A SEQUENTIAL FREQUENT PATTERN MINING FRAMEWORK FOR PERSONALIZED XML RETRIEVAL

Ms. P. Kavitha, Research Scholar, Department of Computer Science, Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore,Tamilnadu,India.

Abstract : With the huge development of internet, the information retrieval became tough and unreliable. Users interest and need is differs at every time. In order to improve the searching experience, several personalized search techniques are proposed. Using the information about the user, their history and query behavior the results will be reproduced. This kind of query reproduction is known as personalized search technique. Nowadays the information on the server is in a semi structured way. XML (eXtended Markup Language) search helps to improve the retrieval effectiveness in the semi structured data. The system proposed a novel personalized search in the XML framework. The system contributes a highly effective XML based personalized technique. This consists of Query reformulation, query expansion, re-ranking and content personalization techniques. The system introduces a new scheme for personalized search of XML documents. The system identifies the weighted terms by knowing the user interests in the search phase. The system performs the index sequential access method (ISAM) along with the improved personalization parameter and fusion re-ranking for better data retrieval.

Keywords: XML, personalization, Divisive Clustering, Pattern mining, PXR

1. INTRODUCTION

Information retrieval is the process of analysis, organization, storage, searching, and retrieval of information form web. It also described as the task of identifying documents in a collection on the basis of properties described to the documents by the user requesting the retrieval. Combination of IR and database will be valuable for the development of probabilistic models for integrity unstructured, semi-structured and structured data, for the design of effective distributed, heterogeneous information systems.

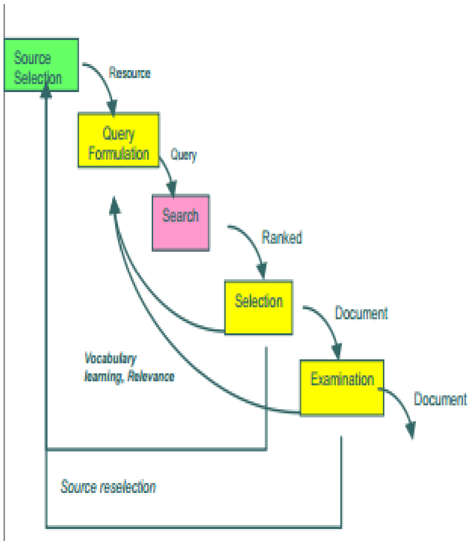


Figure 1.1. IR cycle

In above figure, source selection is first step in IR cycle. Appropriate resource is selected for all valuable data. Then query will be formed according to user needs in query formulation. Query will be processed for searching desired results, rank list will be prepared according to search results and appropriate documents are selected. If selection is not according to user needs, selection criteria goes back to source selection and query formulation. Examination process checks the documents for validation, if its validation is proved then documents will be the result otherwise output of examination goes back to source reselection and query formulation.

Web information extraction is the problem of extracting target information items from Web pages. It includes three ways to extract data which includes

- Manual Approach
- Supervised Learning Approach
- Unsupervised Learning Approach

Manual approach is done by human programmer by observing a web page and its source code. Programmer finds some patterns by writing a program to extract the data. But this approach is not scalable to large number of websites. Supervised Learning approach (Wrapper Induction) which is a semi-automatic extraction method that involves extraction rules to find patterns and to extract data. Extraction rules are learned from manually labeled pages or data records. This extracts target data items from other similarly formatted pages. Unsupervised Learning approach (Automatic Data Extraction) that finds patterns from multiple web pages without human programming. This approach automatically finds patterns or grammar from given a single or multiple web pages for data extraction. This reduces manual labeling effort.

There are two general problems: extracting information from natural language text and extracting structured data from Web pages. Automatic Data Extraction is used in this paper to extract the data from web pages.

2. RELATED WORK:

The chapter discusses about the related work of the proposed personalized XML search system. Initially this provides the details about the different personalization techniques. The system focuses on XML personalization, query expansion and re-ranking so the system- discusses about the above with their related works.

Initially the system defines the concepts about the user profiles and its methods. Next, the system focuses on XML model;

User profiles could be built by combining users' navigation paths with other data features, such as page viewing time, hyperlink structure, and page content [16]. What makes the discovered knowledge interesting had been addressed by several works. The more accurately this information represents the user, the better the retrieved results for this user.

