The Research of Marine Information Clustering Algorithm Based on User-Browsing Path and Associated Query

Yanping Cong
College of Information Science and Engineering
Ocean University of China
Qingdao, China
congyp@ouc.edu.cn

Zhiqiang Wei
College of Information Science and Engineering
Ocean University of China
Qingdao, China
weizhiqiang@ouc.edu.cn

Miaoqing Tian
College of Information Science and Engineering
Ocean University of China
Qingdao, China
tian_miaoqing@qq.com

Yongquan Yang
College of Information Science and Engineering
Ocean University of China
Qingdao, China
i@yangyongquan.com

Abstract—With the enrichment of Marine information on the Internet, the users find that it is hard to find the knowledge they need when they face with a deluge of Marine information resources, they wish to acquire marine information data rapidly which is integrated and desirable. Recently, researchers have been committed to solve these problems with the Web Clustering Mining, according to the user's access action, looking for patterns of behavior or retrieval of users with similar interests. Aiming at the deficiency in the current problem of calculating similarity between users based on user sessions. This paper proposes an algorithm to use associated queries about query logs as the compensation characteristics to measure the similarity between users, combined with user-browsing path algorithm to clustering the user data. The experimental results of the algorithm have a good reference value for users of personalized service and promote the development of marine information.

Keywords- Marine information; Web Clustering Mining; users query log; associated queries; user-browsing path; (key words)

I. INTRODUCTION

The continuous growth in the size and use of the World Wide Web (WWW) requires new methods of design and development of online information retrieval systems[10]. With the increasing of marine information on the Internet, the users have to spend plenty of time to search the marine information they need. How to find these information forms the vast marine information resources quickly and accurately has become a big problem to Internet users. Although the Internet search engines have solved this kind of problem to a certain extent, it is undeniable that traditional information services have been unable to meet the requirements of personalized and intelligent: the demand of Internet is growing fast, and the user had divided into many different types of groups. They will ask the Internet to provide more personalized service according to their knowledge background, interests and hobbies. The Internet is not only just a simple resources platform for the demand of users, but an intelligent resource recommend platform.

The Internet has been popularly known as an extremely huge data repository consisting of much data types as well as a large amount of unseen informative knowledge, which can be discovered via a wide range of data mining or machine learning paradigms [1].The Internet could get information on a simple browsing behavior or matching the keyword from the search engines, but these external information as we have seen, is messy and isolated. In fact the Internet hide a lot of internal information. Data mining is a method to find these potential and important information. Clustering of Web mining is a new technology base on traditional data mining technology, combined with the characteristics of Web information, it could identify the user's interest characteristics and access patterns, and predict user behavior trends, find out the potential and valuable knowledge. Clustering of Web mining can improve the quality of the search engine, provide users with more personalized service, found potential user bases, meet the needs of users and achieve the Internet product business interests.

In this paper, we cluster the user data onto statistical regularities of Web log, and we use a user similarity computing method to get the user similarity matrix based on the associated query analysis[2], then on the basis of this we use user browsing path algorithm for clustering. In the end we will use this clustering algorithm on the search engines of Marine monitoring platforms which developed by our laboratory.

II. RELATED WORK

As an emerging research field of data mining, Web Clustering mining mainly concentrated in two aspects- the clustering of similar user groups and clustering of Web pages. Web user clustering mining mainly through the Web log mining to extract the user browse information on users clustering. Many typical clustering algorithms have been put forward by scholars at home and abroad. Reference [3] presented a simulated annealing approach to clustering. They combined their heuristic with artificial neural networks to improve solution quality. Different from widely-known methods, they preferred to use a similarity criterion function called the XB cluster validity index. Reference [5] proposed a hybridized approach that combines the K-means algorithm, as in [4], Nelder–Mead simplex search and PSO technique. The K-means algorithm is a well-known approach to clustering. Its popularity depends on its simplicity and computational efficiency. However, that approach tends to fixate on local optima near the initial cluster centers, which are assigned randomly. Thus, many researchers have presented heuristic clustering algorithms to overcome this problem. Reference [6]presented the artificial bee colony (ABC) as a state-of-the-art approach to clustering. Reference [7] proposed an approach based on a vector space model, called Random Indexing, to discover the latent factors or hidden relationship among users' navigational behaviour, and the clustering results are used to predict and prefetch Web requests for grouped users. Reference [8]proposed K-path clustering method, they think that user access to Web sites means they are interest in this Web. Reference[9] proposed a personalized recommendation approach joins the user clustering technology and item clustering technology to solve the problems of scalability and sparsity in the collaborative filtering. These clustering methods have its own advantages, and similarity algorithm in user clustering method is also very important. Reference [2] proposed a method of automatic query expansion based on user interactions recorded in query logs. This method through associated queries to get the relationship between two query words .

