Research of Social Opinion Monitoring Mechanism of Baoding City Based on micro-blogging

Yatao Zhu
College of Information Science & Technology,
Agricultural University of Hebei
Baoding, China
zhuyatao@ict.ac.cn

Hua Jin
College of Information Science & Technology,
Agricultural University of Hebei
Baoding, China
jinhua923@163.com

***Abstract*—Recent years have witnessed a tremendous growth of research and application in the field of social opinion monitoring mechanism. Moreover, social tagging has grown to be a particular tool for users to organize and share digital content on many social webpages. Among the knowledge discovery techniques that are applied in social tag recommendation systems, those based on collaborative filtering are achieving widespread success. The similarity measurement is critical to determine the appropriate results recommendation in the collaborative-filtering schema. In the paper, a nugget is introduced as an atomic conceptual entity, to measure the similarity of web con-tent and recommend tags. With nuggets, we can use the conceptual neighbors, rather than the literal ones for collaborative filtering, which consider the common case that the expression varies for a specific concept. The experiments conducted on the dataset from micro-blogging about Baoding city, have shown that the approach is effective and consistently achieves better precision and recall than both baselines.**

Keywords—social tagging; monitoring; recommendation; micro-blogging; nugget

I. INTRODUCTION

With the web technology evolution, especially the arise of Web2.0 applications such as Del.icio.us, Flickr and Citeulike, social tagging has become a popular service on the web due to its effectiveness in organizing and accessing web resources. Although social tags are very useful, lots of webpages have few or no annotations (Heymann et al., 2008). Thus automatically generating social tags for a new webpage is gaining more and more attention (Lu et al., 2009). Among of knowledge discovery techniques, applied in social tag recommendation systems, the collaborative filtering (CF) based ones are achieving widespread success. The underlying assumption of CF approach is that those who agreed in the past tend to agree again in the future.

Nugget which is first proposed in Question & Answering evaluation (Voorhees, 2003), represents a conceptual entity. It should be atomic, in the sense that an assessor should be able to make a binary decision as to whether the nugget appears in webpage content. In the paper, we firstly build both word nuggets and tag nuggets for webpage content and social tags, based on WordNet's synsets. And then we propose a CF approach by using conceptual similarity to improve social tag recommendation. With experiments conducted on the micro-blogging dataset, our method achieves higher precision and recall than the baselines.

The rest of the paper is organized as follows. The next section describes the related work followed by the research. We first describe the development of social tagging and then discuss collaborative filtering approaches implemented in tag recommendation. Finally we provide a brief background of nuggets associated with conceptually neighboring. In section 3, we present our approach of social tag recommendation by using nugget-based neighborhood, where details of the approach are given. Section 4 describes our experimental results. It provides details of our datasets, evaluation process, methodology and results of different experiments. Section 5 provides some concluding remarks and directions for future research.

II. RELATED WORK

In the earlier studies on social tags, Quintarelli (2005) gave a general introduction of social annotation and suggested that it should be taken as an information organizing tool. In (Golder and Huberman, 2006), Golder and Huberman provided empirical study of the tagging behavior and the usage of tags in Del.icio.us.

The developers of one of the first recommender systems, Tapestry (Goldberg et al., 1992), coined the phrase "collaborative filtering", which has been widely adopte d regardless of the facts that recommenders may not explicitly collaborate with recipients and prediction may suggest particularly interesting items. The fundamental assumption of CF is that if users X and Y rate n items similarly, or have similar behaviors (e.g., buying, watching, listening), and hence will rate or act on other items similarly (Goldberg et al., 2001). Nakamoto et al. (2007) considered the context clues through tags and social connectivity among users in the CF approach. Lu et al. (2009) proposed a CF approach to generate tags of a webpage from those tags of its nearest neighbors, according to the literal similarity between their web content.

