Study on User Behavior and Social Tagging *

Yaping Wan [1], Xiaohua Yang [2], Zhiming Liu, Jiayu Ma, Xiaoyun Li , Chunping Ouyang, Ying Yu, Hui Jiang

School of Computer Science and Technology, University of South China, Hunan Province, China

ypwan@yahoo.cn [1], xiaohua1963@yahoo.com.cn [2]

Abstract - We studied the electronic communication of knowledge users collaborating on a community and found that their work and interactions were mediated by the use of tag. Drawing on these, we found social tagging is the process by which many users add metadata in the form of keywords, to annotate and categorize information items (books, pictures, films etc.). Automatic discourse classification according to tag in social information sharing, transfer and knowledge communication provided a higher level of service quality. By investigating user behavior in Douban community, the relationship between social tagging and user behavior was studied. The results showed that, given a set of objects, and a set of tags applied to those objects by users, we can predict whether a given tag could/should be applied to a particular object by user behavior.

Index Terms - Tags; Social Networks; Classification; Discourse

1. Introduction

The advances experienced in the last decades in areas as information, communication, and media technologies have made available a large amount of all kinds of data. This is particularly true for Web community, whose data have grown exponentially since the advent of Web 2.0 and the development of Internet technology. The registered number of social network [1, 2] is growing year by year, and its content is constantly updated, which provides an unprecedented real experimental platform for the study of large-scale social network. The traditional classification techniques use text content build indexing model, but the social network contains a variety of structured and unstructured data (such as the various video, audio, graphics, image information), which makes it difficult to use the traditional classification methods to recommended or obtain information. This situation demands for tools able to ease searching, retrieving, and handling such a huge amount of data. Among those tools, using social tagging classifiers have a particularly important role, since they could be able to automatically index and retrieve all kinds of data in a human-independent way. This is very useful because a large portion of the words used to describe Web content is inconsistent or incomplete.

Tagging has become a popular way to organize content on the Web in order to simplify access to documents. Many systems, such as Delicious, Flickr, Digg and Connotea use social tagging to categorize their information items and help users to share them. However, despite increased interest in tagging, tags are still poorly understood to user and information system. In particular, little is known about the recommendation of tags, because different persons have different choices and the number of tags is too much. Various methods have been discussed and proposed in order to provide tag recommendation on social knowledge sharing and communication. Unfortunately, most of them are seen to be complicated to meet the aspects of information tagging.

Some researchers has developed tag recommendation algorithms [3, 4], which try to exploit tags given by users on specific items. Golder and Huberman[5] analyzed the structure of collaborative tagging systems as well as their dynamic aspects. In contrast to the above ternary relation, many recommender systems use Collaborative Filtering (CF) to recommend items based on preferences of similar users, by exploiting a two-way relation of users and items [6]. YouTube allows users to tag only the videos that they have uploaded. Therefore, collaborative tagging is not directly possible.

Other tag recommendation algorithms are based on conceptual stuctures similar to the hyperlink structures used in Search Engines. For example, Collaborative Tag Suggestions algorithm [7], also known as Penalty-Reward algorithm (PR) uses an authority score for each user. The authority score measures how well each user has tagged in the past. Schmitz et al. is primarily concerned with theoretical properties of mining association rules in tripartite graphs. Schwarzkopf et al. [8] extend Schmitz's association rules work to build full ontologies. However, neither Schmitz et al. nor Schwarzkopf et al. appear to evaluate the quality of the rules themselves aside from generating ontologies. Lastly, there is also much previous work in IR studying query expansion and relevance feedback trying to address similar questions of cross-language and cross-vocabulary queries (see for example a general reference such as Manning et al. [9]). However, we believe that association rules may be the most natural approach to these problems in tagging systems due to user interface issues. Chakrabarti et al. [10] suggest a different way to use local link information for classification that might prove more effective than our domain features, however, we do not evaluate this possibility here.

As can be seen from above those algorithms are too complex to widely applied. The main contributions of the method described in this paper are twofold. First, it excavated the relationship between the different object tags according to the common behavior of different user. Second, it built a multi-dimensional tag model which use vector denoting tag.

* This work is partially supported by Hunan Natural Science Foundation (10JJ6097, 11JJ6047), Hunan province science and technology Program(No2011FJ3087), Hunan Scientific Research Key Foundation funded project (11A105), Doctor Start Fund of University of South China(2011XQD39) and Social Science Fund of University of South China(2012XYB02).

