Extracting user’s interest based on social bookmark tags

Junki Saito
Nagaoka University of Technology
1603-1, Kamitomiokamachi
Nagaoka-shi, Niigata, Japan
junkis@stn.nagaokaut.ac.jp

Takashi Yukawa
Nagaoka University of Technology
1603-1, Kamitomiokamachi
Nagaoka-shi, Niigata, Japan
yukawa@vos.nagaokaut.ac.jp

ABSTRACT

Social Networking Services (SNSs) have been gaining popularity in recent years. However, as the number of users of SNS increases, users are forced to spend a great deal of time in order to find a “good friend”. The user recommendation system is a good tool that solves this problem, but in that case, the system should understand the user’s interests. In the present paper, as a means of extracting a user interest, we construct a hierarchy of words from social bookmarking tags and emphasize characteristic words based on this relation. Then, we evaluate and demonstrate whether it is useful to extract an SNS user’s interest.

1. INTRODUCTION

Recently, web services called Social Networking Services (SNSs) have become popular and many people are using SNS to build social networks among individuals. Here, users are provided with an online space for interacting with real-life friends, acquaintances or other individuals who share common interests and/or activities.

Most SNSs are equipped with a function that users can freely form groups called communities. Communities are mainly formed as gathering places for users who have similar attributes. However, as the number of communities or the number of users in same community increases, users are forced to spend more time and energy to find “good friends”. Moreover, the difficulty increases further in an SNS that does not have a function to make an explicit community, such as Twitter. A user recommender system might be useful as a means of solving this problem. In this case, how the user interest is extracted is a major issue.

Collaborative filtering is widely used in Web page recommendation systems. This is often used in E-Commerce recommendation applications, and consists of the following two procedures [1].

1. Analyze the purchase history of the user and an access log, and extract user groups that have similar purchase and browsing pattern
2. Recommend products that are popular in each group

The method for estimating user interests according to such purchase histories is effective when that the suppliers recommend items to consumers on particular websites. This paper focuses on the recommendation of people who have similar interests. In this case, it is difficult to apply this method because the system is unable to collect sufficient histories.

On the other hand, when messages (diaries, comment, etc.) that the user has posted on the SNS are used as preference data, excerpting characteristic words is necessary. The characteristic word is a word indicating the user interest, and is weighted to reflect the semantic similarity to other words.

In the field of information retrieval, TF-IDF is a standard method for calculating the weights of words. However, TF-IDF computes the weight of each word individually. Therefore, among users who do not have same characteristic words, the similarity of users becomes zero even if the users have interests in similar areas. In addition to this method, if the relationship between words is evaluated, the system will understand the user interest with high accuracy.

Here, the semantic relation of words should be automatically generable from the text or words written by the user. In a social bookmarking service, each bookmark reflects the user interest and a tag is generated by the user at the same time. Thus, we expect that the semantic relation of a word can be extracted based on the co-occurrence relation of the tags in a bookmark.

In the present paper, we propose a method of constructing the hierarchical relation of words based on social bookmarking tags. Then, by emphasizing nouns using this relation, we extract the interests of SNS users.

2. TWITTER, FOLKSONOMY, AND SOCIAL BOOKMARKING

In the present paper, we investigated Twitter as an SNS. Twitter is not equipped with a function to make an explicit community compared with other SNSs and it is more difficult to find users with similar interests.

Social bookmarking services are being used by thousands of users every day. This is a web service using folksonomy, which is related to Semantic Web. As mentioned in the previous section, we propose to construct the hierarchy of words based on social bookmarking tags, which are briefly introduced in this section.

2.1 Twitter

Twitter [2] is a social networking and microblogging service. Twitter users can report their present situation, opinion, etc., by posting a short message of 140 characters or less. These short messages are called "tweets". Moreover, Twitter users can also 'chat' with other users.

By default, the tweets of each user can be viewed by the general public, and these tweets can be read when accessing the profile page of the user. In addition, new tweets of specified users can also be accessed in real time by registering the user as a friend. This registration action is called "follow".

2.2 Folksonomy

In traditional taxonomy, the meaning of an individual word is defined beforehand. This is a top-down type classification scheme in which specialists or the producers (senders) of information decide a classification system and classification words in advance. In contrast, folksonomy [3] is a classification scheme of the bottom-up type in which users (receivers) of the information perform classification themselves.

In other words, folksonomy can be described as a classification method that takes into account the concept that "a lot of people use a large amount of information". Concretely, the information classification and grouping are performed by providing several short words called "tags". Tags are not controlled words and so are freely given based on the vocabulary and the value judgment of each user.

