Clustering of Web Search Result using Meta-heuristic Algorithm

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*— This paper is an attempt to understand the clustering using k means more effectively with actual values. Clustering is a very integral part of mining domain. This paper is going to present the usage k-means to form the clusters of documents. To experiment this we are going to use AMBIENT dataset. We are going to use Balanced Bayesian Algorithm to calculate the probability to decide in which cluster the document should be placed.**

The idea of our project is to create nests or cluster of the items searched by the user. For example, if the user searches the word ORANGE the output that appears on any web browser is either a fruit or a color or may be a data mining tool etc. It is not worth if the user spends the time visiting all the web links to get to the required result. Hence, after applying the clustering we are going to segregate the search results into the form of clusters for easy navigation. Orange as a color will be stored in a particular cluster whereas orange as a fruit will be in another cluster. This will save the user's time and will make searching more efficient. Also the comparison of the results will be much simpler to perform due to formation of clusters.

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.

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

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.[2,5]

ALGORITHM:

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

Return Solution

To obtain good results in web document clustering the algorithms must meet the following specific requirements [3]:

(1) automatically define the number of clusters that are going to be created;

(2) generate relevant clusters for the user and assign the documents to appropriate clusters;

(3) define labels or names for the clusters that are easily understood by users;

(4) handle overlapping clusters (documents can belong to more than one cluster);

(5) reduce the high dimension of document collections;

(6) handle the processing time, i.e. less than or equal to 2 s; and

(7) handle the noise that is frequently found in documents.

A. Balanced Bayesian Algorithm

In the data balancing problem the goal is to update a probability distribution, under the guiding principle that the best inference is the one which takes into account all available information and no other. This principle is operationalized by searching for a posterior distribution that is as close as possible to the prior (in an information sense) and that satisfies the accounting identities, expressed in terms of moment constraints [4].

In Bayes' rule, the product of prior probability π (θ) and the likelihood of data given a parameter. Vector f(y|θ) 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. It is found by integrating prior densities depending on the dimensionality of θ.

B. TF-IDF

The TFIDF is a abbreviation used for term frequency inverse document frequency. It is a numerical approach which intends to how important word is in that document. Mostly used in data and text mining. The value of TFIDF increases proportionally with the no. of times the word has appeared in the document. It basically shows how frequently the word has appeared in the document [2].

Term-By-Document Matrix (TDM):

The TDM matrix is the most widely-used structure for document representation in IR, and is based on the vector space model [6,31]. In this model, the documents are designed as bags of words; the document collection is represented by a matrix of D-terms by n-documents. Each document is represented by a vector of normalized frequency term (tft) by the document inverse frequency for that term, in what is known as TF-IDF value (expressed by Eq. (7)), and the cosine similarity (seeEq. (3)) is used for measuring the degree of similarity between two documents or between a document and cluster centroid.

Wt,i= freqt,i /max(freqi)*log(n/nt)

where freqt,I observed frequency of the term t in document representation in IR [2]. . In this model, the documents are designed as bags of words; the document collection is represented by a matrix of D-terms by n-documents. Each document is represented by a vector of normalized frequency term (tft) by the document inverse frequency for that term, in what is known as TF-IDF value.

CONCLUSION

Thus web search clustering can be successfully implemented using k-means and Balanced Bayesian algorithm. Tf-idf matrix can also be generated using the above discussed formulas and based on its values we can determine the document in which the word can be placed.

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"