Mining typical user profiles [2] and URL associations from the huge amount of access logs is an important component of Web personalization. In this paper this defines the notion of a user session as being a temporally compact sequence of Web accesses by a user. This also defines a dissimilarity measure between two Web sessions that captures the organization of a Web site. To cluster the user sessions based on the pair wise dissimilarities, this introduce the Relational fuzzy c-maximal density estimator (RFC-MDE) algorithm. RFC-MDE is robust and can deal with outliers that are typical in this application.

In another work [3] analysis of contextual information in search engine query logs enhances the understanding of b users' search patterns. Obtaining contextual information on b search engine logs is a difficult task, since users submit few numbers of queries, and search multiple topics. Identification of topic changes within a search session is an important branch of search engine user behavior analysis. The purpose of this study is to investigate the properties of a specific topic identification methodology in detail, and to test its validity. The topic identification algorithm's performance becomes doubtful in various cases.

Query logs record [4] the queries and the actions of the users of search engines, and as such they contain valuable information about the interests, the preferences, and the behavior of the users, as ll as their implicit feedback to search engine results. Mining the all of information available in the query logs has many important applications including query-log analysis, user profiling and personalization, advertising, query recommendation, and more. In this paper the *query-flow graph*, a graph representation of the interesting knowledge about latent querying behavior. The query-flow graph is an outcome of query-log mining and, at the same time, a useful tool for it.

In this [7] paper they present the theoretical developments necessary to extend the existing Context-based Influence Diagram Model for Structured Documents (CID model), in order to improve its retrieval performance and expressiveness. Firstly, they make it more flexible and general by removing the original restrictions on the type of structured documents that CID represents. This extension requires the design of a new algorithm to compute the posterior probabilities of relevance. Another contribution is related to the evaluation of the influence diagram.

3. PROPOSED WORK

3.1. Divisive Clustering Method

The proposed system uses the data clustering technique for further process. The system uses the Hierarchical clustering technique. Specifically divisive technique has been used in web usage mining process. The following are the general introduction to the hierarchical clustering and its types.

3.1.1 Hierarchical Clustering

The proposed system uses the hierarchical clustering. The Cluster Analysis which is also called data segmentation. This has a variety of goals in the system. All relate to grouping or segmenting a collection of objects into subsets or clusters such that those within each cluster are more closely related to one another than objects assigned to different clusters. There are two basic approaches to generating a hierarchical clustering:

Agglomerative: Start with the points as individual clusters and, at each step, merge the most similar or closest pair of clusters. This requires a definition of cluster similarity or distance.

Divisive: Start with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain. In this case, we need to decide, at each step, which cluster to split and how to perform the split

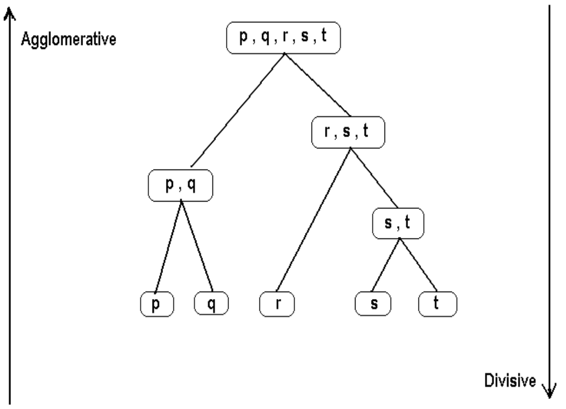
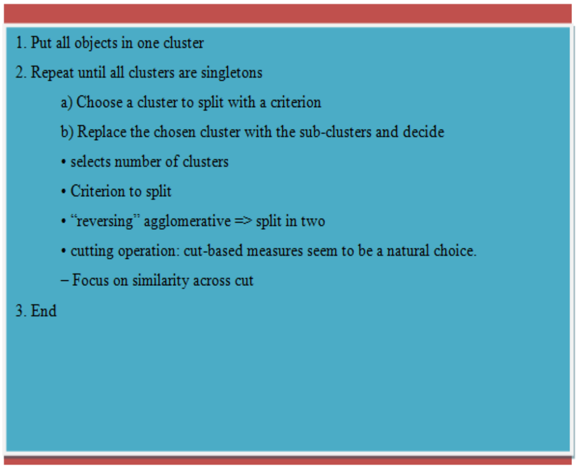


