An Optimized Tag Recommender Algorithm in Folksonomy

Jie Chen, Baohua Qiang, Yaoguang Wang, Peng Wang, Jun Huang

An Optimized Tag Recommender Algorithm in Folksonomy

Jie Chen[1], Baohua Qiang[1,2], Yaoguang Wang[1], Peng Wang[1],Jun Huang[1]

[1]Guilin University of Electronic Technology, Guilin 541004, China
cj134cj@163.com
[2]Guangxi key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China
qiangbh@yahoo.com.cn

ABSTRACT. In the existing folksonomy system, users can be allowed to add any social tags to the resources, but tags are fuzzy and redundancy in semantic, which make it hard to obtain the required information for users. An optimized tag recommender algorithm is proposed to solve the problem in this paper. First, based on the motivation theory, the recommender system uses the model given to calculate the user retrieval motivation before searching information. Second, we use the results in first step to distinguish the user's type and then cluster the resources tagged according to users who have the similar retrieval motivation with k-means++ algorithm and recommend the most relevant resources to users. The experimental results show that our proposed algorithm with user retrieval motivation can have higher accuracy and stability than traditional retrieval algorithms in folksonomy system.

KEYWORDS: Folksonomy, tag recommender system, collaborative filtering, user retrieval motivation, k-means++

1 Introduction

Nowadays, with the development of web 2.0, a large amount of digital resources appear and rise. Accord to them, a new network information classification system folksonomy appears. The term folksonomy is generally attributed to Thomas Vander Wal [1], which is a portmanteau of folk and taxonomy. In folksonomy, the tags and resources are created by users, and users can tag the information according to their own needs and preferences in order to make others to retrieve and share the resources. As shown in Figure 1, the resource "Baidu.com" is tagged by user "Brown" with the tag "Search". In folksonomy, users can easily find other users with similar preferences, resources and tags which have been used.

In the existing folksonomy system, most of tags are lack of semantic precision and not standard, which affect the use of it to a certain degree. In order to better organize

the information resources, the collaborative filtering [2] recommender system was introduced to folksonomy. But, the clustering result is not as good as we imagine. According to this idea, we put forward an optimized tag recommender algorithm with the users' retrieval motivation.

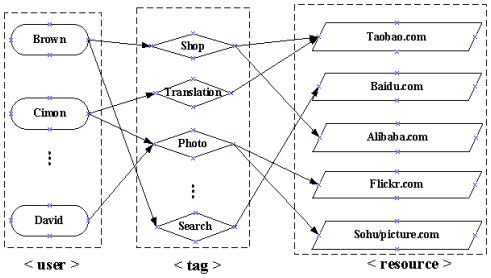

Fig. 1. the structure of a folksonomy system

First, we get a certain amount of datasets from the network, and use the stemmer technology [3] to get more effective and useful text content of tags in the datasets. Second, according to the motivation theory proposed by Strohmaier [4], we get two types of users, Categorizers and Describers [5], which have different retrieval motivations. Third, according to the new user's retrieval motivation, we judge user is a Categorizer or Describer, and then cluster the tags and resources that types of users have tagged with k-means++ algorithm [6]. Finally, according to the similarity, the algorithm give the recommended information which is more accuracy and relevant to users. The experimental results show that the optimized algorithm with user retrieval motivation can have higher accuracy and stability than traditional tag recommender algorithm in folksonomy system.

2 Preliminaries

2.1 Recommender Algorithm

Recommender system is a special system that it is always used to deal with some problems, such as information filtering. It actively provided the options to users that the options are identified as the most suitable results by recommender algorithm. And all we known, the most important part of the recommender system is recommender algorithm, the algorithm can determine the way to work and the recommended strategy. Generally speaking, recommender algorithm has three main types: Content-Based Filtering, Collaborative Filtering and Hybrid Recommender. In this paper, the optimized tag recommender algorithm belongs to the kind of collaborative filtering. The optimized tag recommender algorithm is based on the behavior or

interests of the user groups. Firstly, the recommender system will find the neighbor users, according to the historical data of the target users. Secondly, it can get the object value from the neighbor users' evaluations. Finally, according to the previous data, the recommender system will provide the personalized recommender to the target user.

