Clustering of Web Search Result Using Meta-Heuristic Algorithm

Chetan Mahajan
Assistant professor
Shah and Anchor Kutchhi Engineering College,
Mumbai, India

Rachit Agarwal
Information Technology
Shah and Anchor Kutchhi Engineering College,
Mumbai, India

Deepti Dighe
Information Technology
Shah and Anchor Kutchhi Engineering College,
Mumbai, India

Sayali Gawade
Information Technology
Shah and Anchor Kutchhi Engineering College,
Mumbai, India

Priyanka Pirale
Information Technology
Shah and Anchor Kutchhi Engineering College,
Mumbai, India

***Abstract*— Search engines serve millions of users per day looking for answers to their questions or for solutions to their problems. But there are some words which multiple meanings and user may have to review each and every document till he finds the desired document. The aim of our project is to reduce the time of user in reviewing document. Our tool will be usefully to those who are less friendly with web browser. People just have to type a single word which will retrieve all the documents related to the word and with proper categorization. For doing this task we will be forming clusters of documents having same meaning for example a word orange has multiple meanings like it can be a fruit, a data mining tool and a color. Technologies used to execute this idea are Java Swing and MySQL as database. K-means is used for Clustering of documents and Cuckoo Search Algorithm for selection of best nest. Nest in this scenario means collection of clusters. Islands are formed using nest to increase the efficiency of clusters. Islands means carrying out the same procedure of nesting to get accurate result. BIC ie Bayesian Information Criterion is used to calculate the scores on which the best cluster is selected. We have kept the number of Islands to be 5 and number of nests depends on the user input. Minimum Clusters is kept 2 and Maximum Clusters is calculated using the formula: √n+1 where n is the number of documents retrieved. User will select the number of nest ranging from 1 to Maximum Cluster – Minimum Cluster Key Words: K-means, Balanced Bayesian, Web search result clustering, Meta-heuristic algorithm**

INTRODUCTION

Data clustering is the process of grouping data elements in a way that makes the elements in a given group similar to each other in some aspect.

In this paper we have explained the criteria to check whether the formed cluster is accurate or not. For this we have used the concept of BIC score. Bayesian information criterian .We have also explained the formation of nests and islands.

A. Nesting

Nest are created based on user input. The number of nest can be obtained by using following formula:

Number of nests = maximum number of clusters – minimum number of cluster. [1]

The purpose of performing nesting is to get more efficient result. This is done by using cuckoo search algorithm.

BIC score that is Bayesian Information Criterion [5] is used to compare the clusters. More the score more efficient the cluster is. Nests are compared using BIC score.

Nest with highest BIC score from all 5 islands is considered to be the best cluster with less error rate.

B. Island

Island means carrying out the nesting procedure for multiple number of times.[1] This is done to obtain the most efficient nest. The number of nests in the island is determined by the user. We can create as many as islands until we get the efficient result but this creates load on the system [1]. So we should select a boundary value to limit the number of islands

C. Visualization of cluster

Cluster with highest BIC score is displayed on the user interface using Jtree.

Label is assigned to each cluster using highest term frequency.

Care is taken that the label does not match with the query inserted [3].

CONCLUSION

Clustering of web search results has been studied in the area of Information Retrieval (IR). The goal of clustering search result is to give user an idea of what the result contains. Thus web search clustering can be made more efficient by using concept of nests and islands. Also BIC score is be used to determine the best nest.

REFERENCES

[1] Anil K. Jain, Data Clustering: 50 Years beyond K-Means. Michigan State University, Michigan.

[2] Clustering of web search results based on the cuckoo search algorithm and Balanced Bayesian Information Carlos Cobos a,b, Henry Muñoz-Collazos , Richar Urbano-Muñoz a

[3] C. Carpineto, S. Osin´ ski, G. Romano, D. Weiss, A survey of Web clustering engines, ACM Comput. Surv. 41 (2009) 1–38

[4] Hamerly, G.; Elkan, C. (2002). "Alternatives to the k-means algorithm that find better clusterings" (PDF). Proceedings of the eleventh international conference on Information and knowledge management (CIKM).

[5] Celebi, M. E., Kingravi, H. A., and Vela, P. A. (2013). "A comparative study of efficient initialization methods for the k-means clustering algorithm"