Figure 3.1 Hierarchical Clustering

3.1.2 Steps included in divisive clustering



1. Put all objects in one cluster
2. Repeat until all clusters are singletons
   a) Choose a cluster to split with a criterion
   b) Replace the chosen cluster with the sub-clusters and decide
   • selects number of clusters
   • Criterion to split
   • "reversing" agglomerative => split in two
   • cutting operation: cut-based measures seem to be a natural choice.
   – Focus on similarity across cut
3. End

This is more perfect and fast because the weblog datasets are fixed. So they run much faster than Hierarchal Agglomerative Clustering algorithms, which are at least quadratic. There is evidence that divisive algorithms produce more accurate hierarchies than bottom-up algorithms in some circumstances.

3.2. Pattern discoveries and Analysis

The discovery of user access patterns from the user web history is the main purpose of the proposed system. In the proposed system pattern discovery is performed only after the preprocessing of the weblogs, which contains cleaning the data and after the identification of user transactions and sessions from the access logs. The tools used for this process use techniques based on divisive and data mining algorithms. In the proposed system the patterns are considered as the user query and access log. The system initially collects the query from the user and retrieves data from the XML document based on the user interest. Several existing system used hits, clicked URL and explicit feedbacks are considered for data retrieval and ranking. But the proposed system utilizes the personalized XML retrieval based on the additional parameter which is the session ranking, this will added with the above parameter for effective re ranking.

3.3. PXR (Personal XML Retrieval)

Several existing systems have been proposed an explicit user feedback session clustering which is constructed from user click through logs. From the user click stream the need and interest have been identified. All the existing approaches only concentrates the click count rather than measuring "time count". Identifying user search goals based on the click stream doesn't provide exact interest and need of the user. The proposed system provides the following solution for the existing problems,

- Solution against personalized XML Retrieval problems.
- Effective user search interest detection.

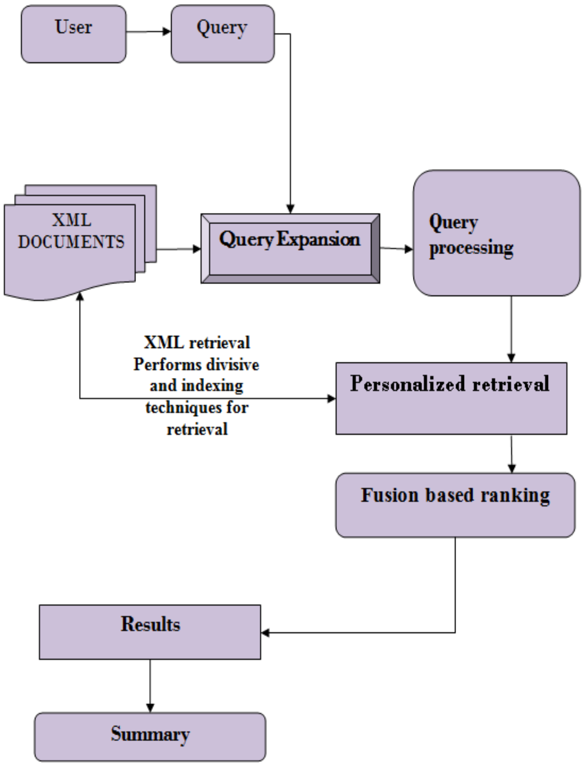
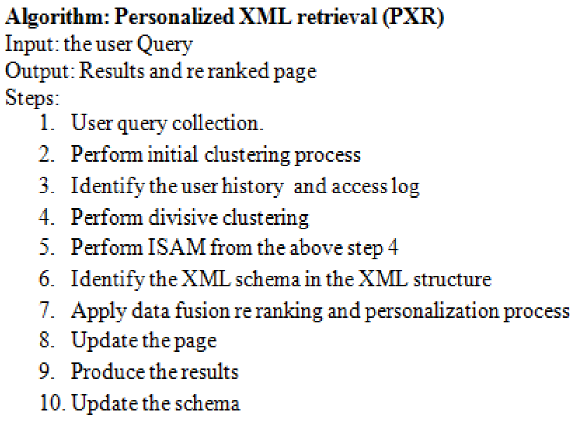


Figure 3.2. PXR architecture

3.3.1 PXR Algorithm

In the proposed method each query of the users will collected and grouped by the relevance with the collection of queries by the same user that are relevant to each other around a common information need, Here the queries are grouped and dynamically organized with updating process. As the request of users the data will be extracted and provides the most relevant links based on the sequence of search. This may sometimes creates a new issue when there no more appropriate queries in the cluster. So the proposed PXR algorithm helps to deal the new queries, and new query group creation problem, that may be solved by creating dynamic self evolved clusters.