User-browsing path a kind of clustering algorithm based on User-browsing path content. This user clustering method in pay attention to user's interests, at the same time reduce the dimension of user, and it is a clustering method to clustering user based on path similarity matrix. On the basis of this method, we found that it will have a better effect if we replace the path similarity matrices by user similarity matrix, then users browsing paths clustering algorithms can be implemented on users clustering.

III. THE CLUSTERING PROCESS

Because of the similarity calculation of data is closely related to data expression of this section we set out from two aspects, the similarity calculation and clustering method.

A. User similarity calculation

Due to clustering algorithm is to achieve similar data division, and each internal data onto clustering is required to as similar as possible, and each clustering should be different as far as possible. So it is very important to define a scale to measure the similarity. In general, there are two ways to define the similarity. The first method is to define the distance between data, describe the difference in data. The second one is to define the similarity between the data directly.

Common measurement methods of the distance between data points are Ming's distance, Markov distance and cosine distance, what we use most commonly used are Ming's distance, its calculating methods as in (1).

$$d_{ij}(q) = \left(\sum_{K=1}^{p} \left| X_{ik} - X_{jk} \right|^{q} \right)^{1/q} \qquad (1)$$

The method commonly used in Euclidean space distance measurement, the formula is Manhattan distance when q=1, and when q=2 the formula is the Euclidean distance, when q is infinite, the formula is the Chebyshev distance. The Markov distance is commonly used to measure the distance between the random variable; the Cosine distance is using cosine of angle between two vectors to represent its similarity. According to the different data calculation method, when the distance between the two data points is d, we can use the following (2) to calculate the similarity:

$$\mathrm{sim1} = \frac{1}{d + 1} \qquad (2)$$

User similarity calculation based on the retrieval log mining is calculated the similarity between user retrieval interests according to the user session in the log information. Each user will have and maintain a query session set is $S=(s_1,s_2,\ldots,s_n)$,and each session of the set is $s=(query,(u_1,u_2,\ldots u_n))$,the *query* is user's query, the u_i on behalf of the results returned by click on the web *URL*. We use the click sequence of result pages as the session characteristics, and we have two different user session $s_i=(q_i , U_i)$ and $s_j=(q_j , U_j)$, the q and U represent the query in the session and web page collection, if $q_i=q_j$, then we can use the Dice coefficient to calculate the similarity, as in (3):

$$sim2 = sim(s_i, s_j) = \frac{2\left|U_i \cap U_j\right|}{\left|U_i\right| + \left|U_j\right|} \qquad (3)$$

When the query of the two user session is different, like $q_i \neq q_j$, its similarity is 0. We use (3) to calculated the similarity when $q_i=q_j$. Then, the similarity of two users, can be gained through the accumulation of its similarity query session.

The defects of calculation method above are only calculated with the same query word similarity between user sessions. For example, the query of two user sessions are "seaweed" and " algae" respectively, obviously the two user retrieval interest is similar. So we need to solve different queries to calculate the similarity between the user's session. We found that the query correlation [2] can find the hidden link between the two queries of user logs, this relationship can be used to measure similar degree of different user session queries.

