American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-6, Issue-10, pp-285-286
www.ajer.org
Research Paper Open Access

Web search result clustering using Meta-heuristic Algorithm

Rachit Agarwal[1],Sayali Gawade[1],Deepti Dighe[1],Priyanka Pirale[1]
[1]*Information Technology, Shah and Anchor Kutchhi Engineering College, Mumbai, India*

ABSTRACT: *Clustering of web search result has become a wide area of research in data mining. Web search result clustering covers all the documents and helps user to review the document directly without wasting time in reading the description given below the url. This paper shows the usage of k-means clustering algorithm along with balanced bayesian to form the proper document clusters. This document clustering is tested on only one dataset that is AMBIENT. K-means algorithm will be used for document clustering and Balanced Bayesian will calculate probability which will be used to put a document into proper nest that is cluster. .The work we can do is improve the quality of clusters and compare results with other algorithm used for clustering and developing algorithm that will meet the real world challenges.*

Keywords: *K-means, Balanced Bayesian, Web search result clustering, Meta-heuristic algorithm*

Date of Submission: 11-10-2017 Date of acceptance: 27-10-2017

I. INTRODUCTION

In recent years, web result clustering has become a very interesting research area among academic and scientific communities involved in information retrieval (IR) and web search [1].This is because it is most likely that results relevant to the user are close to each other in the document space, thus tending to fall into a relatively small number of clusters [2] and thereby achieving significant reductions in search time. The searching begins with query defined by the user. A keyword inserted by the user is then searched throughout the dataset and matching results are retrieved. Pre-processing is the step in which following actions are carried out:(1)removing punctuation marks and converting all words into lowercase from the snippet;(2)removing stop words;(3)determining root word for each word in snippet using lemmatizer;(4)creating TF-IDF matrix. Once the pre-processing is completed the next step is to construct cluster based on meta-heuristic algorithm.Meta-heuristic algorithm uses k-means algorithm for clustering and balanced bayesian information criterion for evaluation of the clusters formed. Finally, in the visualization step, the system displays the results to the user in hierarchically organized folders. Each folder seeks to have a label or title that represents well the documents it contains and that is easily identified by the user.As such, the user simply scans the folders that are actually related to their specific needs [3].needs. The presentation folder tree has been adopted by various systems such as Carrot2, Yippy, SnakeT, and KeySRC, because the folder metaphor is already familiar to computer users. Other systems such as Grokker and Kart004 use a different display scheme based on graphs [1].

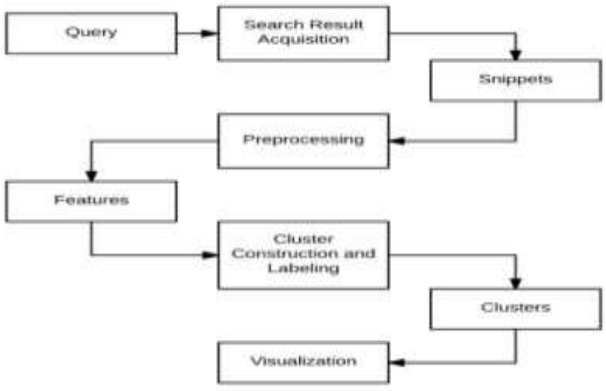

Figure 1: Components of Web Clustering Engine (adapted from [1])

II. MATERIALS AND METHODS

Materials used:

Graphical User Interface: Java Swing
Database used : MySQL.
Dataset : Collected from Ambient (http://credo.fub.it/ambient).

1. K-means Algorithm

The K-means algorithm is the simplest and most commonly used algorithm for clustering employee a sum of squared error (SSE) criterion based on

$$SSE = \sum_{j=0}^{k} \square \sum_{i=0}^{n} Pij||xi - cj||^2$$

where n is the total number of records(documents), k is the number of clusters , P_{ij} equals to 1 when the document x_i belongs to the c_j cluster, otherwise 0. The algorithm is popular because it finds the local minimum (maximum) in the search space, it is easy to implement.

ALGORITHM:

Select an Initial Partition (k centers)
Repeat
Data Assignment: Re-compute Membership
Relocation of "means": Update Centers
Until (Stop Criterion)
Return Solution

III. BALANCED BAYESIAN ALGORITHM

In Bayes' rule, the product of prior probability π (θ) and the likelihood of data given a parameter Vector $f(y|\theta)$ result in the posterior distribution where y is the data and θ are the model parameters. The denominator $m(y)$ is known as the marginal likelihood of the data and found by integrating the Likelihood overprior densities Depending on the dimensionality of θ and the complexity of $f(\cdot)$, the determination of the scaling factor $m(y)$ is not often possible and the Markov Chain Monte Carlo (MCMC) [5] approach can be used in such cases. After a burn-in period which is necessary to converge from an initial parameter vector to the stationary distribution, each iteration of the MCMC approach represents a parameter vector out of the posterior distribution.

IV. CONCLUSION

The web search result clustering can be successfully implemented using k-means and balanced Bayesian algorithm. Pre-processing step is completed with implementation and cluster labelling with nest formation will be done later so that the user is able to search the results without any extra time in reviewing the entire document.

REFERENCES

[1] C. Carpineto, S. Osin´ ski, G. Romano, D. Weiss, A survey of Web clustering engines, ACM Comput. Surv. 41 (2009) 1–38.
[2] C. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, Cambridge, England, 2008.
[3] Data Clustering: Algorithms and Applications. CRC Press; 2014.
[4] Clustering of web search results based on the cuckoo searchalgorithm and Balanced Bayesian Information Criterion.
[5] Balanced Bayesian Mechanisms∗ Claude d'Aspremont CORE, Universite catholique de Louvain, Louvain-la-Neuve, Belgium ´ Jacques Cremer ´ CNRS, IDEI and GREMAQ, University of Toulouse 1, France Louis-Andre G´ erard-Varet ´ EHESS, GREQAM, Marseille, France

Rachit Agarwal "Web search result clustering using Meta-heuristic Algorithm" American Journal of Engineering Research (AJER), vol. 6, no. 10, 2017, pp. 285-286.