The following algorithm represents the overall steps involved in the proposed system.



Algorithm: Personalized XML retrieval (PXR)

```
Input: the user Query
Output: Results and re ranked page
Steps:
    1. User query collection.
    2. Perform initial clustering process
    3. Identify the user history and access log
    4. Perform divisive clustering
    5. Perform ISAM from the above step 4
    6. Identify the XML schema in the XML structure
    7. Apply data fusion re ranking and personalization process
    8. Update the page
    9. Produce the results
    10. Update the schema
```

Therefore, this attempts to design a recommender system that achieves XML Retrieval by automatically estimating user's interest and their needs.

3.4. Sequential Pattern Mining

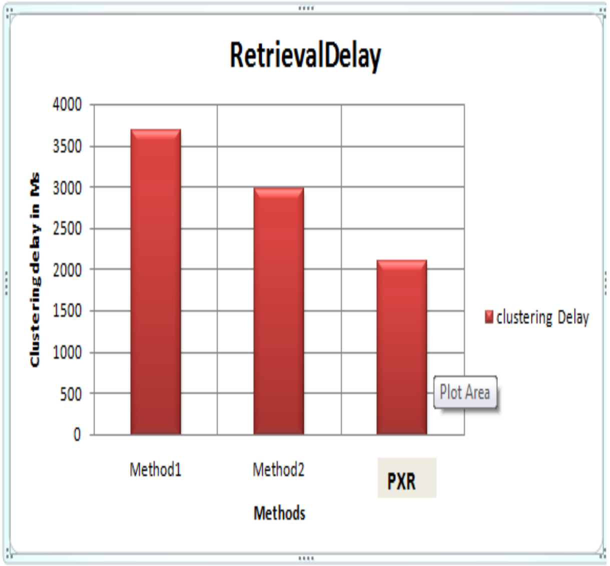
In addition to improving the efficiency of the mining process, another important direction of research is to find other types of pattern in time-related databases. Important research in this category includes finding frequent episodes in event sequences; finding frequent traversal patterns in a web log; finding cyclic patterns in a time-stamped transaction database.

4. PERFORMANCE RESULTS AND COMPARATIVE STUDY

To evaluate the performance of the proposed schemes, execution time and storage are the main measurement of performance evaluation. Without loss of generality, this defines processing delay and Retrieval Delay for deployed clustering. Processing delay indicates the execution time for clustering to produce frequent items and corresponding interest before page load.

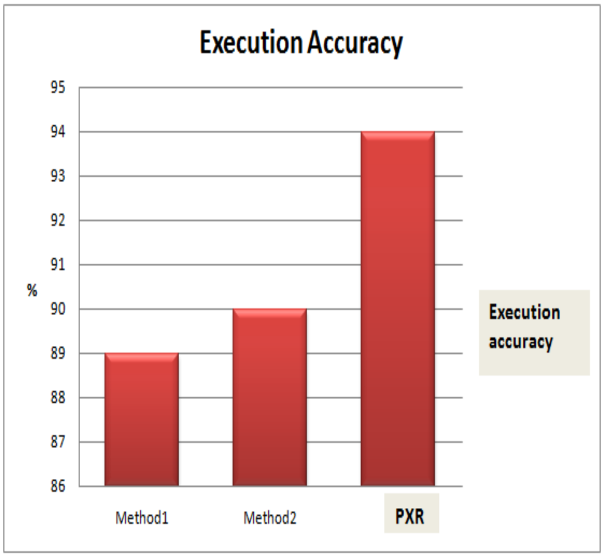
Goal detection delay is also evaluated by measuring time spent on processing time on clustering frequent items and interest in the proposed schemes. Another criterion is cost evaluation. Cost evaluation involves storage and computation aspects. The performance of this proposed work PXR using session clustering and event collection Scheme is compared with two existing approaches method1 and method2.

