Construction of Information Resources Organization System Based on Medical Portal Using Folksonomy

Shuaiqi Yang [a], Runtong Zhang [b]

School of economics and management, Beijing Jiaotong University, Beijing, 100000, China

[a]shuaiqiyoung@163.com, [b]RuntongZhang@bjtu.edu.cn

Abstract. With medical portal as application scenarios, we proposed a method of constructing information resource organization system using folksonomy, the first to use subject terms set from medical portal, clustering labels that the public create under the appropriate subject term, according to the label cluster similarity ranks ,and forming some similar type of cluster, Then we generate labels having different levels of the tree structure, using the improved K-means clustering algorithm clusters within each class, select several nodes of the tree structure and tag number of levels ,last form information resource organization system for medical portal.

Keywords: Medical Portal; Information resources system; Folksonomy; K-means algorithm.

1. Introduction

To build the Information Resource Organization System website by selecting the number of nodes and tag tree level is the effective way to improve website navigation and information resources semantic in nature, tag tree hierarchy reflected in the setting of a site's directory hierarchy. Therefore, many tag tree generation methods have been proposed, M. Strohmaier et al., The current tag tree generation method extensively detailed comparison [12], the results show, P. Heymann-based label refers to the method of [9] semantics were better than the other methods are compared. Ch. Luo, who presented the tag tree generation method as a benchmark, the use of K-means clustering algorithm, clustering analysis between labels, to some extent solves the semantic ambiguity between labels, including classification relationship inaccuracies. However, in this paper, in order for the medical portal as a specific application scenarios, this method still has many deficiencies, specifically in the following aspects:

First, as a medical portal available to the public health and medical information platform, authoritative information and precision, classification of some resources may relate to a specific medical vocabulary, but the method described above in a node by the user tags for a simple cluster analysis, the independent division of a number of class to the user's label as a directory site resources, and to some extent undermined the authority and accuracy of medical information resources portal. Secondly, according to the above method of generating tag tree is complex, multi-characteristic, and performing cluster analysis process, just a tag tree node cluster analysis, which is likely to lead to other similar labels high coupling phenomenon between the trees, which led to two or three information resources directory represents a large number of overlapping resources does not completely cover the parent node represents.

2. Optimization algorithm based on Heymann method

2.1 Improved clustering method based on label.

To make Heymann method better suited for classification of medical information resource portal, forming a complete structure, semantic clear resource classification system, we use clustering techniques to improve methods for Heymann, ideas for improvement:

(1) relating to the use of keywords set taxonomy classifies formed, and these keywords tag tree as a first-level nodes, according to the similarity of each label to the next time you add keywords, which not only ensure the authority and information each class precision also makes cluster internal sense closer, information resource corresponding decrease in the degree of coupling between classes.

© 2016. The authors - Published by Atlantis Press

(2) Construction of the label subtree Heymann methods within each category, but with the "resource coverage" (RC) network instead of the original center of the label as an index refers to the degree.

(3) All the sub-tree tags placed together to form a complete tag tree.

2.2 Basic symbol definitions.

Social tagging data can be represented as a collection of triples, $D=\{\langle u,r,t\rangle | u\in U, r\in R, t\in T\}$.U, R and T respectively users, resources, and label collection. $M_{m\times n}(m=|R|, n=|T|)$, the elements m_{ij} represents the resource i is the number of different users use tags marked j, $m_{ij}=|\{\langle u,r,t | r=r_i, t=t_j\}|$. Each column of M can see the role of the resource vector representation label, two M $m_{ij}=|\{\langle u,r,t | r=r_i, t=t_j\}|$ calculated amount of the cosine similarity, $c(t_j)=|\{m_{ij} | m_{ij}>0\}|$. The similarity may be obtained the corresponding label. Denotes the number of resources are marked with a tag t_j.

2.3 K-means clustering algorithm improvements.

On improved K-means clustering algorithm tag, the process is as follows:

(1) Set the initial value

Each specific medical portal has a unique set of keywords, represented by the A_i, i = 1, 2, 3.....m. The algorithm can be seen in the K is m value, which is divided into m classes, each class as a sub-tree root node denoted Fi. After a free cluster system, eventually identified as the tag number n, the label is defined as T_j, where j = 1, 2, 3 ..... n, the similarity variable RC (A_i, T_j), referred to as RC_{ij}, after clustering category is C_i.

(2) Select the similarity calculation formula

A_1 input keywords and tags T_1, T_1 tag number to obtain resources through dataset D gets M (T_1), M (A_1) and M (A_1, T_1) value, calculated according to the similarity of the above-mentioned formula A_1 T1 and the similarity of RC (A_1, T_1).

(3) calculate the similarity

Accordance with the instructions in Step 2, were calculated similarity RC_{i1} ,namely,T_1 and other m-1 Ge keywords, where i = 2,3,4 ... ..m, depending on RC_{i1} value T_1 is allocated to its most similar a MeSH class. Similarly, calculate the similarity tag T_2, T_3 ..... Tn and m keywords. This step can be the property of each label into a keywords. Under the assumption that each cluster contains keywords Ai Ni after the label, including $\sum_{i=1}^{m} Ni$ = n.

