PERSONALIZED QUERY RESULTS IDENTIFICATION USING AGGLOMERATIVE CLUSTERING ALGORITHM

Chandrababu alavala[1*], Ch. Srinivasarao[2*], Vennakula l s Saikumar[3*]

1. *M.Tech (CSE) Student, Dept of CSE, Pydah College of Engg & Tech, Dist: Vishakapatnam, AP, India.*
2. *Assoc Professor, Dept of CSE, Pydah College of Engg & Tech, Vishakapatnam, AP, India.*
3. *M.Tech (CSE) Student, Dept of CSE, Pydah College of Engg & Tech, Vishakapatnam, AP, India.*

Keywords:
Personalized Query, Clustering Methodology, Agglomerative Approach, Clickthrough

Abstract: *The age of modern technology gives more user synthetic easiness to the user. Query Personalization is the process of dynamically enhancing a query with related user preferences stored in a user profile with the aim of providing personalized answers. These days' people likely to be short and ambiguous to search the data or information in the web, so as it's difficult for search engine to search which xml data and gives to rise some unrelated data. In this paper, we give stress on the identification and then followed but its solution by using Agglomerative clustering approach. Clustering depends critically on density and distance (similarity), but these concepts become increasingly more difficult to define as dimensionality increases. This approach handles many problems that traditionally plague clustering algorithms, e.g., finding clusters in the presence of noise and outliers and finding clusters in data that has clusters of different shapes, sizes, and density.*

1. INTRODUCTION

With the help of search engines, Web queries are becoming a major bridge between Web users and online services provided by search engines such as advertisement and Web page search. Query classification (QC) is a task that aims to classify Web queries into topical categories. Since queries are usually short in length and ambiguous, the same query may need to be classified to different categories according to different people's perspective. In this paper, we propose Personalized Query Classification (PQC) task and develop an algorithm based on user preferences learning. Users' preferences that are
hidden in clickthrough logs are quite helpful for search engines to improve on their understanding of users' queries. We propose to connect query classification with preferences learning from click through log for PQC. To tackle the sparseness problem in user preferences learning, we also propose a collaborative ranking model to leverage similar users' information. Experiments on a real world click through log show that our proposed PQC algorithm can gain significant improvement compared with general QC.

Cluster analysis tries to divide a set of data points into useful or meaningful groups, and has long been used in a wide variety of fields: psychology and other social sciences, biology, statistics, pattern recognition, information retrieval, machine learning, and data mining. Cluster analysis is a challenging task and there are a number of well-known issues associated with it, e.g., finding clusters in data where there are clusters of different shapes, sizes, and density or where the data has lots of noise and outliers. It is an unsupervised learning technique that is widely used in Artificial Intelligence, Data mining etc. It aims to discover patterns taking into account the entire data. There are no pre-defined conditional and decision variables. Members within a cluster are more similar or related to each other and different from members of other clusters. Clustering is crucial to our work, as it helps to

*** *Chandrababu alavala***
M.Tech (CSE) Student, Dept of CSE, Aditya Engg College, Surampalem, Dist: E.Godavari, AP, India

identify related or similar web services. There are many different approaches to clustering. In the present work, we use agglomerative or bottom-up Hierarchical Clustering method. In this method, initially each web service is treated as belonging to a cluster. Then we use similarity matrix of web services in the training dataset to determine the nearest neighbors. Nearest clusters are then merged into one cluster. This process is repeated and in the end all the web services merge to a single cluster.

2. RELATED WORK

Our method can be applied to a wide range of applications including personalized search and online advertising. While "better" notions of distance and density are key ideas in our clustering algorithm, we will also employ some additional concepts which were embodied in three recently proposed clustering algorithms, i.e., CURE, Chameleon and DBSCAN. Although, the approaches of these algorithms do not extend easily to high dimensional data, they algorithms outperform traditional clustering algorithms on low dimensional data, and have useful ideas to offer. In particular, DBSCAN and CURE have the idea of "representative" or "core" points, and, although our definition is somewhat different from both, growing clusters from representative points is a key part of our approach. Chameleon relies on a graph based approach and the notion that only some of the links between points are useful for forming clustering; we also take a graph viewpoint and eliminate weak links. All three approaches emphasize the importance of dealing with noise and outliers in an effective manner, and noise elimination is another key step in our algorithm.

In this related part, we likely stress on the query part and the relevant problematic words to be searched on the web engine by the help google middleware. Query clustering techniques have been developed in diversified ways. The very first query clustering technique comes from information retrieval studies. Similarity between queries was measured based on overlapping keywords or phrases in the queries. Each query is represented as a keyword 'vector'. Similarity functions such as cosine similarity or similarity were used to measure the distance between two queries. One major limitation of the approach is that common keywords also exist in unrelated queries.

