LETTER

Extracting User Interest for User Recommendation Based on Folksonomy

Junki SAITO[†a)], *Nonmember and* **Takashi YUKAWA**[†b)], *Member*

SUMMARY In the present paper, a method for extracting user interest by constructing a hierarchy of words from social bookmarking (SBM) tags and emphasizing nouns based on the hierarchical structure (folksonomy) is proposed. Co-occurrence of the SBM tags basically have a semantic relationship. As a result of an experimental evaluation using the user profiles on Twitter, the authors discovered that the SBM tags and their word hierarchy have a rich vocabulary for extracting user interest.

key words: *user profiling, folksonomy, social bookmarking, SNS, Twitter*

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.

A function that allows users to freely form groups (communities) is provided in most SNSs. Communities are mainly formed as gathering places for users who have similar attributes. However, as the number of communities or the number of the users in same community increases, users are forced to spend more time and energy to find "good friends". 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.

This paper focuses on the recommendation of people who have similar interests. 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. In this regard, however when using only the existing lexicon (example: "Nihongo Goi-Taikei"), it is difficult to deal with an in-vogue word, a new word or an abbreviation that is used in SNS.

Hence, it is considered that the semantic relation of words should be automatically generable from the text or words written by the user. In a social bookmarking (SBM) service, each bookmark reflects the user interest, and a tag is generated by the user at the same time. Thus, in our previous work [1], the semantic relation of a word can be expected to be extracted based on the co-occurrence relation of the tags in a bookmark.

In the present paper, a method of constructing the hierarchical relation of words based on SBM tags is proposed. Then, by emphasizing nouns using this relation, the interests of SNS users are extracted for user recommendation.

2. Twitter, Folksonomy, and Social Bookmarking

In the present paper, Twitter is investigated as an SNS. As discussed in more detail below, Twitter is a microblogging service. Since the message which is posted by the user is very short, it is expected that the text is reflecting user interest strongly than the blog article that be seen in other SNS.

SBM 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, the authors propose to construct the hierarchy of words based on SBM 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".

Manuscript received December 28, 2010.
Manuscript revised February 5, 2011.

[†]The authors are with the Department of Electrical Engineering, Nagaoka University of Technology, Nagaoka-shi, 940–2188 Japan.
a) E-mail: junkis@stn.nagaokaut.ac.jp
b) E-mail: yukawa@vos.nagaokaut.ac.jp

DOI: 10.1587/transinf.E94.D.1329

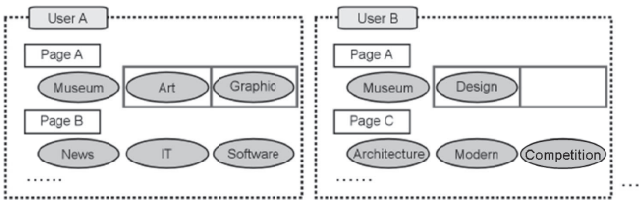

Fig. 1 Tagging example in SBM.

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 Fig. 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 [4].

3. Related Research

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.

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 words that generated from SBM tags.

Table 1 Contingency table for calculating G-Score.

	Tag B	¬Tag B
Tag A	a	b
¬Tag A	c	d

4. Constructing Word Hierarchy and Estimating User Interests

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.
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.

$$\text{MI-score} = \log \frac{freq(T_A \cap T_B) \times N}{freq(T_A) \times freq(T_B)} \tag{1}$$

$$\text{t-score} = \frac{freq(T_A \cap T_B) - \dfrac{freq(T_A) \times freq(T_B)}{N}}{\sqrt{freq(T_A \cap T_B)}} \tag{2}$$

$$\begin{aligned}\text{G-score} &= 2 \sum_{i,j} O_{ij}(\log O_{ij} - \log M_{ij}) \\ &= 2\Big\{a \log \frac{aN}{(a+b)(a+c)} + b \log \frac{bN}{(a+b)(b+d)} \\ &\quad + c \log \frac{cN}{(a+c)(c+d)} + d \log \frac{dN}{(b+d)(c+d)}\Big\}\end{aligned} \tag{3}$$

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 words are constructed from SBM tags.
4. Collect the Twitter user ID, and obtain the user status (tweet count, user description†, etc.) and recent tweets using the Twitter API.
5. Extract nouns from collected description and tweets if they exist as SBM tags.

†Primarily written in the self-introduction.

6. Emphasize the weight of the noun in the description based on the hierarchical relation of words and the appearance frequency of a noun in tweets. The weight of a noun n is calculated as follows:

$$w_n = \sum_{n_{rel} \in RelatedNouns} \frac{freq(n_{rel})}{1 + distance(n, n_{rel})} \quad (4)$$

In Eq. (4), each symbol represents the following meanings.

- $freq(n_{rel})$ is the appearance frequency of a noun n_{rel} in tweets.
- $distance(n, n_{rel})$ is the distance between a noun n and n_{rel} in the layered structure of words.
- $RelatedNouns$ are nouns that $distance(n, n_{rel})$ is three or less.

In this way, a user interest can be extracted by emphasizing the noun related to the genre in which especially the user is interested.

5. Experimental Evaluation of Interest Extraction

In the extraction of the characteristic word, whether the hierarchical structure of words is constructed well is important. In this section, the layered structure of words constructed by the method described in the previous section is evaluated in detail. Then, the effectiveness of their vocabulary for the extraction of user interest is also demonstrated.