The nugget-based paradigm has been previously detailed in a number of papers (Voorhees, 2003; Hildebrandt et al., 2004; Lin and Demner-Fushman, 2005). "Information Nuggets" were firstly used for judging the quality of the question answering

(QA) systems' responses in TREC 2003 (Voorhees, 2003). Afterwards, the focus of evaluation shifted from documents and facts to more elaborate nuggets, and nugget-based evaluation methodology was created (Hildebrandt et al., 2004; Lin and Demner-Fushman, 2005). The idea of "nugget pyramids" was introduced as a refinement to the nugget-based methodology used to evaluate answers to complex questions in the TREC QA tracks (Lin and Demner-Fushman, 2006), and Dang and Lin (2007) evaluated its performance. Lad and Yang (2010) provided a nugget-based approach for learning to rank relevant and novel documents through user feedback, which used observable query and document features (words and named entities) as surrogates for nuggets, whose weights are learned based on user feedback in an iterative search session. In this paper, we borrow the conception "nugget" from above and build nuggets for conceptually neighboring to improve the performance of traditional CF method.

III. NUGGET-BASED COLLABORATIVE FILTERING FOR SOCIAL TAG RECOMMENDATION

A. Preliminaries

According to the bag-of-word assumption, which considers a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order, we use a vector w_i in a word space $\boldsymbol{W}$ to present the textual content of webpage i. Each element $w_{i,j}$ of w_i indicates the frequency of $word_j$ in webpage i. Likewise, we use a vector t_i in a tag space $\boldsymbol{T}$ to signify the social tags of webpage i. Each element $t_{i,j}$ of t_i means the frequency of the tag_j that is used to annotate the webpage by web users. So webpage i with its social tags is represented as a 2-tuple (w_i, t_i), and the corpus consists of such tuples of the webpages. The corpus is divided into training dataset $\boldsymbol{R}$ and testing dataset $\boldsymbol{D}$. As (Li et al., 2008) mentioned, social tags usually differ from the words in webpages literally. Furthermore, we assume that word space $\boldsymbol{W}$ and tag space $\boldsymbol{T}$ are large enough so that there will not arise any new word or tag out of $\boldsymbol{W}$ and $\boldsymbol{T}$ separately. And when training dataset $\boldsymbol{R}$ is big enough, the assumption comes very close to fact. Finally, the social tag recommendation problem can be described as follows.

The social tag recommendation is to predict a ranked tag list, which is applied to a particular webpage of testing dataset $\boldsymbol{D}$, from a set of tags applied to those webpages of training dataset $\boldsymbol{R}$ by users. Given the frequency of social tags follows a power law, there is usually a long "tail" in the ranked tag list. Thus we mainly focus the top-10 frequently annotated tags, and the top-10 tags in testing dataset $\boldsymbol{D}$ are chosen as the ground truth to evaluate the quality of the tag prediction.

B. Building Nuggets

We scan each word of the word space $\boldsymbol{W}$ and build a corresponding nugget when the word firstly appears in one synset of WordNet. The nuggets cells are named $C_1, C2 \ldots C_n$. Meanwhile, we convert the element $w_{i,j}$ of $\boldsymbol{W}$ to $w_{i,j,k}$, where k is the index of C. Furthermore, not all words of webpage i carry the same information, based on the first element, we compute the weight of each word in the nugget as follows. Let $p(A_{j,k})$ be the normalized the weight of word as equation (1) defines.

$$p(A_{j,k}) = \frac{\sum_{i \in R} w_{i,j,k}}{\sum_{i \in R, A_j \in G_k} w_{i,l,k}} \tag{1}$$

where $A_{j,k}$ denotes the word j of the C_k. In the same way, we can build nuggets for tag space $\boldsymbol{T}$ as $G_1, G2 \ldots G_n$ and let $p(B_{j,k})$ be the normalized the weight of tag as equation (3) defines.

$$p(B_{j,k}) = \frac{\sum_{i \in R} t_{i,j,k}}{\sum_{i \in R, B_l \in G_k} t_{i,l,k}} \tag{2}$$

where $B_{j,k}$ denotes the tag j of the G_k.

