IADIS International Journal on Computer Science and Information Systems
Vol. 7, No. 2, pp. 44-56
ISSN: 1646-3692

RELATIONSHIP STRENGTH ESTIMATION IN SOCIAL MEDIA USING FOLKSONOMY

Hidekazu Yanagimoto. *Osaka Prefecture University, Osaka, Japan.*

Michifumi Yoshioka. *Osaka Prefecture University, Osaka, Japan*.

ABSTRACT

We propose relationship strength in social media using folksonomy. We concretely estimate similarity between web pages in social bookmarking services using a tag vocabulary. A social bookmarking service is one of the most famous social media on the Internet. Since the web pages are selected according to users' interests but not their contents, in the services the classification of them is different from one in ordinary content-based approaches. In this paper we focus relationship strength estimation among web pages using a tag vocabulary. Avoiding a problem that tags includes some ambiguities among users, a similarity between web pages is defined in each user at first. The similarities are integrated over all users regarding the frequency of evaluation and a variance of the similarities. At last social bookmarking data are represented as a weighted network and it is easy to capture the relationships among all web pages registered in a social bookmarking service. To evaluate our proposed approach we carry out some experiments using real social bookmarking service, Buzzurl social bookmarking service. Then we confirm the proposed approach is superior to some comparative approaches and our proposed approach can capture more related web pages.

KEYWORDS

Social media mining, Graph analysis, Folksonomy, Bayes statistics

1. INTRODUCTION

A social bookmarking service is one of the most famous social media on the Internet and one of the useful information sources because of including various kinds of information and many persons' evaluating them. Social bookmarking service users register web pages according to their interests and share them among other users. Hence, interesting web pages are propagated in social bookmarking services and are shared among many users. Hence, registration frequency is one of the best criteria to evaluate web pages. Since a user usually has some interests, his/her bookmarking list includes web pages dealing with various kinds of topics. We

have to consider since the registration frequency depends on topic popularity directly, the frequency approach does not capture minor topic web pages appropriately. Moreover, some of the web pages are registered according to his/her temporal interest. For example, there are articles in news sites to read it later regardless of their interests. Such web pages worsen performance of web classification since the web pages are less relationship among other web pages from the viewpoint of their topics.

Many researches construct a network to analyze a social media. The network is analyzed using network mining approaches (Shi, 2000)(Tang, 2010). Analyzing the network appropriately, a precise network structure needs to be constructed. Many researches represent the relationship between nodes as binary value. For example, the value based on whether a pair of web pages deals with the same topic or not. However, the relationships should be represented as a continuous value. Representing strong or weak link between web pages improves social network analysis, community detection or link prediction.

Data registered in a social bookmarking service include various kinds of data types, for example text, audio, movie, image and so on. Though a similarity among the same data type is defined in a pattern recognition research, it is difficult to define the similarity between the different data types. Using tags in the social media the similarity is defined and we can overcome this problem since in this method we do not use content of the data.

The tag is a keyword attached to data and has no limitation in a vocabulary. Though the tag vocabulary does not have a strict vocabulary structure, a moderate structure emerges and is called “Folksonomy” (Mathes, 2004). The tag vocabulary includes a lot of synonyms and polysemic words and different description of words. This feature causes a pseudo relationship between web pages in calculating the similarity using the tags. Paying attention to a set of tags for an individual user, he/she does not use many synonyms and polysemic words since the tag is used to find some social bookmarking data fitting his/her interest. When a user uses some tags to denote the same concept, he/she cannot get all relative information exhaustively. On the other hand, when the user uses the same tags to denote the different concept, he/she cannot get only relative registered information. The different description of a tag generally denotes user's own purpose except a typographical error. Hence, tags that have different description should be processed differently.

In this paper we estimate a similarity between web pages in a social bookmarking service using users' tag vocabulary. The similarity is a continuous value and denotes strength of relationship. Avoiding some previous problems of tags, the similarity between web pages, which a user registers simultaneously, is defined in first step. In next step the similarities are integrated. Concretely, the integration of the similarities is formulated using Bayes theorem. This approach can introduce user's reliance and a priori knowledge for a pair of web pages in integration because of Bayes theorem.

