International Journal of Engineering Works
Kambohwell Publisher Enterprises
ISSN: 2409-2770 Vol. 2, Issue 8, PP. 79-83, August 2015
www.kwpublisher.com

K-Means Clustering For Segment Web Search Results

Hasitha Indika Arumawadu[1], R. M. Kapila Tharanga Rathnayaka[2,4], S. K. Illangarathne[3]

***Abstract*—** Clustering is the power full technique for segment relevant data into different levels. This study has proposed K-means clustering method for cluster web search results for search engines. For represent documents we used vector space model and use cosine similarity method for measure similarity between user query and the search results. As an improvement of K-means clustering we used distortion curve method for identify optimal initial number of clusters.

***Keywords*—** K-means clustering, Web Search, Search Engine, Vector space model, Distortion curve

I. INTRODUCTION

Internet usage is growing day-by-day. As a result web pages, documents, images, videos are growing massively. Web search engines like Google, Yahoo, Baidu retrieve the results when user request their query (e.g. There are totally 641,000,000 matched results by query "food", searched in May of 2010 in Google) [1]. But there are lots of webpages will appear relevant to users query. So, web search engine displays only few most ranked pages as a top results. Sometimes user need to navigate more pages to get his desired result. It's depend on the ranking algorithm of the web search engine. Different search engine use different algorithm for rank search results.

Search engines return millions of pages or documents in an answer to a query. The result becomes even larger if the query is ambiguous as search engines try to retrieve documents for all the possible meaning of a query. Clustering of search result is a way to summarize this large amount of documents in form of groups where group members share similar qualities. There are many clustering engines available like Kartoo, Carrot2, Vivisimo etc. As search result clustering is being widely researched [2]. Clustering data is an important pre-processing task in information mining that can lead to considerable results. Clustering involves dividing a set of objects into a specified

number of groups. The data objects within each group should exhibit a large degree of similarity while the similarity among different clusters should be minimized. Some of the more familiar clustering methods are: partitioning algorithms based on dividing entire data into dissimilar groups, hierarchical methods, density and grid based clustering, some graph based methods, etc [3]. Most popular and widely use partitioning method is K-means clustering method.

A. K-means Clustering

K-Means Clustering is a method used to classify semi structured or unstructured data sets. This is one of the most commonly and effective methods to classify data because of its simplicity and ability to handle voluminous data sets. It accepts the number of clusters and the initial set of centroids as parameters. The distance of each item in the data set is calculated with each of the centroids of the respective cluster. The item is then assigned to the cluster with which the distance of the item is the least. The centroid of the cluster to which the item was assigned is recalculated. One of the most important and commonly used methods for grouping the items of a data set using K-Means Clustering is calculating the distance of the point from the chosen mean. This distance is usually the Euclidean Distance.

$$d_{euc} = \sum_{i=0}^{n} \sqrt{(x_i - c_i)^2} \quad (1)$$

Where; d_{euc} – Euclidean Distance, x_i - i^{th} Point in cluster and i – Number of points in cluster

The next important parameter is the cluster centroid. The point whose coordinates corresponds to the mean of the coordinates of all the points in the cluster. The main objective of the algorithm is to obtain a minimal squared difference between the centroid of the cluster and the item in the dataset [4]. Main drawback of K-means clustering is determining initial cluster centroids. Many techniques proposed to overcome above issue. Here I am going to use distortion curve (Lbow method) method to overcome this problem.

B. Vector Space Model

Vector space model or term vector model is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms. Each point in vector space represent a document. $V = \{v_1, v_2, \ldots, v_t\}$, where v_i is the weight of dimension i in vector space and t is number of terms.

[1]School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China, hasitha87@gmail.com.

[2]School of Economics, Wuhan University of Technology, Wuhan, China kapilar@sab.ac.lk.

[3]School of Management, Wuhan University of Technology, Wuhan, China skillangarathne@gmail.com.

[4]Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka, Balangoda, Sri Lanka, kapilar@sab.ac.lk.

C. Term Frequency–Inverse Document Frequency (TF-IDF)

TF–IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. Term frequency is the number of times the word appear in the document. TF can be represent as different ways like binary, raw frequency, log normalization etc. Here we use TF as log normalization.

$$tf(t,d) = \log(1 + f_{t,d}) \quad (2)$$

$tf(t,d)$ – Number of times that term t occurs in document d
The inverse document frequency is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. IDF can also be represent as many ways like unary, inverse frequency, inverse frequency smooth etc. Here we use inverse frequency smooth method.

$$idf(t,d) = \log(1 + {}^{N}/_{n_t}) \quad (3)$$

Where; N – Number of documents in the corpus, n_t – Number of document where the term t appears.