2.3 Social Bookmarking

Social bookmarking (SBM) is a service for sharing information of Web pages that users register as favorites (bookmarks). When users create bookmarks, they can save the bookmarks with tags and comment, in addition to the page title and URL.

Since a tag is freely assigned in each bookmark, the same page is often expressed by different tags as shown in Figure 1. However, if the more users create the bookmark of a certain page, the tags which occupy a high ratio among those tags become what expressed the contents and the feature of the page more directly [4].

3. RELATED RESEARCH

Peter Mika [5] defined folksonomy as a tripartite graph structure that consists of Actor-Concept-Instance. He extended the traditional bipartite model of ontologies with the social dimension and demonstrated the possibility of building ontology based on folksonomy. As a case study, he constructed the ontology using tags from the bookmarks of del.icio.us. However, whether such a relationship between words is effective for interest extraction of SNS users has not been tested.

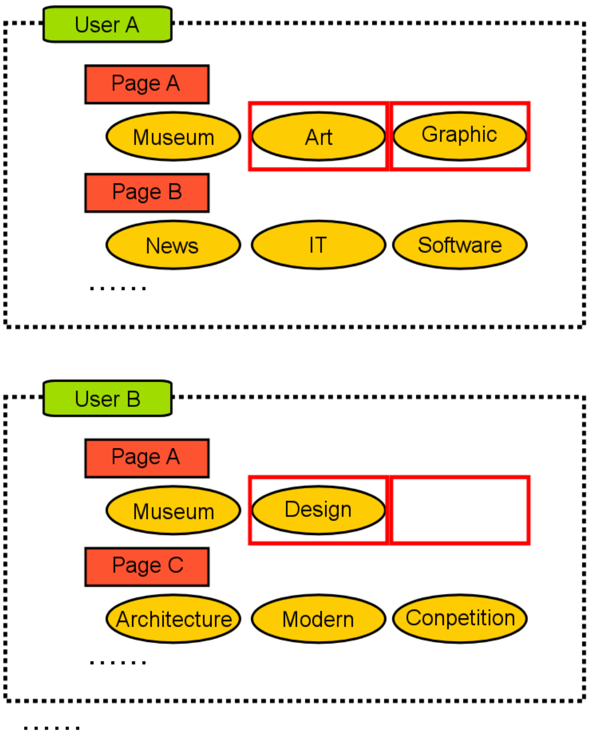

Figure 1: Tagging Example in SBM

Askhay Java et al. [6] analyzed the social network of Twitter and found that such networks have a high degree of correlation and reciprocity. They also considered the intention of the user and the structure of the community and demonstrated the possibility of categorizing a user who has a friend relation. Although they guessed the community of the user using the follower relationship, we estimate user interest using tweets and the hierarchy of SBM tags.

4. SYSTEM OUTLINE

The system proposed in the present paper constructs a word hierarchy and extracts Twitter user interests with the following procedure.

1. Record tags and tag pairs that appear in SBM, and then determine their frequencies. For example, in Page A of Figure 1, the "Museum" tag is counted twice, the "Art"/"Graphic"/"Design" tags are counted once, and the (Museum, Art), (Museum, Graphic), (Art, Graphic), (Museum, Design) combinations are counted one.

2. Calculate the degree of relation between each tag with MI-score, t-score, and G-score (log likelihood [7]), respectively. This is one of the indices for measuring the strength of both co-occurrences, MI-score between tags T_A and T_B is calculated by Eq. (1), and t-score is calculated by Eq. (2). G-score is calculated using Eq. (3) based on a two-by-two contingency table, as shown in Table 1. In each expression, N is the number

of tags in which the co-occurrence pair exists.

$$\text{MI-Score} = \log_2 \frac{freq(T_A \cap T_B) \times N}{freq(T_A) \times freq(T_B)} \quad (1)$$

$$\text{t-Score} = \frac{freq(T_A \cap T_B) - \frac{freq(T_A) \times freq(T_B)}{N}}{\sqrt{freq(T_A \cap T_B)}} \quad (2)$$

$$\text{G-Score} = 2 \times \sum_{i,j} O_{ij}(\log O_{ij} - \log M_{ij})$$

$$= 2 \times \left\{ a \log \frac{aN}{(a+b)(a+c)} + b \log \frac{bN}{(a+b)(b+d)} + c \log \frac{cN}{(a+c)(c+d)} + d \log \frac{dN}{(b+d)(c+d)} \right\} \quad (3)$$

Table 1: Contingency table for calculating G-Score

	Tag B	¬Tag B
Tag A	a	b
¬Tag A	c	d

3. Search and configure the upper-level tag of each tag, which has the highest relationship among all of the tags. The upper-level tag co-occurs more with various types of tags than the lower-level tags of the same category. As a result, the hierarchical categories of tags are constructed.