The rest of the paper is organized as follows. We describe related works on social media analysis and their approach constructing a network in Section 2. In Section 3 we explain a similarity estimation algorithm. In Section 4 we carry out experiments using a real social bookmarking data and discuss the performance of the proposed method. In Section 5 we conclude and point out some future works.

2. RELATED WORKS

In this section we describe some related work on making a network from social media. We especially explain similarity estimation algorithms.

Now we think that social network services are represented as networks. In a social media a node is a user and link is a relationship between users (e.g. friend or not) generally. The relationship is estimated using logs which each social network services stores. Choudhury et. al. (Choudhury, 2010) estimates the similarity between users using communication frequencies of E-mail. The more E-mails users send and receive each other, the higher similarity they have. Horowitzand et. al. (Horowitzand, 2010) calculates the relationship between users based on usage of communication tools, E-mail and Instant Message, and user's profile. Fond et. al. (Fond, 2010) discusses a social influence between users and a time change of a social network. Especially they focus homophily (Wasserman, 1994) and approximate a social network in a smaller network. Kwak et. al. (Kwak, 2010) constructs a network from Twitter using “Follow” relationship. Xiang et. al. (Xiang, 2010) models the relationship strength in social network services using a latent variable model. In this approach latent variable, which denotes the similarity between users, is estimated from users' profile similarity and communication history.

Next we think that social bookmarking services are represented as networks. Rucker et. al. (Rucker, 1997) defines the similarity between web pages using cooccurrence frequency in the same bookmark folder. The web pages, which are classified into the same folder, deal with a similar topic each other. However, since in social bookmarking services users often use tags to classify web pages and the social bookmarking data have flat structure, it is difficult to apply this approach to social bookmarking data directly. Yeung et. al. (Yueng, 2007) creates a relationship between web pages using tags. The web pages where the same tags are added have relationships between them. Using tags users added in web pages, it creates a network reflecting contents of web pages. However, they reported that some pseudo relationships occur because of words of multiple meanings. For example, a tag “sf” means “science fiction” and “San Francisco”. Hence, using the tags to define the similarity between web pages, we have to deal with tags carefully.

3. SIMILARITY ESTIMATION

In social bookmarking services users determine whether web pages fit their interests or not. Hence, the social bookmarking data include user evaluation of the web pages. In this study we construct a weighted graph where nodes denote web pages in the social bookmark service, links denote relationships between the nodes, and weights denote similarities between them.

When we calculate the similarity using tags, we must pay attention to problems of tags. The problems are 1) the tags have many meanings and 2) the tags have description variation (e.g. ``read later" and ``*read later"). Hence, an approach using tags creates many pseudo links between web pages. Our proposed method consists of two steps: 1) Estimating the similarity between web pages in each users and 2) Integrating the similarities and finding a true similarity. Figure 1 shows this process. This approach can avoid some problems of tags.

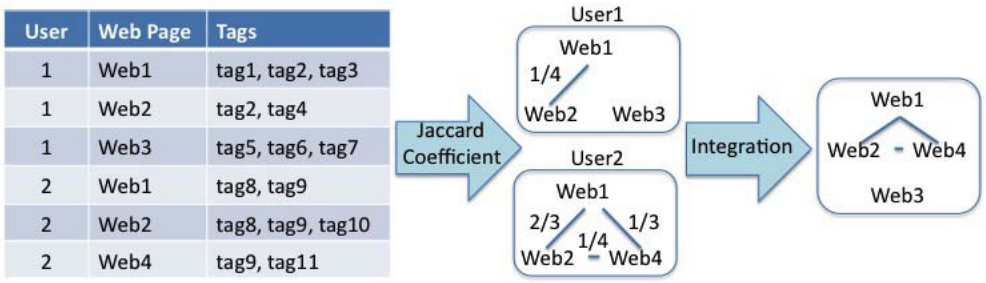

Figure 1. Flow of network construction from social bookmarking data.

