International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org Volume 2 Issue 4 ‖ April. 2013 ‖ PP.20-22

An intelligent user groups Identification based on hierarchical clustering

Zhehuang Huang

School of Mathematics Sciences, Huaqiao University,362021, China

ABSTRACT: *User groups identification is an important task in intelligent personalized information service. In this paper we proposed an intelligent information service model based on hierarchical clustering algorithms. There are two main works in this article: firstly, a vector models which can represent users' interests, preferences and emotional information is introduced. Secondly, a user groups clustering algorithm based on TF/IDF is proposed. The proposed algorithm provides a good practicability and a promising future for pattern identification*

Keywords - *Vector modes, TF-IDF, Hierarchical clustering, User groups.*

I. INTRODUCTION

In recent years, with the popularity of internet, an extensive range of information resources and services is brought to users everywhere. But there have been some problem such as rising demand, massive data growth, and resource heterogeneous distribution. At present, the existing resource service model is simply assigned the resources to users which can not effectively meet the practical needs of users. As the basis and core of personalized information services, the quality of the user model is directly related to the quality of the intelligent personalized service. Only when the information of users' interests, preferences, and access mode are understood better by system, it is possible to assign the information to users based on the characteristics of the user, and achieve the desired service[1,2] .

User modeling method should be automatically constructed according to the user's browsing content and browsing behavior. User service system provides an abstract view of the characteristics of users, thus allowing the system to better understand the user's interests, preferences and emotion. Domestic and foreign scholars have launched a number of user modeling research [3, 4].Another problem of intelligence user server model is how to classifier and indentify the user groups. It can offer better service only when the user groups can be effectively identified. Clustering is a discovery process that groups a set of data such that the similarity is maximized and the cluster similarity is minimized.

Take full account of user's preferences, interest and emotional information, a new user-personalized information service model based on hierarchical clustering is proposed in this paper which can provide an efficient, consistent and personalized unified model for user knowledge resource sharing.

This paper is organized as follows. Firstly, a vector models which can represent users' interests, preferences and emotional is introduced. Secondly, an intelligent information service model based on hierarchical clustering algorithms is presented. Finally some analysis and conclusions are given on the systems.

II. TF/IDF ALGORITHM

The user information can be viewed as a vector, so we can better describe the relationship between user groups. The user information space is show as figure 1.

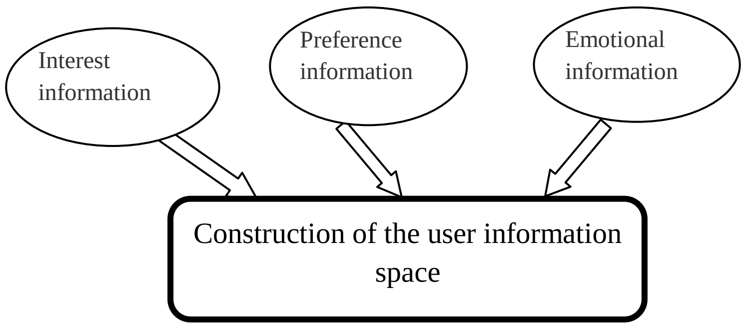

Fig.1. User information

TF-IDF (term frequency-inverse document frequency) is a commonly used weighting technique for information retrieval and text mining. TF-IDF[5,6] is a statistical method used to assess the degree of importance of a word for one of the documents in a set of files or a corpus. The process of constructing the TF-IDF vector is described below.

Supposed the weight vector for document d is $v_d = [w_{1d}, w_{2d} \cdots w_{Nd}]$, Then $w_{td} = \mathrm{TF}_{td}.\mathrm{IDF}_{td}$.

Using the cosine the similarity between document d_j and query q can be calculated as:

$$sim\left(d_j, q\right) = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \sqrt{\sum_{i=1}^{N} w_{i,q}^2}} \cdots\cdots (1).$$

III. HIERARCHICAL CLUSTERING ALGORITHM

Cluster analysis or clustering is the task of grouping a set of objects in the same group. At present, the main research focus on k-means, agglomerative clustering [7], and Information Bottleneck [8] and so on. Hierarchical clustering method set level of decomposition for a given data until certain conditions are met. It can be divided into two methods: cohesion, splitting. clusters should be combined, or where a cluster should be split, a measure of dissimilarity between sets of observations is required. The hierarchical clustering algorithm is shown as Algorithm 1.