5.1 Data Set

In the present study, for the Twitter data set, the status and 200 most recent tweets of 4,161 Japanese users[†] are collected. In addition, the data of a Livedoor clip [8] is chosen as the data set of SBM. This data set includes approximately 1.86 million tagged bookmarks and approximately 184,000 unique tags.

5.2 Results and Discussion

The number of tags per depth in the hierarchical relation of words 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 hierarchical relation of words 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.

Next, Fig. 2 shows the calculated results for how the 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 Fig. 2, 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.

Table 2 Frequency of hierarchically structured tags.

depth	MI-score	t-score	G-score
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	-

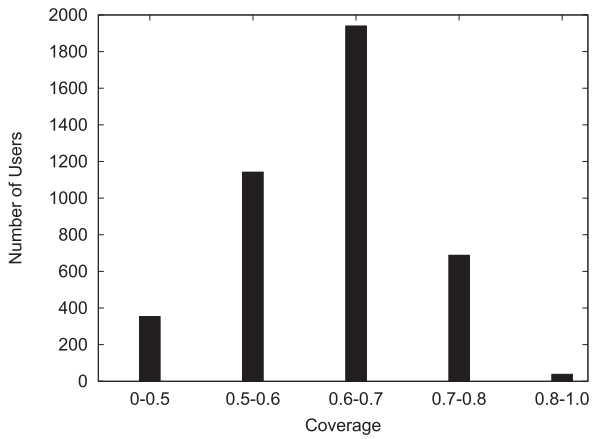

Fig. 2 User distribution with noun coverage by SBM tags.

Table 3 Interest precision with each score.

	MI-score	t-score	G-score
p_i	0.618	0.657	0.628

Table 4 User ratio whose interest precision exceeds 0.5 or 0.8.

	MI-score	t-score	G-score
$p_i \geqq 0.5$	0.747	0.747	0.757
$p_i \geqq 0.8$	0.383	0.420	0.397

Whether the hierarchy of words from SBM tags is effective in extracting of the interests of Twitter users was evaluated as follows. First, 300 users who are not bots were randomly selected from among the collected users. Second, whether the extracted noun was actually related to the interest of the user was manually checked. Tables 3 and 4 show the evaluation results. For the evaluation, "interest precision" is defined as follows:

$$\text{interest precision } p_i = \frac{|W_c|}{|W_e|} \quad (5)$$

where W_e are the words with weight down to top three among emphasized nouns in the description, and W_c are the words that are related to the interests of the user and included in W_e. W_c is the subset of W_e.

Although the proposed method is very simple, the

[†]They appeared on the public timeline on July 19, 2010.

highest interest precision ($p_i = 0.657$) was obtained using t-score. 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 word 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.

6. User Recommendation

Even if the system is able to extract the user interests correctly, whether the system can actually find a "good friend" as a recommended candidate should be confirmed. The validity about the proposed method in user recommendation is briefly investigated as follows.

1. Extract the interests of 9 Twitter users who cooperate with an experiment by the method previously described.
2. Extract the interests of collected users in the same way. In this experiment, for the Twitter data set, the status and 1,000 most recent tweets of 11,104 Japanese users[†] are collected.
3. Generate the interest vector whose components are the weights of each word in user description, and calculate the similarity of the interest vector between cooperators and collected users by cosine similarity.
4. Recommend top 20 candidates to the cooperators. If the cooperator feels that his/her own interests resemble candidate's interests, and cooperator hopes to make friend relation with candidate, the candidate is evaluated as an appropriate recommendation user.

The average of evaluation results by the cooperators is shown in Table 5. The highest precision 0.31 was obtained when t-score was used. On the other hand, the official user recommendation system based on friend relation is already available in Twitter. Those candidates were also evaluated by the cooperators for comparison, then the precision was about 0.14–0.20. Therefore, in the best case, the user recommendation system based on proposed method was about twice as effective as the official system.

[†]They appeared on the public timeline on October 12, 2010.

Table 5 Average of evaluation result.

	MI-score	t-score	G-score
precision	0.25	0.31	0.27

7. Conclusion

In the present paper, as a means of extracting user interest for the purpose of user recommendation, the authors proposed a method for constructing the hierarchy of words based on SBM tags and to emphasize characteristic word by using this relation. Then, the effectiveness of the vocabulary for the extraction of user interest was evaluated. As a result of a survey on Twitter, the authors discovered that the tags in SBM and their hierarchy have a rich vocabulary for extracting the interests of Twitter users. In the case of user recommendation system based on friend relations in the SNS, recommended candidates will most likely be limited to similar users. By using the proposed method in such a case, a user will feel more freshness to recommendation, since a recommended candidate will change according to the contents of the user-posted message on SNS. Therefore, the proposed method is considered to be useful for realizing a user recommendation system.

References

[1] J. Saito and T. Yukawa, "Extracting user's interest based on social bookmark tags," Groundbreaking Intelligent Systems Research Workshop on KES'2010, Sept. 2010.
[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, 2004.
[4] S.A. Golder and B.A. Huberman, "The structure of collaborative tagging systems," J. Information Science, vol.32, no.2, pp.198–208, April 2006.
[5] P. Mika, "Ontologies are us: A unified model of social networks and semantics," J. Web Semantics, vol.5, no.1, pp.5–15, March 2007.
[6] A. Java, X. Song, T. Finin, and B. Tseng, "Why we Twitter: Understanding microblogging usage and communities," Proc. 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pp.56–65, Aug. 2007.
[7] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," Computational Linguistics, vol.19, no.1, pp.61–74, March 1993.
[8] EDGE Datasets, http://labs.edge.jp/datasets/