3.1 Similarity Measure between Web Pages in Each User

In social bookmarking services tags are used for users to manage web pages and retrieve necessary ones from their social bookmarks easily. The tags have two aspects: 1) content representation of web pages and 2) users' intents. For example, “http://www.google.co.jp” has “search engine” and an article has “read later” for users' to read it later. The former is an example that a tag represents a content of a web page and the other is an example that a tag represents user's intent. The same tags are assigned in web pages dealing with the same topic. For example, “http://www.yahoo.co.jp”, has ``search engine", too. To estimate similarity between web pages using tags is a good strategy.

Tags in social bookmarking services are not in limited vocabulary and depend on personal usage. It is difficult to estimate a correct similarity using all tags attached in a web page. Let's discuss a tag that users often use for reading a web page later. Some users add "read later" and other ones add "*read later*". Though these tags are semantically the same ones, these descriptions are different. To overcome this problem tags need to be classified previously according to their semantics. This approach needs high-level natural language processing. Since in this approach in each user the similarity is calculated using the number of shared tag, previous problems do not happen. In user's social bookmark usage of tags are consistent and different description of tags represents different topics or user's different intent. If the usage of tags is inconsistent, it is a hard work for the user to find necessary data from his/her social bookmark. Because of this feature we do not need tag classification and high-level natural language processing. Since we use raw tags to calculate the similarities, we use no natural language processing before similarity estimation.

To define the similarity between web pages using their tags we use Jaccard Coefficient. Since Jaccard Coefficient is a similarity measure between two sets, we can use it to estimate the similarity using the tags. When a web page has some tags, Jaccard Coefficient is defined below.

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

S_i denotes a set and $|\ |$ is the number of elements in the set. When two sets are the same ones, Jaccard Coefficient $J(S_1, S_2)$ is equal to 1. On the other hand, when two sets have no intersection, it is equal to 0. In social bookmarking data T_{ki} denotes a set of tags that the kth

user added in a web page w_i. When a similarity between w_i and w_j is calculated in the kth user's social bookmark, the similarity $sim^k_{i,j}$ is defined below.

$$\mathrm{sim}^k_{i,j} = \mathrm{J}(T_{ki}, T_{kj}) = \frac{|T_{ki} \cap T_{kj}|}{|T_{ki} \cup T_{kj}|}$$

In calculating sim^k_{ij} a tag is a raw one without stemming to recognize difference of contents and users' intents.

In this step many networks is constructed based on each user's social bookmark. The number of constructed networks is equal to the number of all users. The number of web pages in a user's social bookmark is generally less than the number of all web pages in a social bookmarking service. Hence, each network represents a part of a network constructed from all the web pages. It is important to combine all the networks and make a whole network including all the web pages. Then we have to pay attention to overlapping the networks partly. In its intersection similarities between web pages are different because of users' different evaluation frequently. To regard this difference and estimate true similarities the similarity is integrated using Bayes estimation.

3.2 Integration of Similarity

A user can browse are only a part of whole web pages and it is impossible to read all interesting web pages. Hence, all web pages in social bookmark service are not represented as a network from one user's social bookmark even if the user registered the most web pages in the service. It is important to integrate each user's network and make the whole network. In integrating the networks the main problem is that the similarity between web pages, which is in intersection of some networks, has different values in each network. Hence, we have to estimation the true similarity from observed similarities in each network.

We discuss the similarities that calculated from each user's social bookmark. Now we assume there is a true similarity between web pages. In this situation a model is introduced that the similarities are regarded as observation of the true similarity. This model can be represented below.

$$\mathrm{sim}^k_{i,j} = \mathrm{sim}^*_{i,j} + \varepsilon_k$$

ε_k is an observation error and denotes a normal distribution N(0, σ_k^2). $sim^*_{i,j}$ denotes the true similarity between web pages, w_i and w_j. Using the model, a distribution of $sim^k_{i,j}$ is defined below.

$$\Pr(\mathrm{sim}^k_{i,j} \mid \mathrm{sim}^*_{i,j}) = \mathrm{N}(\mathrm{sim}^*_{i,j}, \sigma_k^2)$$