(1) Initialize number of user groups n 、 number of clusters m

(2) clusters $C_i(w_{i1}, w_{i2} \cdots), i = 1, 2 \cdots m$

(3)Set $k = n$

(4)Set $k = k - 1$

(5)Find the nearest clusters C_i and C_j , Merge C_i and C_j

(6)If $k > m$, goto step 3, otherwise goto step 7

(7) Return m clusters

Algorithm 1. Hierarchical clustering algorithm

IV. THE USER GROUP HIERARCHICAL CLUSTERING ALGORITHM BASED ON TF-IDF

We proposed a new group clustering algorithm based on user information vector. How to create a measure of similarity between user groups is a core problem in group clustering. We use similarity calculation method based on TF-IDF algorithm

Hierarchical clustering combine data objects in different levels according to certain rules clustered into classes of different sizes and improve the clustering effect in the decomposition or consolidation process.

The users groups clustering algorithm based on TF/IDF are shown as Algorithm 2.

1) First, the user information vector pretreatment;
2) Use TF-IDF method analysis the semantic similarity between the lexical items, combined with the CS (cosine similarity) method to calculate the degree of similarity between two vectors;
3) User groups hierarchical clustering on the training set;
4) Note user groups.

Algorithm 2.Users group clustering algorithm
The user groups clustering algorithm based on TF/IDF are shown as figure 2.

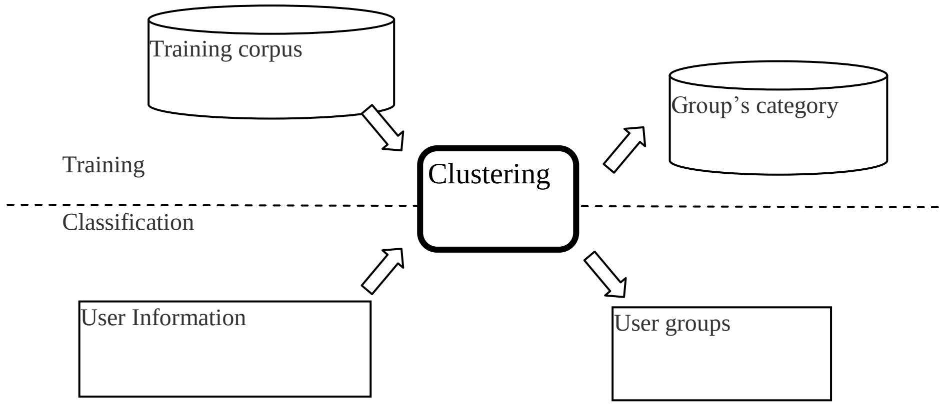

Fig.2. **Overall logical structure**

V. CONCLUSION

In this paper, we proposed a user groups clustering algorithm based on TF/IDF which provides a good practicability and a promising future for pattern Classification. We will do further investigation to inspect the inherent similarity among pattern, and apply it to other areas.

ACKNOWLEDGEMENTS

This work was supported by the science and technology project of Quanzhou (Grant No.2012Z91).

REFERENCES

[1]. Navigli R.,Velard P.,Gangemi A, Ontology learning and its application to automated terminology translation, IEEE Intelligent Systems, 18(1),2003,18(1),22-31.

[2]. Vijayan Sugumaran,Veda C.Storey,Ontologies for conceptual modeling:their creation,use,and management.Data & Knowledge Engineering, 42(3), 2002,251-271.

[3]. Kerschberg,L.,Kim,W.,Scime,A, A Semantic Taxonomy_Based Personalizable Meta-Search Agent, Proceedings of 2th International Conference on Web Information System Engeering.Kyoto,Jappan,2001,53-62.

[4]. Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll, Finding Predominant Senses in Untagged Text, In: Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, 2004, 280-287.

[5]. HUANG Cheng Hui, YIN Jian, HoU Fang. A Text Similarity Measurement Combining Word Semantic Information with TF/IDF Method. Chinese journal of computers,34(5),2011,856-864.

[6]. Ramiz M A． A new sentence similarity measure and sentence based extractive technique for automatic text summarization． Expert Systems with Applications． 36(4),2009,7764-7772.

[7]. F. Murtagh, Expected-time complexity results for hierarchic clustering algorithms which use cluster centres. Information Processing Letters, 16, 1983, 237-241.

[8]. Ying Zhao, and George Karypis, Hierarchical Clustering Algorithms for Document Datasets, Data Mining and Knowledge Discovery, 10, 2005,141–168.