(4) Sort by similarity

Step 3 under each tag has a number of keywords, tags and keywords based on similarity in descending order of size, choose keywords with the highest degree of polymerization of a label, referred to as Tmax, and place it in the child Fi under the tree, except for Tmax label denoted Tk [max], calculated Tk [max] subtree Fi similarity of all the nodes, and the current label and put it under a node having the greatest similarity. After calculation, to each tag are placed in a particular subtree Fi, forming a complete structure, highly polymerized tag tree.

In the construction of the medical portal, considering the user's habits, information organization, classification of finesse and other factors, general web directory is set to 3-4 layers, so only part of the interception of the organizational system tag tree configuration information resource website. Select five highest similarity level sub-set of labels under various keywords as a secondary directory portal, denoted Pik, where i = 1, 2, 3, ...m, k = 1,2,3 ... ... 5, the order of similarity label is RC (Pi1> Pi2> Pi3> ..... Pi5), these five tag and its corresponding keywords Ai combined into one group, denoted by Ci, in the same manner, for each a secondary directory Pik, selection and similarity ranks high as three five tabs directory portal, denoted by Qik, where where i = 1, 2, 3, ...m, k = 1, 2, 3 ..... 5, the order of similarity label is RC (Qi1> Qi2> Qi3> ..... Qi5), then Ci, Pik finally be expressed as:

$$C_i = \begin{Bmatrix} C_1 \\ C_2 \\ C_3 \\ . \\ . \\ C_m \end{Bmatrix} = \begin{Bmatrix} A_1 & P_{11} & P_{12} & . & . & P_{1k} \\ A_2 & P_{21} & P_{22} & . & . & P_{2k} \\ A_3 & P_{31} & P_{32} & . & . & P_{3k} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ A_m & P_{m1} & P_{m2} & . & . & P_{mk} \end{Bmatrix}, \quad P_{ik} = \begin{Bmatrix} P_{1k} & Q_{11} & Q_{12} & . & . & Q_{1k} \\ P_{2k} & Q_{21} & Q_{22} & . & . & Q_{2k} \\ P_{3k} & Q_{31} & Q_{32} & . & . & Q_{3k} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ P_{mk} & Q_{m1} & Q_{m2} & . & . & Q_{mk} \end{Bmatrix}$$

3. Simulation and Competition

3.1 Evaluation Method and Evaluation Index.

In view of research focused on clustering method improvement in the method validation process, data collection is not limited to a specific set of information resources. Therefore, this method validation data set used is A. Zubiaga, who collected the Social - ODP - 2k9, the data set is widely used, has a certain universality, the results are more convincing. In carrying out the original Heymann method uses three different index refers to that degree, betweenness and closeness, which correspond respectively to give the tag tree denoted H_DEG, H_BET, and H_CLO.

(1) Semantic Evaluation Criteria

Evaluation Semantics has a lot of practice, chose M. Reference (reference-based) based approach is about to get tagged with the existing tree, recognized classification were compared Strohmaier et al used in the study.

$$tp(c, AT, RT) = \frac{\left|\mathrm{ce(c,AT)} \cap \mathrm{ce(c,RT)}\right|}{\left|\mathrm{ce(c,AT)}\right|} \tag{1}$$

$$tr(c, AT, RT) = \frac{\left|\mathrm{ce(c,AT)} \cap \mathrm{ce(c,RT)}\right|}{\left|\mathrm{ce(c,RT)}\right|} \tag{2}$$

$$to(c, AT, RT) = \frac{\left|\mathrm{ce(c,AT)} \cap \mathrm{ce(c,RT)}\right|}{\left|ce(c, AT) \cup \mathrm{ce(c,RT)}\right|} \tag{3}$$

$$TP(AT, RT) = \frac{1}{C_{AT} \cap C_{RT}} \sum c \in \left|C_{AT} \cap C_{RT}\right| tp(c, AT, RT) \tag{4}$$

The final TP, TR, TF and TO all the parameter values of the concept of the mean:

$$TR(AT, RT) = \frac{1}{C_{AT} \cap C_{RT}} \sum c \in \left|C_{AT} \cap C_{RT}\right| tr(c, AT, RT) \tag{5}$$

$$TF(AT, RT) = \frac{2 \times TP(AT, RT) \times TR(AT, RT)}{TP(AT, RT) + TR(AT, RT)} \tag{6}$$

$$TO(AT, RT) = \frac{1}{C_{AT} \cap C_{RT}} \sum c \in \left|C_{AT} \cap C_{RT}\right| to(c, AT, RT) \tag{7}$$

(2)Navigation Evaluation

Measure the success rate and efficiency of navigation is defined as follows, obviously SR (Ci) the greater the success rate, the higher the NE (Ci) smaller navigation efficiency.

$$SR(Ci) = N_{suc} / N \tag{8}$$

$$NE(C_{\mathrm{i}}) = \frac{1}{N_{suc}} \sum_{i=0}^{N_{suc}} W_i \tag{9}$$

In this study, the author select tags in M cosine value of the corresponding vector to calculate the similarity, so as to realize simulated users browse search process.