The traditional collaborative recommender algorithm usually has some problems, such as concept drift and sparse, because the limitation of a algorithm is a fact. In this paper, an optimized tag recommender algorithm is proposed to improve these disadvantages based on the user's retrieval motivation. At the same time, this optimized tag recommender algorithm is proved to feasible after the experiment.

2.2 The user's tagging motivation

In this paper, we add the users' motivation theory to folksonomy system and use the theory proposed by Strohmaier. In this theory, there are two kinds of users in the datasets, and they are called categorizers and describers. Categorizer is the users who are motivated by categorization and view tagging as a means to categorize resources according to some high-level characteristics. For example, when a popular music is tagged by categorizers, they will use tag 'music' rather the tag 'song', 'tune' even if they have the similar meaning. Describers are the users who are motivated by description view tagging as a means to accurately and precisely describe resources. For example, when a popular music is tagged by describers, the tag "music", "popular", and "favorite" can be used to describe the resource. We consider the tags tagged respectively rather than the tags that can't be classified, which can give the benefit to us and improve the quality of result. In next section, we will give the model to calculate the users' motivation.

2.2.1. Measures

In this part, we use four indicators to measure the users' tagging motivation. The four measures are respectively Tags per Post (TPP), Tag Resource Ratio (TRR), Low-frequency tagging ratio (LFTR) and Interrogative adverbs tagging ratio (IATR). According to the results obtained by the above measures, we can get the type of the users, a categorizer or a describer. The detail description of the measures can be found in literature [5].

- Tags per Post (TPP)

$$\mathrm{TPP}(u) = \frac{\sum_{i=1}^{r} |T_{ui}|}{R_u} \tag{1}$$

Where R_u is the number of resources tagged by the user u, T_{ui} is the numbers of tags annotated by user u on resources i, r is the total number of resources. This measure relies on the verbosity of users. So, if the TPP reflects in a higher score, the user is more likely a Describer.

- Tag Resource Ratio (TRR)

$$TRR(u) = \frac{T_u}{R_u} \tag{2}$$

Where T_u is the number of tags annotated by the users, R_u is the number of resources annotated by the users. Because a typical Categorizer would apply only a small of tags to his resources and score a low number on this measure.

- Low-frequency tagging ratio (LFTR)

$$LFTR(u) = \frac{|T_u^0|}{|T_u|} \ , \ T_u^0 = \{t||R(t \le n)|\} \ , \ n = \left\lceil \frac{|R(t_{max})|}{100} \right\rceil \tag{3}$$

Equation 3 shows the calculation of the final measure where are seldom used tags. T_u are all tags of the given user. t_{max} denotes the tag which was tagging the most by the user. n means the critical value. If the LFTR reflects in a lower score, the user is more likely a Categorizer.

- Interrogative adverbs tagging ratio (IATR)

$$IAIR(u) = \frac{Card\ (t \in T_{str})}{|T_u|} \tag{4}$$

Where T_{str}={ what, who, when, where, …} is a set of Interrogative adverbs, $Card(t \in T_{str})$ is the number of interrogative adverbs tags annotated by user, T_u is the number of tags annotated by the users. Obviously, $IATR(u) \in (0,1)$, and if the IATR reflects in a lower score, the user is more likely a Categorizer.

2.2.2. An evaluation model for user's motivation

According to the four indicators above to distinguish the users' type, categorizers or describers, we can construct a reasonable evaluation model [7] to calculate the orientation of users' motivation below. M=a*TPP(u)+b*TRR(u)+c*LFTR(u)+d*IA--TR(u),Where, a,b,c,d $\in$(0, 1). According to the results of the experiment, we find that each datasets has a different optimal coefficient. So, we choose the average score of M and use M' to denote the value of it.

$$M' = (TPP(u) + TRR(u) + LFTR(u) + IATR(u))/4 \quad (5)$$

According to the formula 6, we find that M' is a monotonic function and it also has a threshold M_t. If M' is larger than M_t, we consider the user a Describer; On the contrary, the user is a Categorizer. If both are equal, he has not a special motivation and we treat the user either Describer or Categorizer; if the user who have little tagging behaviors and it is hard to get the users' retrieval motivation, we call them the users without motivation.