4. Collect the Twitter user ID, and obtain the user status (tweet count, user description [1], etc.) and recent tweets using the Twitter API.

5. Extract nouns from collected description and tweets.

6. Emphasize the weight of the noun in the description based on the hierarchical relation of tags and the appearance frequency of a noun in tweets. In the example of Figure 2, when the noun "MMORPG" is in the description, if the noun "Online Game" appears in a tweet, the weight of MMORPG will increase by *freq*("Online Game")/1. Moreover, if the noun "Nintendo" appears in a tweet, the weight of MMORPG will increase by *freq*("Nintendo")/3.

5. EXPERIMENTAL EVALUATION

In the extraction of the characteristic word, whether the hierarchical structure of words is constructed well is important. In this section, we evaluate in detail the layered structure of tags constructed by the method described in the previous section. We also demonstrate how effective the vocabulary is for the extraction of the interest of the Twitter user.

5.1 Data set

In the present study, for the Twitter data set, we collected the status and 200 most recent tweets of 4,161 Japanese users who appeared on the public timeline on July 19, 2010. In addition, we chose the data of a Livedoor clip [8] as the data set of SBM. This data set includes approximately 1.86 million tagged bookmarks and approximately 184,000 unique tags.

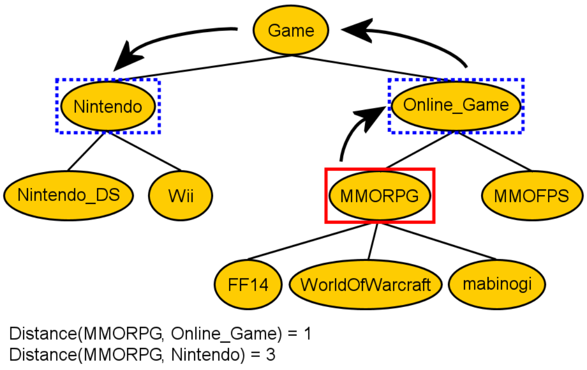

Figure 2: Emphasizing noun weight based on tag hierarchy

5.2 Results and Discussion

The number of tags per depth in the hierarchical relation of the tag built from the above-mentioned SBM data set is shown in Table 2. For each score, the top-layer (depth = 1) tag has the most frequency, and the tags in SBM are divided into a large number of categories. In particular, when MI-score is used as the degree of relation between tags, the tendency is remarkable, and between the tags in a same category, a relationship except parent and child or sibling hardly exists.

Table 2: Frequency of hierarchically structured tags

depth	MI	t	G
1	93,670	61,441	69,059
2	74,330	36,394	49,691
3	434	37,296	32,192
4	12	22,173	13,400
5	1	8,366	3,414
6	-	2,220	603
7	-	470	85
8	-	82	3
9	-	5	-

Next, the average size (number of tags) of categories is shown in Table 3, and the top five tag categories for each score are shown in Tables 4, 5, and 6. In Tables 4 through 6, the actual tags are written in Japanese.

[1] Primarily written in the self-introduction.

Table 3: Average sizes of tag categories

score	value
MI	1.798
t	2.742
G	2.438

Table 4: Top five tag categories (MI-score)

top layer tag	number of tags
movie	721
recruiting ad	399
contents	267
reputation	229
effect	215

Table 5: Top five tag categories (t-score)

top layer tag	number of tags
web	5,757
news	5,302
review	4,249
javascript	3,383
word of mouth	2,649

Table 6: Top five tag categories (G-score)

top layer tag	number of tags
review	1,790
web	1,655
society	1,351
javascript	1,143
recruiting	1,043

These tables reveal a significant difference in the size of the tag category for each score. For MI-score, since the average size of the tag category is less than two, the category is not sufficiently constructed. This is a problem with the characteristics of the MI-score formula. When the appearance frequency of a word is low, MI-score cannot compare co-occurrence relations appropriately.

On the other hand, although the sizes of tag categories are different in t-score and G-score, the tag categories of "web", "review", and "javasctipt" are highly ranked in both scores. Moreover, when the top 50 tag categories were compared, 41 tags were common between t-score and G-score. In contrast, 13 tags were common between MI-score and t-score and 15 tags were common between MI-score and G-score.

Figure 3 shows the calculated results for how the nouns and the proper nouns that appear in the tweets of each user were equated to tags in SBM. The "coverage" is the rate at which a noun in a tweet exists in SBM as a tag. In Figure 3, the coverage of 3,808 users exceeds 0.5, which corresponds to 91.5 percent of all users. As a result, an SBM tag covers a wide range of noun used in Twitter.