Next we discuss a distribution of $sim^*_{i,j}$. When the similarity between w_i and w_j is estimated previously, it is favorable to adjust the estimation according to the information in determining the similarity from observations. For example, we can determine it previously comparing their contents using a cosine similarity. To achieve this aim we introduce a prior distribution of $sim^*_{i,j}$. The distribution is defined below.

$$\Pr(sim^*_{i,j}) = \mathrm{N}(\mu, \sigma_0^2)$$

μ is an estimation value based on previous knowledge and σ_0^2 denotes reliance for the estimation.

Finally we describe an estimation method after some observations are obtained. Combining observations with previous knowledge on the similarity, Bayes theorem is used. Now assume that n users register w_i and w_j web pages and n similarities $\mathrm{sim}^k_{i,j}$ is calculated. A posterior distribution $\Pr(\mathrm{sim}^*_{i,j}|\mathrm{sim}^1_{i,j}, \ldots, \mathrm{sim}^n_{i,j})$ is defined below.

$$\begin{aligned}\Pr(\mathrm{sim}^*_{i,j} \mid \mathrm{sim}^1_{i,j},\cdots,\mathrm{sim}^n_{i,j}) &= \frac{\prod_k \Pr(\mathrm{sim}^k_{i,j} \mid \mathrm{sim}^*_{i,j}) \Pr(\mathrm{sim}^*_{i,j})}{\sum_{\mathrm{sim}^*_{i,j}} \prod_k \Pr(\mathrm{sim}^k_{i,j} \mid \mathrm{sim}^*_{i,j}) \Pr(\mathrm{sim}^*_{i,j})} \\ &\propto \prod_k \Pr(\mathrm{sim}^k_{i,j} \mid \mathrm{sim}^*_{i,j}) \Pr(\mathrm{sim}^*_{i,j}) \\ &= N(\mu' \mid \sigma'^2)\end{aligned}$$

μ' and σ'^2 are calculated below.

$$\mu' = \frac{\frac{\mu}{\sigma_o^2} + \sum_k \frac{\mathrm{sim}^k_{i,j}}{\sigma_k^2}}{\frac{1}{\sigma_0^2} + \sum_k \frac{1}{\sigma_k^2}}$$

$$\frac{1}{\sigma'^2} = \frac{1}{\sigma_0^2} + \sum_k \frac{1}{\sigma_k^2}$$

In Bayes theorem we determine a similarity using the posterior distribution. Estimation of the similarity between w_i and w_j, $\mathrm{sim}_{i,j}$, is a mode of the posterior distribution $\Pr(\mathrm{sim}^*_{i,j}|\mathrm{sim}^1_{i,j}, \ldots, \mathrm{sim}^n_{i,j})$ in this study. Since the posterior distribution is the normal distribution, the mode of the normal distribution is a mean.

The variation σ_k^2 is used to denotes reliance of the kth user. The bigger σ_k^2 is, the less reliable the kth user is. If the kth user is a spammer, σ_k^2 should be bigger and his evaluation is hard to reflect the final similarity estimation.

4. EXPERIMENTS

In this section we describe evaluation experiments using real social bookmarking data, Buzzurl social bookmarking data that are provided by EC Nave Company. We calculate the similarity between all web pages from all social bookmarking data using the proposed approach and discuss results from the viewpoint of correctness of link creation. The proposed method is compared with comparative approaches that are a cooccurrence approach and an all-tag approach. In Table 1 we show content of Buzzurl social bookmarking data. This data are gathered from October/2005 to October/2008. The social bookmarking data includes 3 elements; “User”, “URL”, “Tag”. Figure 2 shows how many tags are added in web pages. This graph shows about quarter of all web page have no tags. Hence, our proposed approach

cannot use such web pages but a comparative approach that does not use tags uses all web pages. In approach using tags we use about 860,000 web pages to construct a network.