2.3 Porter Stemmer and K-means++

In the existing folksonomy system, most of tags and resources are created by users without restriction, so they are short of semantic precision and standardization, which affect the use of it to a certain degree. In order to reduce these interference factors, we use porter stemmer to deal with these tags.

Porter stemmer has five steps and each step defines a set of rules. To stem a word, the rules are tested sequentially, if one of these rules matched the current word, then the conditions attached to that rule are tested. Once a rule is accepted; the suffix is removed and the control moves to the next step. If the rule is not accepted then the next rule in the same step will be tested, until either a rule from that step is accepted or there are no more rules in that step. And the control passes to the next step. In the last step, the resultant stem is returned by the stemmer. After that, the irregular tags reduce the interference factors to understand and they can help us to improve the quality of retrieving information.

At the same time, according to the model of the users' motivation, we can get the types of users. Then, we need to deal with the resources tagged by users who have the similar motivation and find the similar resources which have been tagged. In this paper, we choose the k-means++ clustering [8].

The k-means++ is an algorithm for choosing the initial values for the k-means clustering algorithm. It was proposed in 2007 by David Arthur, as an approximation algorithm for the NP-hard k-means problem—a way of avoiding the sometimes poor clustering found by the standard k-means algorithm. And the k-means problem is to find cluster centers that minimize the intra-class variance, the sum of squared distances from each data point being clustered to its cluster center.

3 The optimized algorithm

According to the knowledge given above, we propose an optimized algorithm with users' retrieval motivation in folksonomy system. The detail procedure of the algorithm is shown in Figure 2, and the detail algorithm is shown in Table 1.

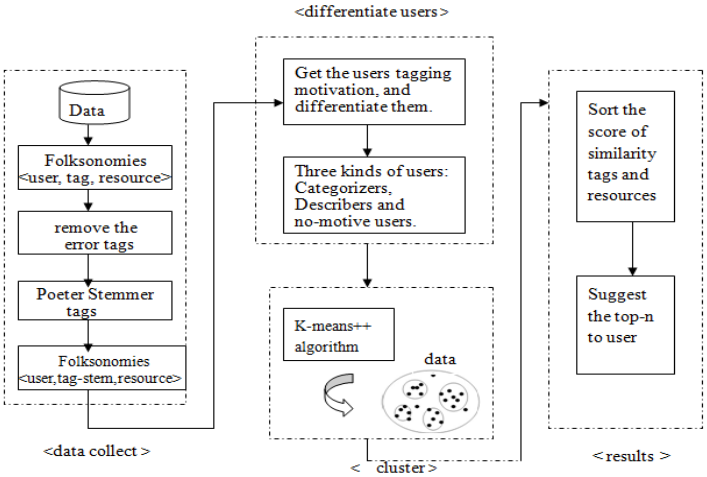

Fig. 2. The procedure of the algorithm

Table 1. The optimized algorithm

An optimized algorithm with users' retrieval motivation:

Input ： a folksonomy datasets，the cluster center k，a retrieved tag t_u by user u.
Output：some tags(to user u).

1) get $T(t_1, t_2, \ldots, t_n)$ from a folksonomy datasets;
2) for each t_i ($t_i \in T$) ;
3) if { t_i has some problem;
4) remove; }
5) else { porter stemmer;
6) get the stem t_j of t_i;
7) $Q \leftarrow t_j$; }
8) end;
9) for each t_j, $t_j \in Q$;
10) Switch(the $t_j(user_j)$ tagging motivation M_{uj}')
11) { case 'categoritizers': k-means++($Q'(t_j \in$ categoritizers));
12) find(sim max(Q_{uj}',t_j))→L;
13) tf-idf L, return the top-n to u ;
14) case 'describers': k-means++($Q'(t_j \in$ describers));
15) find(sim max(Q_{uj}',t_j))→L;
16) tf-idf L, return the top-n to u ;
17) case 'no-motive users': k-means++($Q'(t_j \in Q)$);
18) find(sim max(Q_{uj}',t_j))→L;
19) tf-idf L, return the top-n to u;
20) deefault: system.out.print("Absence recommender.")
21) end;