2. Research Method

A. Stability and diversity of the tag

Tag originally has been considered to be in the domain of literature and has widely been discussed in literary theory. There are some other scholars who identify tag with speech event. The most influential one may be Swales [11]. He proposes tag comprises a class of communicative events, the members of which share some set of communicative purposes. The purposes are recognized by the expert members of the parent discourse community, and constitute the tag rationale. This rationale shapes the schematic structure of the discourse and influences and constrains choice of content and style. The tag names inherited and produced by communities and imported by others constitute valuable ethnographic communication, but typically need further validation. For example, we find there is a total of forty kinds of tag in movie community "douban", such as suspense, action, love, fiction et al. Any movie has a variety of tags. For instance, the film "Let the bullets fly" includes four tags, drama, comedy, action, western.

B. Formal definition of the tag

Many social system use tag to classify objects. While tag exists the essential features, which are known as the basic tag. Social discourse tag is a cross-mmixing of the basic tags. Discourse tag in social network comes from a variety of basic tags. Therefore, we can obtain the following definition,

Definition 1.Given the social network discourse space U, G={g1,g2,…,gn} is the collection of the basic tags of the U, the tag Tg of the discourse T is an n-dimensional vector group (tg1,…,tgn), in which tgi ∈ [0,1],

tgi=0，if T has no the characteristic of the basic tag gi,

tgi>0，if T has no the characteristic of the basic tag gi.

Definition 2.If the social network Discourse T's n n-dimensional vectors are linearly independent, n Independent vectors in this group become a basis sector to measure the n-dimensional discourse space.

Property 1. Among any two discourse T1 and T2 in discourse space of social networking, define using tag vector distance measure tag similarrity between T1 and T2. The smaller the distance is, the higher the tag similarity of T1 and T2 is. Assume that the tag vectors of two discourses T1, T2 were,

Tg1=(tg1,1,tg1,2,tg1,3,…,tg1,n,)

Tg2=(tg2,1,tg2,2,tg2,3,…,tg2,n,)

Where, n is the number of basic tags.

Discourse T1 and T2 tag similarrity is

$$\mathrm{sim}(Tg1,Tg2)=\cos(Tg1,Tg2)=\frac{\sum_{k=1}^{n}(tg_{1,k}\times tg_{2,k})}{\sqrt{\sum_{k=1}^{n}tg_{1,k}^{2}}\times\sqrt{\sum_{k=1}^{n}tg_{2,k}^{2}}} \quad (1)$$

In formula (1), we use the cosine vector to measure the discourse tag similarity. To social discourse which tag is known, using the traditional vector space model calculate the tag similarity. For those unknown tags of discourses by analyzing the potential common characteristics between tags and user common behavior, the use of user behavior to measure the tag similarity, and therefore leads to the following Property 2,

Property 2. The number of co-occurrence among community users behavior can quantify the strength of the tag relationship between discourse.

3. Results and Discussion

This paper selects Douban's data as the basis of experiment, in which the different labels, the number of users and user behavior data are selected to statistics and analysis. And we use MySql database to stored the download data in the location.

A. Douban film community experiments

This experiment in which we collected a total of 1024 films using the 36 tags, on average, over 98.2% films had multiple tags. These 36 tags are as shown in Table 1. The digital in table 1 following each tag expresses the total number of the movies of this tag in the Douban film community. Purpose of the experiment: a survey on the number of people that any two films are common seen and the common tags that two films had verified the potential commonalities relationship of the tag and user behavior.

TABLE I *Douban* film tags

Love (3417826)	Comedy (2581116)	Animation (2476843)	Classic (1670352)
Fiction (1463871)	Action (1403183)	Youth (1356516)	Plot (1165285)
Suspense (977295)	Thriller (751505)	Documentaries (592406)	Inspirational (579749)
Crime (545789)	Terror (524562)	Funny (504163)	Magic (482861)
War (480216)	Literature &art (471861)	Short (434910)	Animated Short Film (432140)
Erotic (383577)	Black humor (339219)	Childhood (321907)	Biography (280779)
Comrade (271822)	Violence (261329)	Music (256832)	Gang (226566)
Female (184902)	Romantic (183303)	Moving (177904)	Family (143504)
Fairy tale (140546)	Epic (126130)	cult (99423)	Shock (51330)

Firstly, we download original movie data, users, user behavior on the movie (mainly four kind of user behavior data: see, read, want to see, the critic). Secondly, we filter out the users who had seen a certain amount of movies and had kept a certain preference. Finally we filter films.

Step 1: we filter out all films which constituted the film set F1 and were seen by the set of users U1;

Step 2: The films which are selected from F1 which had been seen over a certain value by U1 constituted the film set F2;

Step 3: That the number of users who had seen (or the user behavior of seeing, film critic) any two movies which were selected from the collection F2 and the number of common tags are gotten. The entire experimental results are shown in Figure 1-3.