C. Collaborative Filtering with Nuggets

The basic idea of collaborative filtering for tag prediction is that the webpages with similar content have high probability to share their tags. Thus similarity computation between items or users is a critical step in neighborhood-based collaborative filtering algorithms. Firstly, based on the section 3.4.1, we not only take into account the frequency of word, but also consider the weight of each word in the nugget. Thus the textual content of webpage i is denoted by a new vector u_i in the nuggets set C, where each element $u_{i,k} \in u_i$ indicates the nugget C_k weight in webpage i and it is defined in equation (3) and equation (4). Since the total number of words in a webpage follows a power law, we take its logarithm to avoid being over-weighted. The const value 1 is an adjustable parameter to avoid the result is zero when the total number of some word is one.

$$u_{i,k} = \sum_{j \in C_k} T_{i,j,k} \tag{3}$$

$$T_{i,j,k} = \begin{cases} \ln(w_{i,j,k} + 1) \times p(A_{j,k}), w_{i,j,k} \neq 0 \\ 0 \qquad\qquad\qquad\quad , w_{i,j,k} = 0 \end{cases} \tag{4}$$

The social tags of webpage i, likewise, are represented by another new vector v_i in the nuggets set G, where each element $v_{i,k} \in v_i$ means the weight of the G_k that is used to annotate the webpage by web users and it is defined in equation (5) and equation (6).

$$v_{i,k} = \sum_{j \in G_k} E_{i,j,k} \tag{5}$$

$$E_{i,j,k} = \begin{cases} \ln(t_{i,j,k} + 1) \times p(B_{j,k}), t_{i,j,k} \neq 0 \\ 0 \qquad\qquad\qquad\quad , t_{i,j,k} = 0 \end{cases} \tag{6}$$

Thus webpage i with its social tags is represented as a new 2-tuple (u_i, v_i) and the dataset $\boldsymbol{R}$ and $\boldsymbol{D}$ consists of such tuples of the webpages. Secondly, we employ the cosine similarity $Sim_{i,j}$ defined in equation (7) to measure the content similarity of webpage i and j.

$$Sim_{i,j} = \frac{u_i \cdot u_j}{\|u_i\|\|u_j\|} \tag{7}$$

where u_i and u_j are the vectors denoting webpage i and j. Thus, we can find the k-nearest neighbors (k-NN) N_i of an unannotated webpage i based on the result of equation (7).

Next, we include the weight of each tag nuggets and the similarity of webpage i and its neighbor as two factors to adjust the tag nuggets significance. And the CF method generates the ordered tag nuggets list according to the following weight $\boldsymbol{w}_{i,k}$ defined in equation (8).

$$w_{i,k} = \sum_{n \in N_i} (v_{n,k} \times s_{i,n}) \tag{8}$$

where $\boldsymbol{s}_{i,n}$ is the normalized similarity in neighbors N_i as defined in equation (9).

$$s_{i,n} = \frac{Sim_{i,n}}{\sum_{n' \in \kappa_i} Sim_{i,n'}} \tag{9}$$

where n' is each webpage in the k-nearest neighbors N_i of webpage i.

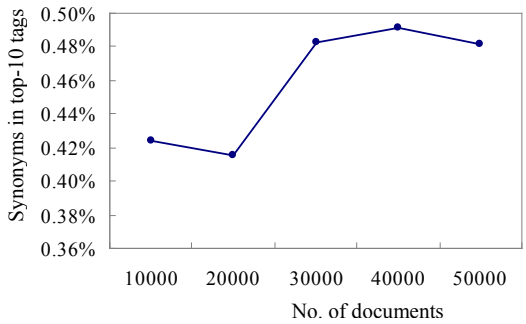

Fig.1. The percent of synonyms in top-10 tags of different amount of documents

As Fig.1 shows, the proportion of synonyms in top-10 tags is below 0.5 percent. That is to say, web users prefer to select tags from different tag nuggets to annotate the webpage. Thus, the top-k tags for the unannotated webpage i are recommended from the top-k tag nuggets based on the ordered tag nuggets list above. Given the weight of each tag in corresponding tag nugget and the frequency of the tag that is used to annotate the k-nearest neighbors N_i of the unannotated webpage i by web users, the each recommended tag j from the corresponding tag nugget G_k is generated according to the following weight $r_{i,j}$ defined in equation (10). For the same considerations, we take logarithm of tag frequency to avoid being over-weighted.

$$r_{i,j} = p(B_{j,k}) \times \ln \sum_{n \in N_i} (t_{n,j,k} + 1) \tag{10}$$

where n is each webpage in t the k-nearest neighbors N_i of webpage i.