Table 1. Content of Buzzurl social bookmarking data

Data	1,626,869
Unique URL	864,574
Unique user	25,597
Unique tag	352,016

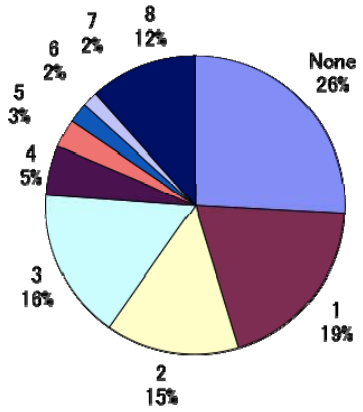

Figure 2. The number of web pages with respect to the number of added tags.

The cooccurrence approach uses cooccurrence frequencies to define the similarity between web pages. Concretely we use the number of users that registered a pair of web pages as the similarity. The relationship occurs in all web pages that a user registered in the same social bookmark though they do not have common tags. Hence, this approach tends to create more pseudo relationships than the proposed approach.

The all-tag approach defines a similarity between web pages using Jaccard Coefficient of tags. This approach is different from the proposed approach because a set of tags for a web page is different. In the all-tag approach the set of tags is constructed by tags all users added for a web page. Since the set of tags includes tags that represent various aspects of a web, it is easy to share tags with web pages dealing with different topics. Hence, this approach tends to create more pseudo relationship than the proposed method, too.

4.1 Comparative Approaches

In this section we explain two comparative approaches to compare them with our proposed approach. They are called a cooccurrence approach and an all-tag approach.

At first we explain the cooccurrence approach. It considers that all web pages registered by the same user have relationship each other because they are selected according to users' interest. A similarity between web pages is the number of users that register a pair of them. The more users register the web pages simultaneously, the more similar they are. Since this approach does not use tags to calculate the similarity, the similarities among all web pages in a

social bookmarking service are calculated easily. An advantage of this approach is to define the similarities among all web pages regardless of their tags. On the other hand, a disadvantage is to create many pseudo relationships. Especially when a user has multiple interests, this approach makes many pseudo relationships between web pages including different topics.

Next we explain the all-tag approach. It calculates the similarity using all tags added in a web page, a set of all tags, T_i is created with a next formulation.

$$T_i = \cup_k T_{ki}$$

The set of tags is a union of tags that each user attached in the web page. The similarity is defined with Jaccard Coefficient. In this approach the number of tags in a web page is more tags than in our proposed approach and more links tend to be generated than the proposed approach. An advantage is that tags added in a web page are generated from various viewpoint and they represent content of it correctly. On the other hand, a disadvantage is that since various tags are added in a web page, it creates relationships easily. Hence, it generates many pseudo relationships.

4.2 Results and Discussion

We make a network from Buzzurl social bookmarking data using the proposed approach and two comparative approaches.

Table 2. Parameters in the proposed approach.

Parameter	Value
μ	0.5
σ_0^2	10
σ_k^2	1

In the proposed approach we determine some parameters previously. Table 2 denotes values of all parameters in the proposed approach. In the experiments all users have the same variance and their reliance is not different. This situation shows we have no knowledge on social bookmarking service users. As a variance of a prior distribution, σ_0^2, is a big value, the prior distribution is similar to a uniform distribution in $0 \le \mathrm{sim}^*_{i,j} \le 1$. This setting denotes that we have no knowledge on $\mathrm{sim}^*_{i,j}$ previously. Using these setting, experiments are carried out under the situation without heuristic knowledge. Hence, we determine the similarities between all web pages from social bookmarking data itself.

Table 3. The number of links created with each approach

Method	Degree
Cooccurrence approach	2,300,343,319
All-tag approach	2,046,287,002
Bayes approach	284,561,877

Table 3 denotes the number of links created with three approaches. Two comparative approaches create more than two billion links and the proposed approach does 200 million links. As we explained that the comparative approaches tended to make too many links, they make about 10 times as many link as the proposed approaches.