3.2 The Evaluation Results Analysis.

(1)Semantic Analysis Evaluation

Improvement methods are verified through the experiment many times has a better effect, to run multiple times of the improved method, the results give the worst result, the average results, and as a result, the best record for CLU_H_MIN, CLU_H_AVG, CLU_H_MAX respectively, using the aforementioned semantic evaluation index formula (4), (5), (6), (7) and matrix, the generated data processing elements of mij said resources I used by different users tag j indicate the number of times, that is, through calculation results as shown in table 1.

Table 1 improved algorithm evaluation index compared with the original algorithm

	TP	TR	TF	TO
CLU_H_MAX	0.018	0.033	0.024	0.012
CLU_H_AVG	0.017	0.032	0.022	0.010
CLU_H_MIN	0.016	0.031	0.020	0.008
H_RC	0.015	0.030	0.019	0.008
H_DEG	0.013	0.030	0.017	0.007
H_BET	0.012	0.026	0.016	0.007
H_CLO	0.007	0.019	0.011	0.005

Table1 for semantic evaluation result, it can be seen that H_RC effect on four indexes are better than H_DEG, H_BET and H_CLO, suggesting that "resources coverage" as a label "to denote degree" refers to Mark has a good performance. Can be seen from the table, CLU_H corresponding TO the maximum, minimum, average three generate results in TP, TF and TO be higher than the original Heymann method, it shows that the improved method effectively improves the accuracy of semantics.

(2) Analysis of the results of the evaluation navigation

The results obtained by calculation shown in Table 2 and Table 3.

Table 2 Improved navigation algorithm and the original algorithm success rate Comparison

	CLU_ H_MAX	CLU_ H_MIN	CLU_ H_AVG	H_RC	H_DEG	H_BET	H_CLO
$SR(C_i)$	0.100	0.080	0.065	0.043	0.025	0.021	0.038

Table 3 Improved navigation algorithm and the original algorithm efficiency Comparison

	CLU_ H_MAX	CLU_ H_MIN	CLU_ H_AVG	H_RC	H_DEG	H_BET	H_CLO
$NE(C_i)$	50	50	52	54	52	70	60

Table 2 and Table 3 shows the results of the evaluation were the success rate and efficiency of navigation, including SR (Ci) the greater the success rate, the higher the NE (Ci) smaller navigation efficiency. According to the above table, the results can be obtained:

(1) H_RC respect H_DEG, H_BET, H_CLO has a high success rate and efficiency, which explains at the same tag tree-building method, using RC as the label refers to the index has better performance;.

(2) CLU_H on the navigation success rate was better than H_RC, H_DEG, H_BET, H_CLO, and on the navigation efficiency can achieve very good level, indicating that the improved method has better navigability.

4. Summary

By clustering tag, a large number of related semantic tags, each comprising a plurality of the following keywords with the highest similarity keywords tag, on the one hand, the original label disorder increases the relevance and level of semantic making the label more closely linked. On the other hand, the label from the label user behavior, fully reflects the actual needs of users, it can provide users more detailed categories, easy for users to find information. In short, both to ensure the authority and accuracy of medical information portal of the organization of information resources, and improve the degree of involvement of users, both combined with each other, influence each other, work together to build a website of information resource classification system.

Acknowledgments

Supported by the State Key Program of National Natural Science of China (Grant No. 71532002).

References

[1] Cui Jianwei, Liu Hongyan, He Jun, et al. TagClus: A random walk-based method for tag clustering [J]. Knowledge and Information Systems, Vol. 27(2011) No. 2, p. 193-225.

[2] D. Ramage, P. Heymann, C. D. Manning, Clustering the tagged Web, Proceedings of the Second ACM International Conference on Web Search and Data Mining. New York: ACM, 2009, p.54-63.

[3] Zubiaga A, Fresno V. getting the most out of social annotations for Web page classification. Proceedings of the 9th ACM Symposium on Document Engineering. New York: ACM, 2009, p. 74-83.

[4] P. Heymann, G. Koutrika. Can social bookmarking improve Web search. Proceedings of the International Conference on Web Search and Web Data Mining. New York: ACM, 2008, p. 195-206.

[5] G. Begelman, P. Keller, F. Smadja. Automated tag clustering: Improving search and exploration in the tag space. Collaborative Web Tagging Workshop at WWW2006. New York: ACM, 2006.p. 15-33.

[6] Song Yang, Qiu Baojun, Farooq U. Hierarchical tag visualization and application for tag recommendations, Proceedings of the 20th ACM International Conference on Information and Knowledge Management. New York: ACM, 2011, p. 1331-1340.