We defined the query correlation degree as a conditional probability between the query words, such as one query is q_i, another query is q_j, and the correlation q_j relative to q_i is conditional probability $P(q_j \mid q_i)$. The

calculation of the conditional probability can be derived by (4):

$$P(q_j|q_i) = \frac{P(q_j, q_i)}{p(q_i)} = \frac{\sum_{\forall d_k \in D} P(q_j, q_i, d_k)}{P(q_i)} = \frac{\sum_{\forall d_k \in D} P(q_j|q_i, d_k) \times P(q_i, d_k)}{P(q_i)} \quad (4)$$

Here we assume that $P(q_j \mid q_i, d_k) = P(q_j \mid d_k)$, the reason is that the link between the query is built by document and query relationship, namely d_k split q_i and q_j. Then we can get (5):

$$P(q_j|q_i) = \frac{\sum_{\forall d_k \in D} P(q_j|d_k) \times P(d_k|q_i) \times P(q_i)}{P(q_i)} = \sum_{\forall d_k \in D} P(q_j|d_k) \times P(d_k|q_i) \quad (5)$$

In formula(8), $P(d_k \mid q_i)$ is the conditional probability of clicked document dk, when the query is qi in the (5), and $P(q_j \mid d_k)$ is the conditional probability of q_j, when the clicked document is d_k. The two conditional probability can be estimated by following, as in (6)

$$P(d_k|q_i) = \frac{f(q_i, d_k)}{f(q_i)} \qquad P(q_j|d_k) = \frac{f(q_j, d_k)}{f(d_k)} \quad (6)$$

In formula (6), $f(q_i, d_k)$ is the number of user session contains document d_k when the query is q_i, and the $f(q_i)$ is the number of user session when the query is qi. The $f(q_j, d_k)$ is the number of user session contains document d_k when the query is q_j, and $f(d_k)$ is the number of document d_k include all the user session.

Due to the asymmetry of the query correlation calculation, we define the similarity between the two queries as in (7):

$$sim(q_i, q_j) = \frac{p(q_i \mid q_j) + p(q_j \mid q_i)}{2} \quad (7)$$

When two user session query words are different, but related to each other, we used this formula instead of similarity of user sessions. When two sessions of the query words are neither the same nor association, we define the similarity of 0.

According to the definition above, when the user session set is $S=(s1,s2,\ldots,sn)$ and $S'=(s1',s2',\ldots,sn')$ respectively, the similarity calculation formula is (8):

$$Sim(S, S') = \frac{1}{m * n} \sum_{i=1}^{n} \sum_{j=1}^{m} sim(s_i, s_j') \quad (8)$$

According to the formula, we can get the user similarity matrix for clustering.

B. User clustering

After we get the similarity matrix of users, then we need for user clustering, in this paper we use the User-browsing path clustering algorithm. The user-browsing path clustering algorithm is a clustering method of user paths, it's a path clustering according to the path similarity matrix. In this paper we use it for user clustering, we found that users browsing paths clustering algorithm can be implemented on user clustering and has good effect when we use user similarity matrix replace the path similarity matrix.

The user-browsing path clustering algorithm is a kind of clustering algorithm with high practicability ,it can deal with high degree of data and processing a large number of Web users paths. We use S represent matrix, and S_{ij} is the elements in the matrix, The *Sij* is the similarity between objects i and j. .Given a threshold θ to structure similar class, each row of the matrix will produce a class *Ci*.

$$C_i = \{S_{ij} | S_{ij} \in S_i, S_{ij} \geq \theta\} \quad (9)$$

Because the user similarity matrix does not have transitivity, so the initial result is similar class not equivalence class, that to say, possible intersection may exist in these similar class , or there may be repeated, such as $Sij \in Si$. So we must remove the repetition of similar classes, and all kinds of intersection between items.