There are three modules in this procedure. First, in the module of data collection, we get the data from the internet and some preprocessing will be done to get the more useful data which have less interference factors. When we get the data, we need to remove the error tags. Then, we can reduce the number of the tags which have the similar semantics by using the stemming technique that we introduced above. And then, some relatively good data is selected for the next module, with which we can ensure the effectiveness of the text data and reduce the number of the text data which has the little influence on the results in the experiment. Second part, after we get the relatively good data, we classify the users into different types. According to the model given above, we can get the M' score, the value of orientation of users' motivation, with which we can get a ranked lists of users, where Categorizers rank high, and Describers rank low. With the two types of users and the tags they have tagged, we can easily find the resources which have tagged by the users with the similar motivation. And this can improve the efficiency of retrieving information when the system recommends resources to users. Third part, after we get the types of users, we need to get the smaller range of search resources, we use the k-means++ algorithm to cluster these resources tagged by users who have the similar motivation. After the data cluster, we calculate the similarity between the retrieved information and the clustered data and then return the resources which have the high similarity to the user. But, for users who have little tagging behaviors and it is hard for us to get the users' retrieval motivation, we only use TF-IDF algorithm to find the similar resources, and then recommend them to the user.

4 Experiments

According to the optimized algorithm proposed above, we need to collect the datasets first. We use the datasets downloaded from http://www.flickr.com in our experiments. This datasets has 4 classifications and 3000 documents. The classifications contain the following types: Resources, Tags, users, Messages of the photo.

In this experiment, we just use three classifications which are Resources, Tags and User. After data processing in the module of data collection, we get the 89 resources, 2537 users and 8478 tags.

4.1 Evaluation Measures

To evaluate the proposed approach, we introduce three measures, precision, recall and F-measure to evaluate the quality of retrieval by different methods, which are defined as follows.

Precision. In the field domain of information retrieval, precision is the fraction of retrieved instances that are relevant, and it is also used with recall. The definition is:

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\text{retrieved documents}|} \tag{6}$$

Recall. Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved. The definition is:

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\text{relevant documents}|} \tag{7}$$

F-measure. Generally speaking, a good retrieval algorithm should be able to have a higher score in precision and recall. Therefore, we use the compromise value to measure, such as F-measure(β=1), that it indicates the precision and recall are equally important now. The formula is shown:

$$F_{(\beta=1)} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{8}$$

If F-measure score is higher, it means that this retrieval algorithm is better than the other.

4.2 Results and Discussion

Compare the F-measure score of three different algorithms to measure them. In Algorithm 1, we use the traditional collaborative recommender algorithm. In Algorithm 2, we add the user motivation theory to the algorithm and use the value of orientation of users' motivation to distinguish the users' type, and then cluster the resources tagged by the users who have the similar retrieval motivation with k-means++ algorithm, and then use the same way as Algorithm 1 to get the results f-measure. In Algorithm 3, different with Algorithm 2, the value of orientation of users' motivation is not used for distinguishing the users' type, but it is used as an auxiliary value of the calculation of the similarity between relevant the tags and the tags which will be retrieved. Finally, we use the same way as Algorithm 1to get the f-measure of result.

The three algorithms can be regarded as the three models of the tag recommender systems. Then we choose the three types of users, Describers, Categorizers and users with no motivation. Let the three models recommend the required source retrieved by three types of users respectively. For the results the models recommend to the users, we take the average of the results after five times' test in order to reduce the experimental error. We can see the results from Figure3, Figure4 and Figure 5. In Figure 3, the users are all the Describers. At the same time, we can see that the F-measure value in Algorithm 2 and 3 have the higher score than the Algorithm 1, which means that for the user of Describers, the algorithms with motivation theory

have the better results. In Figure 4, the users are all the Categorizers Meanwhile, and the F-measure value in Algorithm 2 and 3 also have the higher score than the Algorithm 1, which means that the algorithm with users' motivation have a better result too. In Figure 5, the users are all those without retrieval motivation. So, we cannot use the users' motivation as an auxiliary value in Algorithm 2 and 3. For this reason, the three algorithms have the similar results the recommender system gives to the users and the value of F-measure stay the same level intuitively.