At the same time, the statistical of user behavior data which user are seeing films also had been collected, but the amount of online data is too little to discover the laws, so this paper temporarily does not analyze and discuss them.

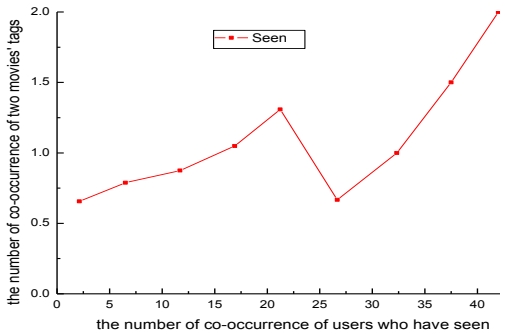

Fig. 1 Relationship of user seen behavior and tags

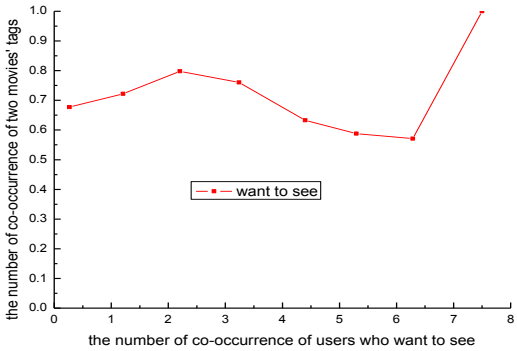

Fig. 2 Relationship of user hope to see behavior and tags

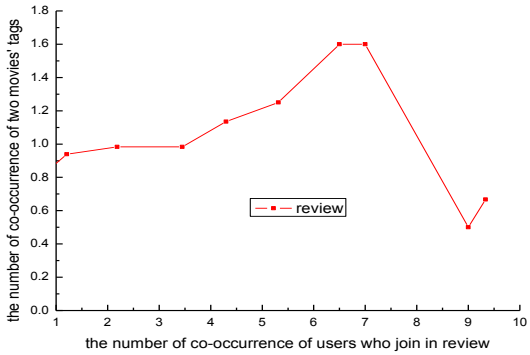

Fig. 3 Relationship of user film critic behavior and tags

As can be seen from figure 1 to 3 the behavior which users had seen these films is of most consistent with our basic design ideas. This is due to the kind of data is the most in all four behaviors we collected. Statistical thinking reveals that the greater the capacity of the sample is, the smaller the sampling error. On the other hand, users want to see and film critic behavioral data also are basically in line with our hypothesis. This shows that the negotiations between the social network and Linguistics discourse have common characteristics and user behavior in social networks and tag has great similarities. In other words, the more similar discourse tag is, the more similar behavior of users of social network is.

B. Douban books community experiments

This experiment in which we collected a total of 2783 books using the 36 popular tags which are selected from hundreds of tags. The same as above, these 36 tags are as shown in Table 2.

TABLE Ⅱ *Douban books tag*

Novel (5614836)	Essay (1251173)	Prose (3776563)	Fairy tale (2383568)
Poem (3263941)	Masterpiece (5203323)	Cartoon (2354168)	Picture book (1758621)
Suspense (1245872)	Reasoning (1452145)	Romance (985647)	Science fiction (758649)
Martial arts (1287321)	Fantasy (653281)	History (751893)	Philosophy (561892)
Biography (358968)	Design (238957)	Memoirs (365879)	Music (587624)
Travel (128597)	Inspirational (389576)	Workplace (128765)	Education (752416)
Food (1869421)	Devotional (468597)	Health (2486273)	Home (385694)
Management (853695)	Financial (178546)	Business (235847)	Marketing (156234)
Financial management (215546)	Stock (456231)	Popular Science (52648)	Internet (1025844)

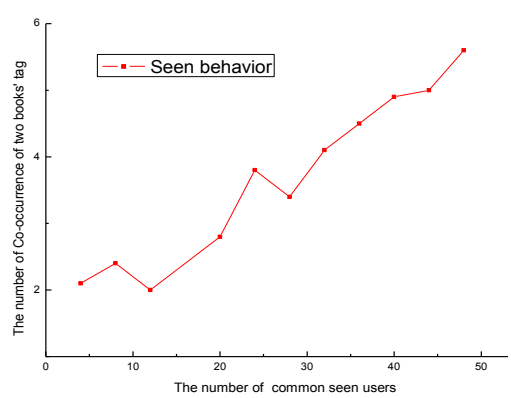

Fig. 4 Relationship of user seen behavior and books tag

In this experiment, we selected 36 popular tags from hundreds of tags in Douban books community. we download original Douban book data, users, user behavior data on the books (mainly three kind of user behavior data about book community: seen, hope to enjoy,comment). We download million of raw data about book and the main purpose is focus on mining the relationship between user behavior and tags. The results are shown in Figure 4-6.