Then TF-IDF is calculated as,

$$tfidf(t,d) = tf(t,d) \times idf(t,d) \quad (4)$$

D. The similarity matrix

In the vector space model, the cosine similarity matrix is the most commonly used method to compute the similarity between two documents.

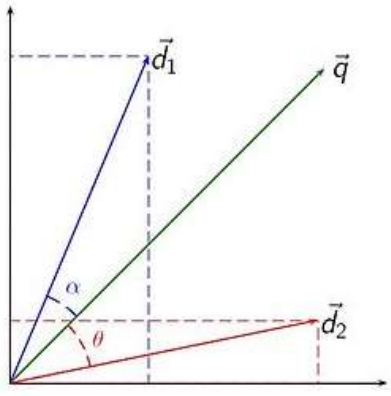

Figure 1: Document representation in vector space

Cosine similarity of the two documents can be represent as dot products of two normalized document vectors in vector space.

$$\cos\theta = \frac{d_2 . q}{\|d_2\|\|q\|} \quad (5)$$

Where $d_2 . q$ is the intersection (i.e. the dot product) of the document (d_2 in the figure to the right) and the query (q in the figure) vectors, $\|d_2\|$ is the norm of vector d_2, and $\|q\|$ is the norm of vector q. The norm of a vector is calculated as such:

$$\|q\| = \sqrt{\sum_{i=1}^{n} {q_i}^2} \quad (6)$$

II. LITERATURE REVIEW

Clustering the web documents is one of the most important approaches for mining and extracting knowledge from the web. Recently, one of the most attractive trends in clustering the high dimensional web pages has been tilt toward the learning and optimization approaches. Mahdavi et al. (2007) proposed novel hybrid harmony search (HS) based algorithms for clustering the web documents that finds a globally optimal partition of them into a specified number of clusters. By modeling clustering as an optimization problem, first, they proposed a pure harmony search-based clustering algorithm that finds near global optimal clusters within a reasonable time. Then, they hybridize K-means and harmony clustering in two ways to achieve better clustering. Experimental results revealed that the proposed algorithms can find better clusters when compared to similar methods and also illustrate the robustness of the hybrid clustering algorithms [3].

Clustering of search result is undoubtedly a tool that can provide the summarization of the millions of documents in a way where a user can easily locate his/her information. To guide user to the right cluster of documents, cluster labels should be meaningful and correctly representing the clusters. However significant a cluster is, if the label is not proper, user will never select it. Mansaf Alama & Kishwar Sadaf (2014) presented a method to label clusters based on their linking information. Their cluster labeling method is independent of any clustering method but restricted to only search result documents. They used heuristic search method to find all the linked documents of a cluster. If all or some documents of a cluster share hyperlinks, then they deduced label from these linked documents' titles using famous Apriori algorithm for frequent item-set mining. This removes the requirement of reviewing other members of a cluster in labeling process [2,3, 4 and 5].

About the web document clustering, a major part of efforts have been concerned to the learning methods such as optimization techniques. This is mostly owing to the lake of orthogonally, and existing high dimension vectors. Optimization techniques define a goal function and by traversing the search space, try to optimize its value. Regarding this definition, K-means can be considered as an optimization method. In addition to the K-means algorithm, several algorithms, such as genetic algorithm (GA) [6, 7], self-organizing maps (SOM) [8] and ant clustering [9,10] have been used for document clustering.

As one of the most fundamental yet important methods of data clustering, center-based partitioning approach clusters the dataset into k subsets, each of which is represented by a centroid or medoid. In 2014 Jian-Ping Mei & Lihui Chen proposed a new medoid-based k-partitions approach called Clustering Around Weighted Prototypes (CAWP), which works with a similarity matrix. In CAWP, each cluster is

ISSN: 2409-2770

characterized by multiple objects with different representative weights. With this new cluster representation scheme, CAWP aims to simultaneously produce clusters of improved quality and a set of ranked representative objects for each cluster. An efficient algorithm is derived to alternatingly update the clusters and the representative weights of objects with respect to each cluster [11,12 and 13].