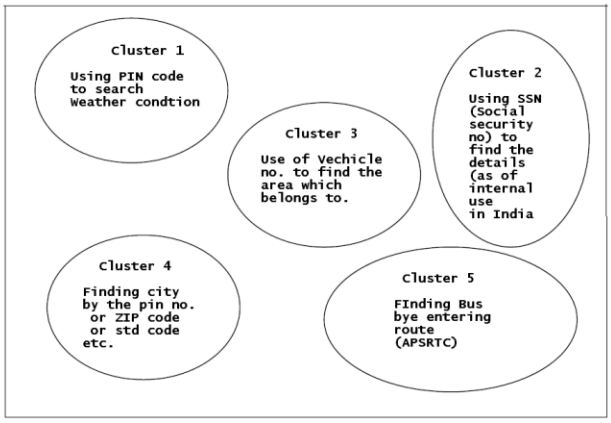

Fig. 2.1 showing the various clustered Approach

We came up with a novel idea of representing the clusters by the most similar operations of web services in that cluster. If there is only one web service in a cluster, then we select an operation that is very dissimilar to operations in other clusters. It lists the characteristic operations for clusters represented in Fig. 2.1.

In another approach, direct similarity is to define the similarity between a pair of points in terms of their shared nearest neighbors. That is, the similarity between two points is confirmed by their common (shared) near neighbors. If point A is close to point B and if they are both close to a set of points C then we can say that A and B are close with greater confidence since their similarity is "confirmed" by the points in set C. This idea of shared nearest neighbor was first introduced by Jarvis and Patrick.

3. METHODS

Technology Changes with human requirement likewise these people are worried about the searched word on the web. As of the search engine 'google' is concerned whenever we enter any word to get the adequate or relevant information , the search engine works based on 'soundex' keyword of oracle data type**.**

A method for determining the mutual nearest neighbors (MNN) and mutual neighborhood value of a sample point, using the conventional nearest neighbors, is suggested. A nonparametric, hierarchical, agglomerative clustering algorithm is developed using the above concepts. The algorithm is simple, deterministic, no iterative, requires low storage and is able to discern spherical and no spherical clusters. The method is

applicable to a wide class of data of arbitrary shape, large size and high dimensionality. The algorithm can discern mutually homogenous clusters. Strong or weak patterns can be discerned by properly choosing the neighborhood width.

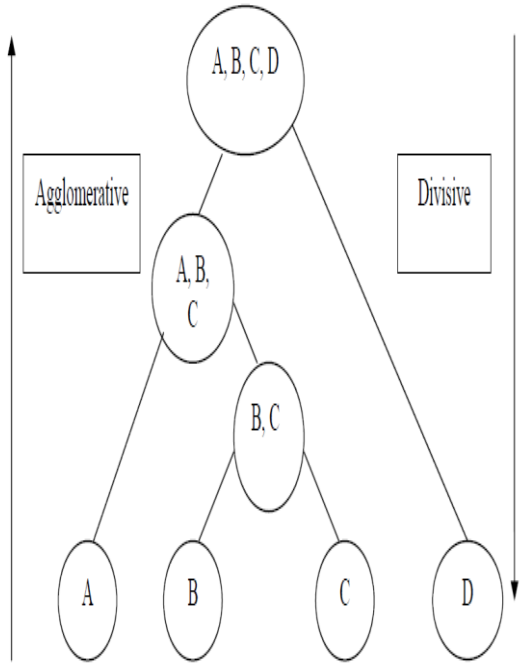

Fig. 3.1Showing Hierarchical Clustering Method

In the above figure 3.1, we have clustered the data as of spanning tree which shows both as agglomerative and divisive, in order to get the direct and indirect or related data.

ALGORITHM FOR AGGLOMERATIVE CLUSTER:

The algorithm forms clusters in a bottom-up manner, as follows:

1. Start: put each article in its own cluster.
2. From all current clusters, pick the two clusters with the smallest distance.
3. Get these two clusters with a new cluster, formed by merging the two original ones.
4. Repeat the above two steps until there is only one remaining cluster in the pool.

Thus, the agglomerative clustering algorithm will result in a binary cluster tree with single article clusters as its leaf nodes and a root node containing all the articles.