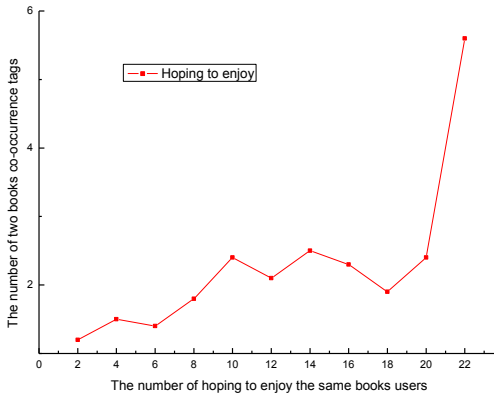

Fig. 5 Relationship of hoping to enjoy behavior and books tag

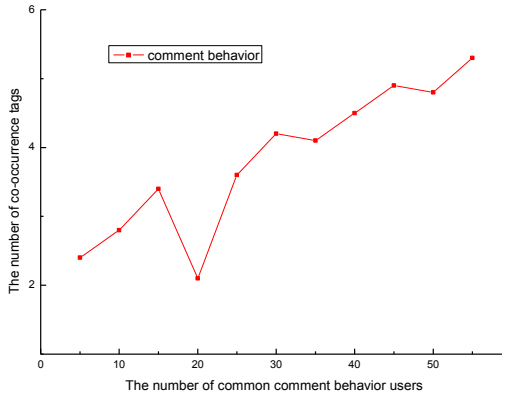

Fig. 6 Relationship of comment behavior and books tag

Figure 4-6 reveals the potential relationship between the user behavior and books label.Those data indicated that, given a user and an object, the purpose is to help whether and how much the user is likely to tag this item with a specific tag which most likely comes from the same team member tags. This suggests that they may serve as a way to link disparate vocabularies among users. This makes tags appropriate for corpora like the web and user contributed video and photo collections where the distribution or type of content may change rapidly.

4. Conclusion

With the enhancement of network penetration and the popularity of network applications, the scale of social networking users will be further expanding, and more and more users will extend real life relationships to network. This study shows that users in the social network information behavior have a high degree of user viscosity. Users always focus on the discourse tag of their own interest, and the discourse set most users are interested is greater similarity in the tag. Also due to the use of different purposes, the user behaviors exhibited. Overall, the community user behavior on social networking sites are more dispersed, but the discourse tag embodied communication in a fixed time when a particular discourse tag is concerned is still the center of user behavior.

References

[1] KONG Weize, LIU Yiqun, ZHANG Min, MA Shaoping. Answer Quality Analysis on Community Question Answering. Journal of Chinese Information Processing. 2011; 25(1):3-8.

[2] WANG Yuxiang, QIAO Xiuquan, LI Xiaofeng, MENG Luoming. Research on Context-Awareness Mobile SNS Service Selection Mechanism. Chinese Journal of Computers. 2010; 33(11):2126-2135.

[3] H. Wang and N. Ahuja. A tensor approximation approach to dimensionality reduction. International Journal of Computer Vision. 2007; pages 217-229.

[4] A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme.Information retrieval in folksonomies: Search and ranking. In The Semantic Web: Research and Applications, pages 411 – 426, 2006.

[5] Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. Collaborative Web Tagging Workshop, 2006.

[6] S. Golder and B. Huberman. The structure of collaborative tagging systems. In Technical Report, 2005.

[7] J. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In Proc. Conf. on Uncertainty in Artificial Intelligence, pages 43 – 52, 1998.

[8] E. Schwarzkopf, D. Heckmann, D. Dengler, and A. Kroner. Mining the Structure of Tag Spaces for User Modeling. Workshop on Data Mining for User Modeling (ICUM'07).

[9] Chakrabarti, B. Dom, and P. Indyk. Enhanced Hypertext Categorization Using Hyperlinks. SIGMOD'98.

[10] C. D. Manning, P. Raghavan, and H. Schutze. Introduction to Information Retrieval. Cambridge University Press, 2008.

[11] M. Swales, J. M.. Genre Analysis:English in Academic and Research Settings. Cambridge:Cambridge University Press,1990.