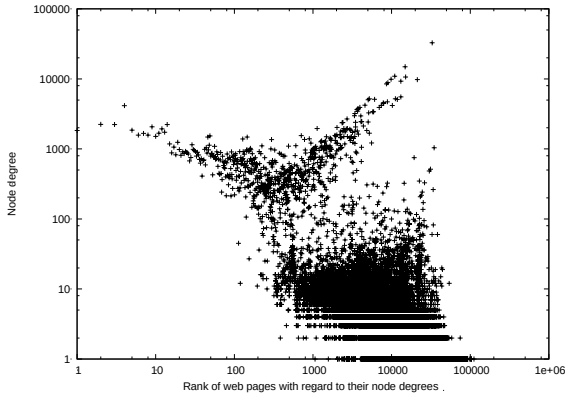

Figure 3. Distribution of node degree in the cooccurrence approach.

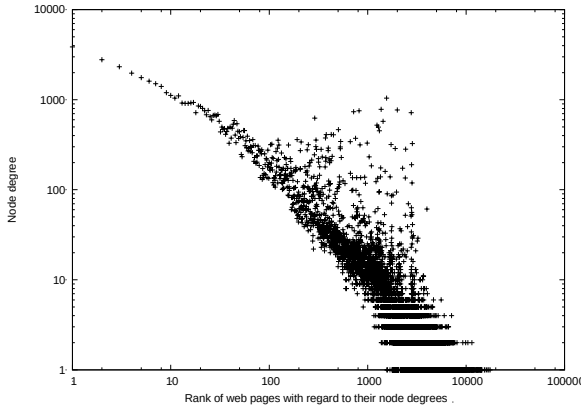

Figure 4. Distribution of node degree in the cooccurrence approach with a threshold.

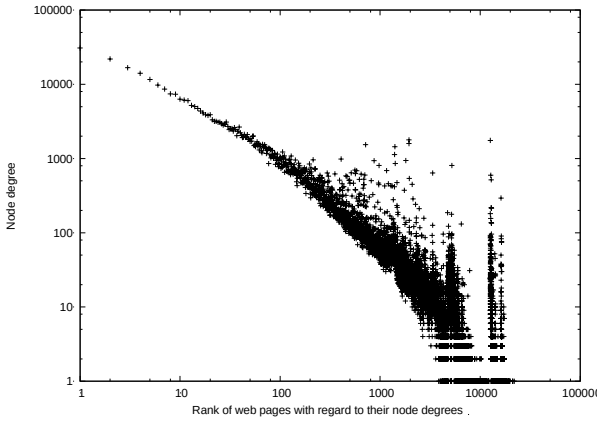

Figure 5. Distribution of node degree in the proposed approach with a threshold.

In Figure 3 node degree in the network constructed with the comparative approach is scattered and it is difficult to find some regularities. Since popular topics are in many web pages, such web pages have many links between other web pages. On the other hand, there are a few web pages dealing with specific topics and they have a few links. We assume that there are many web pages including specific topics in social bookmarking service. From the viewpoint of the assumption the left figure is not good. Especially we should decrease the data that has large node degree since we cannot think there are so many popular topics.

In Figure 4 we remain links that denotes more than 2 users registered both nodes (web pages). This threshold constrains too many links creation and remains links that are evaluated by many users. We hope the threshold deletes pseudo relationships. A distribution in the center figure represents a power-law distribution. However, this approach needs the threshold to delete pseudo links and it is difficult to determine the threshold previously.

In Figure 5 node degree in a network constructed with the proposed approach represents a power-law distribution. Under our assumption the distribution is suitable. Though the approach does not introduce new threshold, the distribution of node degree forms the power-law distribution.

Table 4. Node degree in three approaches for “http://www.google.co.jp”

Method	Degree
Cooccurrence approach	15,565
All-tag approach	25,949
Bayes approach	1,137

These results focus features of the networks constructed with three approaches and does not focus a quality of links. Next we discuss what kind of links the approaches make. Table 4 denotes node degrees for “http://www.google.co.jp”. In Table 4 the all-tag approach makes the most links and the order in Table 3 is different from the order in Table 4. There are “search engine”, “advertisement”, “promotion” and “encyclopedia” as tags for “http://www.google.co.jp”. Especially “advertisement”, “promotion”, and “encyclopedia” represent different aspects of Google. Hence, through these tags “http://www.google.co.jp” connects with many other web pages. This causes too many links between web pages and makes it difficult to find related web pages.

