns, there were two kinds of users – one kind was the people who take advantage of the group's information to the large extent and another kind was the people who just neglect the group information and try the information seeking strategy of their own with very specific information needs.

As the future direction, it is necessary to examine the timely change of information similarity and dynamics of memberships. Using different kinds of social networks such as friendships or unilateral relationships (as 'following' in twitter), it may be possible to see how personal similarities flows to the formation of groups or the similarity of a group to the members. In order to reinforce our findings in this study, we plan to add different data sets as well.

7. REFERENCES

[1] Backstrom, L., D. Huttenlocher, et al. (2006) Group formation in large social networks: membership, growth, and evolution. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Philadelphia, PA, USA.

[2] Guy, I., N. Zwerdling, et al. (2009) Personalized recommendation of social software items based on social relations. *Proceedings of the third ACM conference on Recommender systems*, New York, New York, USA.

[3] Hotho, A., R. Jäschke, et al. (2006). Information Retrieval in Folksonomies: Search and Ranking**:** 411-426.

[4] Hung, C.-C., Y.-C. Huang, et al. (2008). Tag-based User Profiling for Social Media Recommendation. *Workshop on Intelligent Techniques for Web Personalization and Recommender Systems, AAAI 2008*.

[5] Lund, B., T. Hammond, et al. (2005). Social Bookmarking Tools (II). *D-Lib Magazine* 11(4): 1-1.

[6] O'Hara, K., H. Alani, et al. (2002) Identifying Communities of Practice: Analysing Ontologies as Networks to Support Community Recognition. *In Proceedings IFIP World Computer Congress. Information Systems: The E-Business Challenge.*, Montreal, Canada.

[7] Zhou, D., E. Manavoglu, et al. (2006) Probabilistic models for discovering e-communities. *Proceedings of the 15th international conference on World Wide Web*, Edinburgh, Scotland.