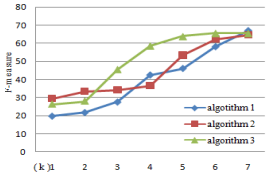

Fig. 3. Compared the results based on Describers

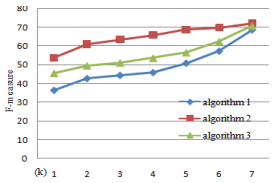

Fig. 4. Compared the results based on Categorizers

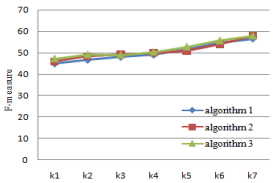

Fig. 5. Compared the results based on the sparse users

And then, we random chose a few users to test the optimized algorithm and no longer deliberately distinguish the type of the users. Then, we can see the results from table 2.

Table 2. The F-measure score(%) results at random

The F-measure score(%)	User1(k=260)		User2(k=260)		User3(k=260)	
	Tag1	Tag2	Tag3	Tag4	Tag5	Tag6
Algorithm 1	56.2	57.0	62.4	71.6	46.2	58.2
Algorithm 2	56.3	57.0	69.1	79.4	53.3	62.3
Algorithm 3	56.3	57.1	66.1	77.7	65.7	65.5

From the analysis above, we can conclude that in tag recommender system, the optimized algorithm with users' retrieval motivation can have the better results than the traditional algorithm without users' retrieval motivation when recommend the resource to the users.

5 Conclusions

In folksonomy system, the traditional tag recommender algorithm [9] need to deal with the resources tagged by all the users. For improving the efficiency of the recommender, the users' motivation is added to our tag recommender algorithm. When users retrieved the information, their motivations were given to the model. Then, we just need to deal with the resources tagged by the users who has the similar

motivation, and this need less time just to deal with the more relevant resources, meanwhile, the accuracy of resources recommend was improved. The experimental results show that this algorithm with user retrieval motivation can have higher accuracy and stability than traditional retrieval algorithms when recommend the resources to users.

6 Acknowledgement

This work is supported by National Natural Science Foundation of China (grant 61163057), Guangxi Nature Science Foundation (grant 2012jjAAG0063), Open Fund of Guangxi Key Laboratory of Trusted Software (kx201308). The authors would also like to express their gratitude to the anonymous reviewers for providing helpful suggestions.

References

1. Vander Wal Thoms. “Folksonomy Coinage and Definition”. Retrieved 2013-03
2. J.A.Konstan, J.Riedl. “Recommended for you,” Spectrum, IEEE, vol.49, no.10, pp.54-61, October 2012.
3. Wahiba Ben Abdessalem Karaa. “A new stemmer to improve information retrieval” [J]. International Journal of Network Sccurity (IJMIT), Vol No.4, July 2013
4. M.Strohmaier, C.Körner, and R.Kern. “Why do users tag? Detecting users' motivation for tagging in social tagging systems”. In International AAAI Conference on Web blogs and Social Media (ICWSM2010), Washington, DC, USA, May 2010
5. A.Zubiaga, C.Körner and M.Strohmaier. Tags vs Shelves: From Social Tagging to Social Classification [C]. In: Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia. New York, NY, USA: ACM, 2011: 93-102
6. J.GEMMELL, RAMEZANIM, et al. The impact of ambiguity and redundancy on tag recommender in folksonomies [C]// Proceedings of the 2009 ACM Conference on Recommender Systems. New York: ACM Press, 2009: 23-25
7. C.Trattner, C.Körner, D.Helic. Enhancing the Navigability of Social Tagging Systems with Tag Taxonomies. In Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, ACM, New York, NY, USA, 2011
8. Ma Li. Immune Network Based Text Clustering Algorithm. [M].Software Engineering, Artificial Intelligence, Networking and Parallel & Distributed Computing (SNPD), 2012 13th ACIS International Conference on,2012
9. HARVEY M, BAILLIE M, RUTHVEN I,et al. Tripartite hidden topic models for personalized tag suggestion [C] // Proc of the 32nd European Conference on IR Research. Spring Berlin:2010:432-443