Table 5. Top 6 web pages linked to "http://www.google.co.jp/"

Cooccurrence approach	Bayes approach
http://www.yahoo.co.jp/	http://www.excite.co.jp/
http://www.asahi.com/	http://www.yahoo.co.jp/
http://www.rakten.co.jp/	http://www.hatena.ne.jp/
http://www.livedoor.com/	http://find.2ch.net/
http://jp.msn.com/	http://ask.jp/
http://www.vector.co.jp/	http://jp.msn.com/
All-tag Approach	
http://japan.cnet.com/news/biz/story/0,2000056020,20123367,00.htm	
http://ask.jp/	
http://jp.msn.com/	
http://www.cuil.com/	
http://qooqle.jp/	
http://markezine.jp/a/article/aid/1065.aspx	

Table 5 shows the 6 most relative web pages with “http://www.googl.co.jp”. The cooccurrence approach includes non-relative web pages, “http://www.asahi.com/”, “http://www.vector.co.jp/”. “http://www.asahi.com/” is a site of a Japanese newspaper and “http://www.vector.co.jp” is a site of a freesoft indexing service. Since this approach focuses only cooccurrence frequency, the list includes many famous Japanese web pages and almost all web pages are not relative to “http://www.google.co.jp” directly.

The all-tag approach includes “http://japan.cnet.com/news/biz/story/0,2000056020,20123367,00.htm” and “http://markezine.jp/a/article/aid/1065.aspx’’. The sites are a news site and a blog site and is not relative to “http://www.google.co.jp" directly.

Table 6. Top 6 web pages linked to "http://www.youtube.com/watch?v=P6uFXSE3ARM9"

Cooccurrence approach
http://headlines.yahoo.co.jp/hl?a=20070510-00000015-rec_r-ent
http://www.popxpop.com/archives/2007/06/post_278.html
http://vision.ameba.jp/watch.do?movie=255203
http://www.youtube.com/watch?v=pUcn45Q2dec
http://www.youtube.com/watch?v=cwIsG6O_Iik
http://www.youtube.com/watch?v=Q8ZD4_FL0ms
All-tag approach
http://www.youtube.com/watch?v=mFxXgYiZ6Pw
http://vision.ameba.jp/watch.do?movie=22182
http://www.youtube.com/watch?v=aA3dHi_o7Yw
http://www.youtube.com/watch?v=N_-_kT0apw0
http://www.youtube.com/watch?v=hqjAn0z9yJA
http://www.sorainu.com/archives/50577750.html

Bayes approach
http://vision.ameba.jp/watch.do?movie=580333
http://vision.ameba.jp/watch.do?movie=560990
http://www.youtube.com/watch?v=ULlsVcW5bRI
http://www.youtube.com/watch?v=cpVRaPhVSj4
http://www.youtube.com/watch?v=hdtpfDy0DwA
http://vision.ameba.jp/watch.do?movie=584364

However, the list is better than in the cooccurrence approach. Using tags to define the similarity, web pages that are relative to search services have high similarities. However, since some tags are attached from the other viewpoint, the tags create some pseudo links.

In the proposed approach web pages that are evaluated by only one user are excluded because such similarities are not reliable. The proposed approach can select only relative web sites in the list. All sites are search services and are the famous ones in Japan. Hence, the approach makes an appropriate network.

Table 6 shows the 6 most relative web pages with "http://www.youtube.com/watch?v=P6uFXSE3ARM9". The web page is a movie which is relative to "cat" in YouTube and it is difficult to find similar web pages using content based similarity because the content based approaches have to compare movies. In this work the similarity is calculated regardless of their contents because the approach uses only social bookmarking data. Since social bookmarking data are classified according to users' interests, we can capture the contents of web pages without checking the web pages themselves. It is favorable to connect with web pages that are movies and deal with "cat". In this case the cooccurrence approach is the worst performance because it includes some non-relative web pages. For example, "http://headlines.yahoo.co.jp/hl?a=20070510-00000015-rec_r-ent" is a news article but not movie. Other approaches using tags to calculate similarities are better because listed web pages are movies and relative to "cat". The result denotes the approaches using tags capture true relationships between web pages that deal with the same topic.