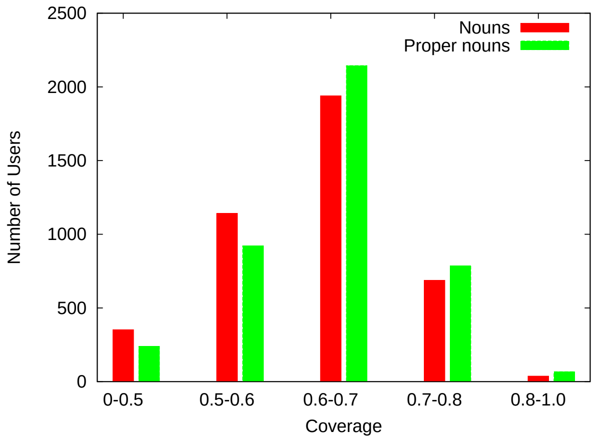

Figure 3: User distribution with noun coverage by tags

Whether the hierarchy of tags is effective in extracting of the interests of Twitter users is evaluated as follows. First, we randomly selected 300 users who were not bots from among the collected users. Second, we manually checked whether the extracted noun was actually related to the interest of the user. Tables 7 and 8 show the evaluation results. For the evaluation, we define "tag precision" as follows:

$$\text{tag precision } p_t = \frac{N_c}{N_e} \quad (4)$$

where N_c is the number of tags that are related to the interests of the user and appear in the user description, and N_e is the number of tags extracted from user description.

In addition, in Tables 7 and 8, its result is evaluated based on the following cases:

- case (A): list of words that perfectly match the SBM tags among nouns in the description
- case (B): list of words with weight down to top three among emphasized nouns in the description

Table 7: Precision of matched/emphasized nouns in description

case	tag precision
case (A)	0.607
case (B) by MI-score	0.618
case (B) by t-score	0.657
case (B) by G-score	0.628

Table 8: Ratio of users whose tag precision exceeds 0.5 or 0.8

	case (A)	case (B)		
		MI	t	G
$p_t \geqq 0.5$	0.733	0.747	0.747	0.757
$p_t \geqq 0.8$	0.327	0.383	0.420	0.397

Although the proposed method is very simple, for case (A), a precision of $p_t = 0.607$ was obtained. For case (B), which emphasizes weights according to hierarchical relations among tags, the highest precision ($p_t = 0.657$) was obtained using t-score. This is attributed to the tendency for users who are writing description describe their own interest by the enumeration of words like tag.

Furthermore, the characteristic word of the user was widely emphasized in t-score and G-score. On the other hand, in MI-score, the emphasis of the characteristic word was slight when the same noun did not appear in a tweet. Therefore, when collecting tweets continuously while guessing user interest, it is desirable to use t-score or G-score, rather than MI-score.

In the present paper, the tag categories are constructed based on only the co-occurrence frequency of the tags. In addition, if the tag categories are constructed by providing further semantic information for a tag, the accuracy should increase. Moreover, if two or more parental generation tags are set for each tag, the parental generation tags reduce disjuncture between categories and inhibit the influence of misclassification of the tag.

6. CONCLUSION

In the present paper, as a means of extracting user interest for the purpose of user recommendation, we proposed a method by which to construct the hierarchy of words based on SBM tags and to emphasize characteristic word by using this relation. We then evaluated the effectiveness of the vocabulary for the extraction of the interest of the SNS user. As a result of a survey on Twitter, we discovered that the tags in SBM and their hierarchy have a rich vocabulary for extracting the interests of Twitter users. Therefore, the proposed method is considered to be useful for realizing a user recommendation system. The proposed method is also applicable to general SNS if the description is replaced by the user profile and tweets are replaced by (blog) articles.

7. REFERENCES

[1] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 22(1):5–53, January 2004.

[2] Twitter. http://twitter.com/.

[3] A. Mathes. Folksonomies - cooperative classification and communication through shared metadata. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html, 2005.

[4] S. A. Golder and B. A. Huberman. The structure of collaborative tagging systems. Journal of Information Science, 32(2):198–208, April 2006.

[5] P. Mika. Ontologies are us: A unified model of social networks and semantics. In Proceedings of the 4th International Semantic Web Conference (ISWC2005), pages 522–536, November 2005.

[6] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pages 56–65, August 2007.

[7] T. Dunning. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(2):61–74, March 1993.

[8] Edge datasets. http://labs.edge.jp/datasets/.