Performance comparison of proposed PXR using event collection with existing approaches based on Retrieval Delay



From the chart it shows the performance measure based on the Retrieval Delay and the proposed approach PXR took less time while comparing the other methods and the worst time complexity is method1.

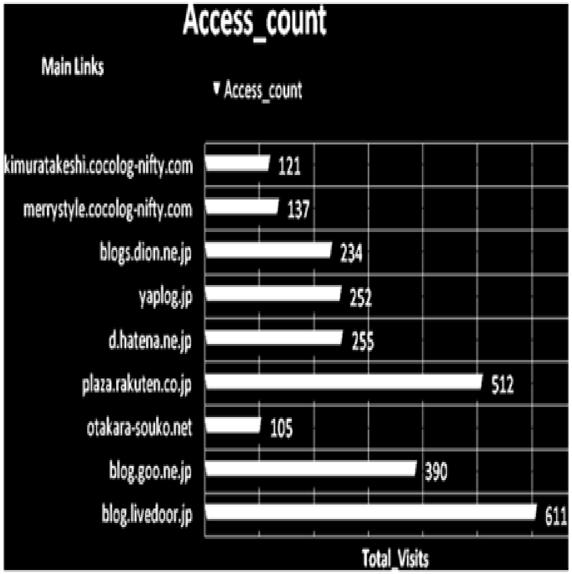
Performance comparison of proposed PXR using event collection with existing approaches based on Execution accuracy



From the chart it shows the performance measure based on the accuracy of detected cluster and the proposed approach PXR took less time while comparing the other methods and the worst based on the accuracy is method1.

Results from Divisive:

This is a "top down" approach. All explanation start in one cluster and splits are performed recursively as one move down the hierarchy. Here the datasets are clustered using divisive analysis; the clustered datasets are split into a single cluster.



Using the methodology and metrics presented above, performed experiments to evaluate the three cluster methods. The results presented in this section provide a detailed analysis and benefits of the proposed approach which has been created to personalizing Web directories.

The results define and show the proposed system is working well in the high dimensional and huge dataset. Even the proposed system obtained good performance by all methods the use of session based clustering and PXR with divisive for the personalization of Web directories appears to be the most promising. It helps identifying latent information in the users' choices and derives high-quality community directories that provide significant benefits to their users. The results presented here provide an initial measure of the benefits that this can obtain by personalizing the user web directories to the needs and interests of user communities.

5. CONCLUSION AND FUTURE WORK

The system proposed a new technique for personalized XML retrieval. To improve the data searching and retrieving process, the system provides an effective query expansion method, which deals with the personalized web mining techniques. The proposed personalized XML retrieval system studies how to select the configuration parameters which includes the number of terms from the profile to use and the normalization factor of their weights depending on the characteristics of the query, in order to obtain better personalized results.

Future work

The proposed query expansion and personalized XML using divisive clustering improved the effectiveness in the personalization. In future this may expanded with some other clustering technique in order to verify the effectiveness. And the XML retrieval will be implemented using ontology and UN structured document domains.

REFERENCES

[1] O. Nasraoui, C. Rojas, and C. Cardona, "A Framework for Mining Evolving Trends in Web Data Streams Using Dynamic Learning and Retrospective Validation", Computer Networks, special issue on Web dynamics, vol. 50, no. 14, Oct. (2006).

[2] Pengyi Zhang, Jiannan Xia, Ruiji Li, "Personalized Multimedia Information Retrieval based on User Profile Mining",Journal of Networks, Vol 8, No 10 (2013).

[3] Ozmutlu, H. Cenk, and Fatih Çavdur, "Application of automatic topic identification on excites web search engine data logs", Information Processing & Management 41.5: 1243-1262, (2005).

[4] Boldi, Paolo, et al, "Query suggestions using query-flow graphs", Proceedings of the 2009 workshop on Web Search Click Data. ACM, (2009).

[5] Feng, Juan, Hemant K. Bhargava, and David M. Pennock, "Implementing sponsored search in web search engines: Computational evaluation of alternative mechanisms", INFORMS Journal on Computing 19.1 (2007): 137-148.

[6] Chernishev, George, "Personalization of XML Text Search via Search Histories", SYRCoDIS. (2008).