In the clustering algorithm, we use a distance measure based on log likelihood. For articles *X* and *Y*, the distance is defined as

Dist (X, Y) =L(X) +L(Y)-L (XU Y)

The log likelihood *L*(*G*) of an article or cluster *X* is given by a unigram model:

$$L(G) = \log \prod_{t \in G} P^*(t)^{c_x(t)}$$

$$= \sum_{t \in G} c^*(t) \log c_x(t) - M_x \log M_x$$

Here, P*(t) is the count and probability, respectively, of word t in cluster *X*, and Mx is the total number of words occurring in cluster *X*.

The agglomerative cluster is given as follows.

$$Dist'(X,Y) = (M_X + M_Y)K(XU\ Y) - (M_X K(X) + M_Y K(Y)$$

Where

$$K(X) = -\sum_{t \in G} P_x(t) \log P_x(t)$$

Dist=Distance

Hence, in this approach of clustering, and in the above algorithm which gives distance based approach in the bottom way cluster.
As we move and move in further of web based search, its uncountable no. of related documents are available in the web. But, there would like to create a problem of URL, in such context, we have gone through the click through based method is that the number of common clicks on URLs for different queries is limited. This is because different queries will likely retrieve very different result sets in very different ranking orders. Thus, the chance for the users to see the same results would be small, let alone clicking on them. It was reported that in a large click through data set from a commercial search engine the chance for two random queries to have a common mechanism of data and vice versa.

4. CONCLUSION

The concept of web based search engine on a personalized document and filtration become a typical task to the search engine due to the huge mass of data. As

search queries are ambiguous, we have studied effective methods for search engines to provide query suggestions

on Agglomerative related queries in order to help users formulate more effective queries to meet their diversified needs. In this paper, we have proposed a new personalized concept-based clustering technique that is able to obtain personalized query suggestions for individual users based on their conceptual profiles. The technique makes use of click through data and the agglomerative clustered data which grouped and also sub grouped based on various factors like to be dimension and neighborhood graphic approach. Both of which can be captured at the back end and as such do not add extra burden to users. An adapted agglomerative clustering algorithm is employed for finding queries that are conceptually close to one another. In this approach which keeps for further research to find AI based intelligence of the search engine.

5. REFERENCE

[1] He, Hao (2003), "What is Service-Oriented Architecture", Retrieved 5/5/2007 from http://webservices.xml.com/pub/a/ws/2003/09/30/soa.html

[2] E. Christensen, F. Curbera, G. Meredith, and S. Weerawarana (2001), "Web Services Description Language (WSDL) 1.1", W3C Recommendation, 2001, Retrieved 4/4/2007 from http://www.w3.org/TR/2001/NOTE-wsdl-20010315.

[3] L. Clement et al. (Ed.) (2004), "UDDI Version 3.0.2", Retrieved 5/5/2007 from http://uddi.org/pubs/uddiv3.0.2-20041019.htm.

[4] J.-R. Wen, J.-Y. Nie, and H. Zhang, "Query clustering using user logs," *ACM Trans. Inf. Syst.*, vol. 20, no. 1, pp. 59–81, 2002.

[5] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. ACM SIGKDD, 2000.

[6] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to Rank Using Gradient Descent," Proc. Int'l Conf. Machine learning (ICML), 2005.

[7] K.W. Church, W. Gale, P. Hanks, and D. Hindle, "Using Statistics in Lexical Analysis," Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon, Lawrence Erlbaum, 1991.

[8] Z. Dou, R. Song, and J.-R. Wen, "A Largescale Evaluation and Analysis of Personalized Search Strategies," Proc. World Wide Web (WWW) Conf., 2007.

[9] S. Gauch, J. Chaffee, and A. Pretschner, "Ontology-Based Personalized Search and Browsing," ACM Web Intelligence and Agent System, vol. 1, nos. 3/4, pp. 219-234, 2003.

[10] T. Joachims, "Optimizing Search Engines Using Clickthrough Data," Proc. ACM SIGKDD, 2002.

[11] K.W.-T. Leung, W. Ng, and D.L. Lee, "Personalized Concept- Based Clustering of Search Engine Queries," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 11, pp. 1505-1518, Nov. 2008.

[12] B. Liu, W.S. Lee, P.S. Yu, and X. Li, "Partially Supervised Classification of Text Documents," Proc. Int'l Conf. Machine Learning (ICML), 2002.

[13] F. Liu, C. Yu, and W. Meng, "Personalized Web Search by Mapping User Queries to Categories," Proc. Int'l Conf. Information and Knowledge Management (CIKM), 2002.

[14] http://research.microsoft.com/apps/pubs/default.aspx?id=102410