5. CONCLUSION

We proposed similarity estimation between web pages in social bookmarking service regarding users' tag vocabulary. The proposed approach consists of two steps: 1) Estimating the similarity between web pages in each users and 2) Integrating the similarities and estimating the true similarity.

We carry out some experiments using Buzzurl social bookmarking data and confirm that the proposed approach is superior to the comparative approaches, which are the cooccurrence approach and the all-tag approach. From the viewpoint of node degree a network that the proposed approach constructs represents a power-law distribution. Though there are a few topics that many users are interested in, there are many web pages dealing with such topics. On the other hand, though there are many topics that a few users are interested in, there are a few web pages dealing with them. Data including previous natures tend to form a power-law distribution easily. We discussed web pages linked with some specific web pages, "http://www.google.co.jp" and "http://www.youtube.com/watch?v=P6uFXSE3ARM9".

In a future work we have to discuss influence of users' variances. Some users are regarded as spammers because they usually register web pages on illegal information and porno sites. As their variances are big, we discuss how their scores reflect final similarities. Next we have to analyze the network constructed and find new knowledge from the network. A network constructed with our proposed method remains some pseudo links though their similarities are low. Since it is difficult to remove them using only relationship between two web pages, we need to remove them based on the network structure. Now we try to use manifold learning approaches as a network approximation method. In the manifold learning approaches keeping a neighborhood of data in original space, they approximate a data distribution in lower-dimension space. They are similar to our goal that we want to remain essential strong relationship (links between related web pages).

ACKNOWLEDGEMENT

We thank EC Navi Company for giving us social bookmarking data.

REFERENCES

Shi, J. and Malik J., 2000. Normalized cuts and image segmentation. *In IEEE Transactions on Pattern Analysis and Machine Learning*, Vol.22, No.8, pp.888-905.

Tang, L. and Liu, H., 2010. *Community Detection and Mining in Social Media*. Morgan & Claypool Publishers, UK.

Mathes A., 2004. Folksonomies – Cooperative Classification and Communication Through Shared Metadata. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

Choudhury, M. D. et al., 2010. Inferring Relevant Social Networks from Interpersonal Communication. *Proceedings of the 19th International Conference on World Wide Web*. Raleigh, NC, USA, pp.301-310.

Horowitzand, D. and Kamvar, S. D., 2010. The Anatomy of a Large-Scale Social Search Engine. *Proceedings of the 19th International Conference on World Wide Web*. Raleigh, NC, USA, pp.431-440.

Fond, T. L. and Neville, J., 2010. Randomization Tests for Distinguishing Social Influence and Homophily Effects. *Proceedings of the 19th International Conference on World Wide Web*. Raleigh, NC, USA, pp.601-610.

Wasserman, S. and Faust K., 1994. *Social Network Analysis*. Cambridge University Press, UK.

Kwak, H. et al., 2010. What is Twitter, a Social Network or a New Media?. *Proceedings of the 19th International Conference on World Wide Web*. Raleigh, NC, USA, pp.591-600.

Xiang, R. et al., 2010. Modeling Relationship Strength in Online Social Networks. *Proceedings of the 19th International Conference on World Wide Web*. Raleigh, NC, USA, pp.981-990.

Rucker, J. and Polanco, M. J., 1997. Siteseer: personalized navigation for the Web. *Communications of the ACM*, Vol.40, No.3, pp.73-76.

Yeung, C. A.. et al., 2007. Understanding the Semantic of Ambiguous Tags in Folksonomies. *Proceedings of the International Workshop on Emergent Semantics and Ontology Evolution at ISWC/ASWC 2007*. Busan, Korea, pp.108-121.