[7] De Campos, Luis M., Juan M. Fernández-Luna, and Juan F. Huete, "Improving the Context-based Influence Diagram Model for structured document retrieval: removing topological restrictions and adding new evaluation methods", Advances in Information Retrieval, Springer Berlin Heidelberg, (2005). 215-229.

[8] Sieg, Ahu, Bamshad Mobasher, and Robin D. Burke, "Learning Ontology-Based User Profiles: A Semantic Approach to Personalized Web Search", IEEE Intelligent Informatics Bulletin 8.1 (2007): 7-18.

[9] Schenkel, Ralf, and Martin Theobald, "Structural feedback for keyword-based XML retrieval", Advances in Information Retrieval, Springer Berlin Heidelberg, (2006).326-337.

[10] Belkin, Nicholas J, "Some (what) grand challenges for information retrieval", ACM SIGIR Forum. Vol. 42. No. 1. ACM, (2008).

[11] Haveliwala, Taher H, "Topic-sensitive pagerank", Proceedings of the 11th international conference on World Wide Web. ACM, (2002).

[12] Bertier, Marin, et al, "Toward personalized query expansion", Proceedings of the Second ACM EuroSys Workshop on Social Network Systems. ACM, (2009).

[13] Eirinaki M., Vazirgiannis M, "Web mining for web personalization", ACM Transactions On Internet Technology (TOIT), 3(1), 1-27 (2003).

[14] Agrawal R. and Srikant R, "Privacy preserving data mining, In Proc. of the ACM SIGMOD Conference on Management of Data, Dallas, Texas, 439-450 (2000)..

[15] Berendt B., Bamshad M, Spiliopoulou M., and Wiltshire J, "Measuring the accuracy of sessionizers for web usage analysis", In Workshop on Web Mining, at the First SIAM International Conference on Data Mining, 7-14 (2001)..

[16] Mobasher, B., "Web Usage Mining and Personalization", in Practical Handbook of Internet Computing, M.P. Singh, Editor, CRC Press. p. 15.1-37 (2004),.

[17] Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, and Eduardo Vicente-López, "Using Personalization to Improve XML Retrieval", IEEE Transactions On Knowledge and Data Engineering, Vol. 26, No. 5, May(2014).

[18] Iyad Batal, Alexandros Labrinidis.QuickStack: "A Fast Algorithm for XML Query Matching", Department of Computer Science University of Pittsburgh, June 6, (2008).

[19] D.Bujji Babu, Dr. R.Siva Rama Prasad, M.Santhosh, "XML Twig Pattern Matching Algorithms and Query Processing ", International Journal of Engineering Research & Technology Vol.1 - Issue 3, May(2012).

[20] Mylonas Ph., Vallet D., Castells P., "Personalized information retrieval based on context and ontological knowledge", Knowledge Engineering Review, 23(1) pp. 73-100, (2008).

[21] Oussalah M., Khan S., Nefti S., "Personalized information retrieval system in the framework of fuzzy logic", Expert Systems With Applications, 35(1-2) pp. 423-433, (2008).

[22] Yoo Donghee, "Hybrid query processing for personalized information retrieval on the Semantic Web", Knowledge-based Systems, 27 pp. 211-218, (2012).

[23] Pereira Celia da Costa, Dragoni Mauro Pasi Gabriella, "Multidimensional relevance pp. Prioritized aggregation in a personalized Information Retrieval setting", Information Processing & Management, 48(2) pp. 340-357, (2012).

[24] Carmel David, Zwerdling Naama, Guy Ido, Ofek-Koifman Shila, Har'El Nadav, Ronen Inbal, Uziel Erel, Yogev Sivan, Chernov Sergey, "Personalized social search based on the user's social network", In Proceedings of the 18th ACM conference on Information and knowledge management, pp.1227-1236,(2009).

[25] Bo Ning, Zhang Ji, "Research on web information retrieval based on vector space model", Journal of Networks, 8(3) pp. 688-695, (2013).

[26] Godoy D, Amandi A, "Modeling user interests by conceptual clustering", Information Systems, 31(4-5) pp. 247-265, (2006).

[27] Chi Ming-Chieh, Yeh Chia-Hung, Chen Mei-Juan, "Modeling user multiple interests by an improved GCS approach", Expert Systems With Applications, 29(4) pp. 757-767,(2005).