Users browse path clustering algorithm as shown in Fig. 1:

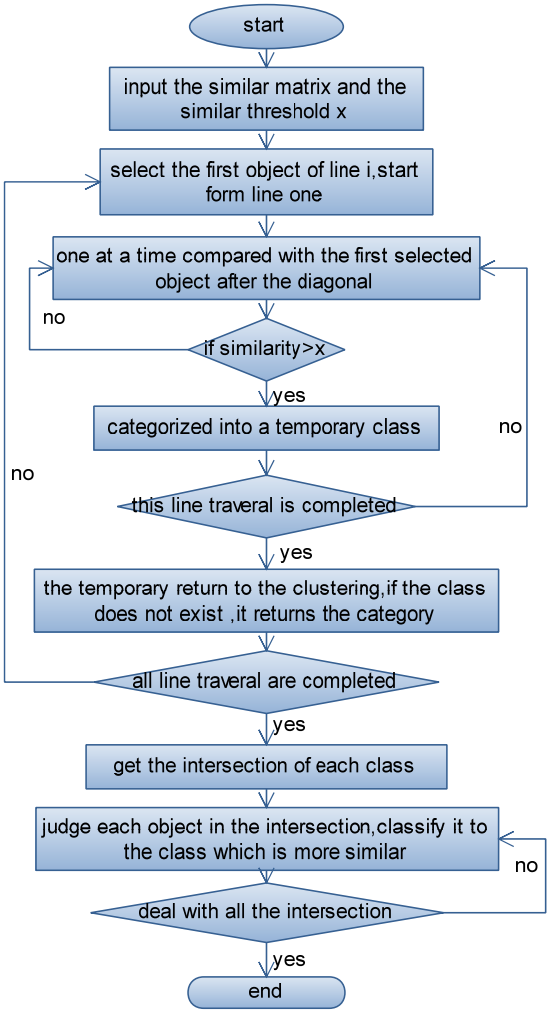

Figure 1. The clustering process flow diagram

IV. EXPERIMENTAL RESULTS

This paper conduct the validation of the algorithm on window 7 operating systems and MySQL 5.6.16. First of all, we structure experimental data setted from the Sogou user query logs, To compare the results of clustering, we choose five themes drawn from the log session information, they are "seaweed", "scallop", "tidal power", "pilchard" and "red tide", etc. The principle of extracting is that these themes have clearly meaning, and it can highlight the user interest, used by high frequency, with a large amount of historical use of information, and with more related query. Then for each theme class ,we add same topic interest query words, such as "seaweed" class ,we add "varek", "algae" etc. which interested in seaweed information query. Then choose 200 user sessions according to the query words for each topic class. Because the Sogou log does not have IP information, then randomly divided the 200 user sessions into 20 set, to simulate 20 users that are interested in this topic. Then five theme classes could generate much 100 users access information.

In order to compare the performance of user similarity calculation method. We use users-web access matrix, and the matrix constructed by accessing information, rows and columns represent the users and web pages. The elements in the matrix which is the frequency of user access to the web pages, then the degree of similarity between the user can be calculated by the cosine of row vector angle, we called it cosine similarity. We use the user similarity method which retrieves the session information without using query association compares with the similarity method using query association. Then we use the User-browsing path clustering algorithm to cluster the two group of samples data, the clustering results as show in Fig. 2:

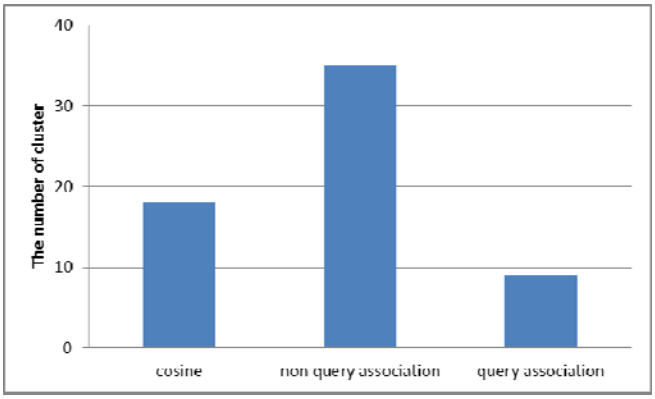

Figure 2. Compare the three clustering methods

In order to compare the performance of clustering, we use Fowlkes - Mallows (FM) index to evaluate the clustering results. The results as shown in Fig. 3:

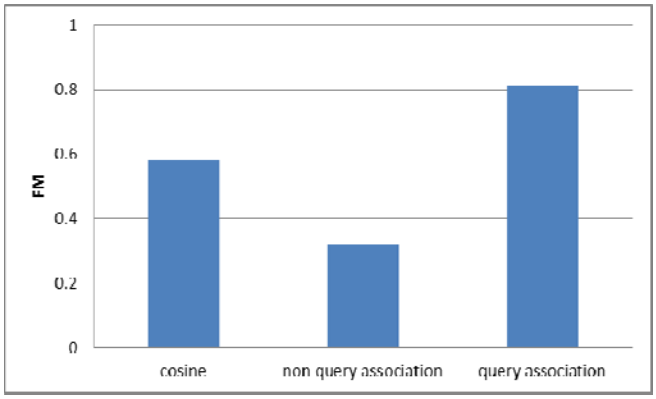

Figure 3. Fowlkes - Mallows (FM) of three clustering methods

The results from Fig. 2 and Fig. 3 show that the clustering method with query association are better than the method without query association and the cosine similarity method. The result of clustering method of query association is close to the ideal result, the reason is that this clustering method increased the similarity between different query words with the same theme of the session, thus improve the degree of similarity between users of the same interests. The result of method without query association is poor, because it can only be calculated with the same query session, and dissever the relationship of different queries about the same theme of the session, then it can't clustering the users with same interests but used different query words. The method of cosine similarity is better than the method without query association, but this kind of users-web access matrix lack of global information, and unable to effectively discover and mining query relevance between words, then the method still has a gap with the massive log mining, so that its performance can not exceed the method of query association.

V. CONCLUSION

Cluster analysis has a high value of the mining of user logs. In this paper, we analysis and research the user clustering method based on the user log mining. First we introduced existing clustering method, and then introduces the User-browsing path clustering algorithm in the application of user clustering, and studied the method of similarity involved in clustering algorithm, and propose a user similarity method with query association combined with user-browsing path algorithm to clustering the user data based on the user's session of Marine information. We analyzed on the experimental data generated by simulation and verify the effectiveness of the proposed method.

ACKNOWLEDGMENT

This work is supported by Shandong science and technology project (2013GHY11519) and other projects in Qingdao(13-CX-2) and strategic emerging industry in Qingdao development project（13-4-1-45-hy）

REFERENCES

[1] Z. Zhang,O. Nasraoui. "Mining Search Engine Query Logs for Query Recommendation," Proceedings of the 15th International Conference on World Wide Web, May 23-26,2006.

[2] Kunpeng Z, Xiaolong W, Yuanchao L. A new query expansion method based on query logs mining[J]. International Journal on Asian Language Processing, 2009, 19: 1-12..

[3] Maulik U, Mukhopadhyay A. Simulated annealing based automatic fuzzy clustering combined with ANN classification for analyzing microarray data[J]. Computers & Operations Research, 2010, 37(8): 1369-1380.

[4] Jain A K. Data clustering: 50 years beyond K-means[J]. Pattern recognition letters, 2010, 31(8): 651-666.

[5] Kao Y T, Zahara E, Kao I W. A hybridized approach to data clustering [J]. Expert Systems with Applications, 2008, 34(3): 1754-1762.

[6] Zhang C, Ouyang D, Ning J. An artificial bee colony approach for clustering [J]. Expert Systems with Applications, 2010, 37(7): 4761-4767.

[7] Wan M, Jönsson A, Wang C, et al. Web user clustering and Web prefetching using Random Indexing with weight functions [J]. Knowledge and information systems, 2012, 33(1): 89-115.

[8] S. Wang, W. Gao. Path Clustering: Discovering the Knowledge in the Web Site. Computer Research & Development, 2001, 38(4): 482-486

[9] Gong S. A collaborative filtering recommendation algorithm based on user clustering and item clustering[J]. Journal of Software, 2010, 5(7): 745-752.

[10] R. Cooley, B. Mobasher and J. Srivastava. "Web mining: Information and pattern discovery on the World Wide Web," in Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, 1997. pp. 055S-567.