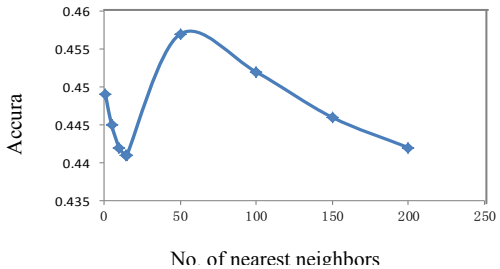

Fig.2. The accuracy with different number of nearest neighbors in CF method.

Additionally, with different number of neighbors, we check the accuracy of top-10 predicted tags. Based on 10-fold cross-validation, the test is performed on $\boldsymbol{R'}$ with 5,000 webpages from training dataset $\boldsymbol{R}$, and the k-nearest neighbors are found from the left 45,000 webpages. As Fig.2 shows, the accuracy reaches maximum with around 50 neighbors. Thus, we choose 50 nearest neighbors in the CF-based tag recommendation method.

IV. EXPERIMENTS

In the experiments, we firstly employed a hybrid crawling strategy that combines tag, URL and user-centered crawling strategies (Heymann et al., 2008) to crawl a relatively unfiltered view of the data from sina website during October and November, 2014. The original data crawled includes 167,958,659 bookmarks made by 825,402 different users on 57,813,581 different URLs, with 5,916,196 different tags. Secondly, in order to reduce the noises to the training procedure of CF and Corr-LDA, we filtered out those webpages annotated by less than 100 users. Next, we randomly selected 50,000 tagged webpages and another 10,000 ones as training dataset $\boldsymbol{R}$ and testing dataset $\boldsymbol{D}$ separately. Finally with suffix stripping and stemming of the words in webpage content and the tags annotated to the relevant webpage, the size of word space $\boldsymbol{W}$ reaches 67,146, while that of tag space $\boldsymbol{T}$ is 12,669.

A. Evaluation Metrics

In the experiment, we compare both CF and Corr-LDA approaches with our method in the following metrics (Song et al., 2008). They are the accuracy of top-k, exact-k, recall and precision of recommendation tags. The accuracy of top-k is the percentage of webpages correctly annotated by at least one of the top-k predicted tags. The accuracy of exact-k is the percentage of webpages correctly annotated by the k-th predicted tag, which gives the indication that whether the tags ranked higher in prediction list are more likely to annotate webpages. Recall is the percentage of correctly predicted tags in the user-annotated ones. And precision is the percentage of the correct tags in the predictions by an algorithm.

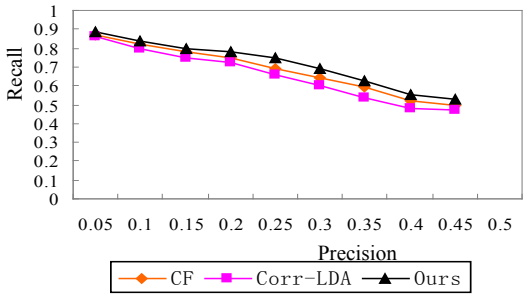

Fig.3. precision-recall curve

B. *Experimental Results*

The last row in Table 2 illustrates the improvement of our approach in Top-10 accuracy and Exact-10 accuracy, compared with other both well-known approaches. In addition, the precision-recall curve is drawn as well in Fig.3 to show that our method could constantly gain higher recall while keeping higher precision compared with CF and Corr-LDA methods. In summary, the experimental results show that our method indeed generates a well ordered tag list, which outperforms the two well-known methods.

TABLE I. Compares the two approaches with our approach in Top-k accuracy and Exact-k accuracy

	CF		**Corr-LDA**		**Our Method**	
	Top-k	*Exact-k*	*Top-k*	*Exact-k*	*Top-k*	*Exact-k*
1	81.8%	82.2%	80.1%	80.1%	82.1%	84.3%
2	89.0%	69.5%	89.6%	67.0%	89.9%	72.2%
3	92.4%	61.5%	93.3%	58.4%	93.6%	64.1%
4	94.1%	54.5%	95.4%	49.9%	95.7%	55.3%
5	95.2%	47.1%	96.6%	43.8%	96.6%	49.7%
6	96.1%	41.2%	97.4%	38.3%	97.8%	42.3%
7	96.7%	36.0%	97.9%	33.1v	98.1%	36.4%
8	97.2%	31.4%	98.3%	28.7%	98.4%	32.6%
9	97.5%	27.7%	98.6%	25.6%	98.7%	28.1%
10	97.8%	24.7%	98.8%	23.2%	98.9%	25.2%
Imp	**1.2%**	**1.4%**	**0.5%**	**4.2%**	–	–

V. CONCLUSION

In the paper, we investigate the problem of social tag prediction, aiming at generating tags automatically for webpages by considering the varieties of the expression for a specific meaning. We build word nuggets and tag nuggets via a well-known lexical tool WordNet, which reduces web content and social tags from the literal space into the conceptual space. Then, we propose a new algorithm for automated social tagging by using nugget-based neighborhood in collaborative filtering framework. Experimental results show that both the precision and recall of our predictions consistently outperforms both of the baselines, content-based collaborative filtering and Corr-LDA. A further study on building nuggets through topic model with the honor of the sparsity is our focus in the future.

ACKNOWLEDGMENT

This work has been partially supported by Plan Project of Research and Development of Science and Technology of Baoding under Grant No.13ZF098 and No.13ZN025 and Youth Foundation of Science and Technology of College of Hebei Province with Grant No.Z2012142.

REFERENCES

[1] Bao. S., Xue. G., Wu. X., Yu. Y., Fei. B., and Su. Z. Optimizing web search using social annotations, In WWW ’07: Proceedings of the 16th international conference on World Wide Web. 501–510.

[2] Blei, D. M. and Jordan, Modeling annotated data. InSIGIR ’03: Proceedings of the 26th annual international ACM SIGIR conference on Re-searchand development in information retrieval. 127–134.

[3] Conde. J. M., Vallet. D., and Castells. P. Inferring user intent in web search by exploiting social annota-tions, In Porc. of international ACM SIGIR confer-ence on Research and development in information retrieval. 827–828.

[4] Dang. H. and Lin. J. Different structures for eval-uating answers to complex questions: Pyramids won’t topple, and neither will human assessors. In Proceedings of the 45th Annual Meeting of the Asso-ciation for Computational Linguistics (ACL 2007).

[5] Dang. H. and Lin. J., and Kelly. D. Overview of the TREC 2006 question answering track. In Proc. of TREC 2006.

[6] Ellen M. Voorhees. Overview of the TREC 2003.Question Answering Track. In Proc. of TREC 2003

[7] Fellbaum, WordNet: An Electronic Lexical Data-base,Christiane, MIT Press.

[8] Goldberg. D., Nichols. D., Oki. B. Using collaborative filteri ng to weave an information tapestry, In Communications of the ACM.

[9] Goldberg. K., Roeder. T., Gupta. D. Eigentaste: A constant time collaborative filtering algorithm, Information Retrieval, 4(2).

[10] Golder. S. A. and Huberman. B. A. 2006. Usage pat-terns of collaborative tagging systems, J. Inf. Sci. 32, 2, 198–208.

[11] Guo. J., Cheng. X., Xu. G., and Shen. H. A struc-tured approach to query recommendation with social annotation data, In Proc. of the ACM conference on Information and knowledge management. 619–628.

[12] Heymann. P., Koutrika. G., and Garcia-Molina. H. Can social bookmarking improve web search? In WSDM’08: Proc. of international conference on Web search and web data mining. 195–206.

[13] Li. X., Guo. L., and Zhao. Y. E. Tag-based social interest discovery. In WWW ’08:Proceeding of the 17th international conference on World Wide Web. 675–684.

[14] Lin. J. and Demner-Fushman. D. Automatically evaluating answers to definition questions. In Proc. of HLT/EMNLP 2005.

[15] Lin. J. and Demner-Fushman. D. Will pyramids built of nuggets topple over? In Proc. of HLT/NAACL 2006.

[16] Lu. Y.-T., Yu. S.-I., Chang. T.-C. content-based method to enhance tag recommen-dation. In In Proc. of IJCAI’09. 2064–2069.

[17] Nakamoto. R., Nakajima. S., Miyazaki. J., and Uemura. S. Tag-based contextual collaborative filtering. IAENG International Journal of Computer Science 34. 2. 214–219.

[18] Vallet. D., Cantador. I., and Jose. J. M. Personal-izing web search with folksonomy-based user and document profiles. In ECIR 5993. 420–431.