Fersini, Messina & Archetti (2009) presented a document clustering method, which takes into account both contents information and hyperlink structure of web page collection, where a document is viewed as a set of semantic units. They exploited this representation to determine the strength of a relation between two linked pages and to define a relational clustering algorithm based on a probabilistic graph representation. The experimental results show that the proposed approach, called RED-clustering, outperforms two of the most well-known clustering algorithm as k-Means and Expectation Maximization [14, 15 and 16].

Both general-purpose and text-oriented techniques exist and can be used to cluster a collection of documents in many ways. Carullo et al. (2009) proposed a novel heuristic online document clustering model that can be specialized with a variety of text-oriented similarity measures. An experimental evaluation of the proposed model was conducted in the e-commerce domain. Performances were measured using a clustering-oriented metric based on F-Measure and compared with those obtained by other well-known approaches. The obtained results confirmed the validity of the proposed method both for batch scenarios and online scenarios where document collections can grow over time [17].

Major drawback of the K-means clustering is to determine initial centroids. To find an appropriate value for k, we can use the method in to generate a distortion curve for the input data by running a standard K-means operation on all k values between 1 and Kmax. Then computed the resulting clustering distortion to find a specific range that features a minimal decrease in average diameter [18, 19].

III. METHODOLOGY

Based on concept of K-means clustering below discussed the steps,

Step 1: Select K initial cluster centroids, C1, C2, C3, ..., Ck. K number of observations is randomly selected from all N number of observations, according to the number of clusters, and these become centers of the initial clusters. In here we use different K values and repeat the process.

Step 2: Proceed through the list of items. For each of the remaining N-K observations, assign an item to the cluster whose centroid is nearest (distance is computed by using Euclidean) and re-calculate the centroid for the cluster receiving the new item or for the cluster losing the item.

Step 3: Repeat Step 2 until no more reassigning. Rather than starting with a partition of all items into K preliminary groups in Step 1, we could specify K initial centroids (seed points) and then proceed to Step 2 [20, 21 and 22].

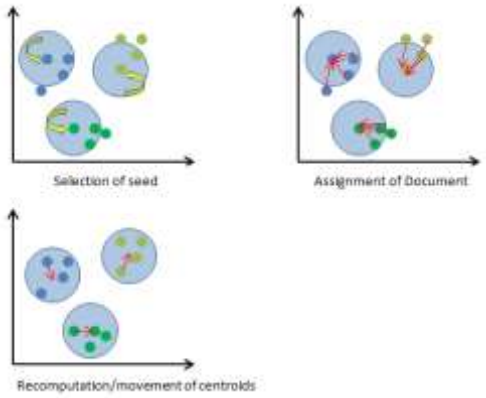

Figure 02: Clustering

Sample R code for document clustering using K-means.

```
library(tm)
file <- "reut2-000.xml"  ### download reuters21578 data first (use first 1000 documents; 1984/85)
reuters <- Corpus(ReutersSource(file), readerControl = list(reader = readReut21578XML))
reuters <- tm_map(reuters, as.PlainTextDocument) ## Convert to Plain Text Documents
reuters <- tm_map(reuters, tolower) ## Convert to Lower Case
reuters <- tm_map(reuters, removeWords, stopwords("english")) ## Remove Stopwords
reuters <- tm_map(reuters, removePunctuation) ## Remove Punctuations
reuters <- tm_map(reuters, stemDocument) ## Stemming
reuters <- tm_map(reuters, removeNumbers) ## Remove Numbers
reuters <- tm_map(reuters, stripWhitespace) ## Eliminating Extra White Spaces
dtm <- DocumentTermMatrix(reuters)  ## create a term document matrix
inspect(dtm[1:10, 5001:5010])
findFreqTerms(dtm, 100)
findAssocs(dtm, "washington", .4)
dtm_tfxidf <- weightTfIdf(dtm) ## do tfxidf
inspect(dtm_tfxidf[1:10, 5001:5010])

## do document clustering ##

m <- as.matrix(dtm_tfxidf) ### k-means (this uses euclidean distance)
rownames(m) <- 1:nrow(m)
norm_eucl <- function(m) m/apply(m, MARGIN=1, FUN=function(x) sum(x^2)^.5)   ###normalization
m_norm <- norm_eucl(m)
cl <- kmeans(m_norm, 10) ### cluster into 10 clusters
cl
table(cl$cluster)
```


ISSN: 2409-2770

plot(prcomp(m_norm)$x, col=cl$cl) ### show clusters using the first 2 principal components
findFreqTerms(dtm[cl$cluster==1], 50)
inspect(reuters[which(cl$cluster==1)])
findFreqTerms(dtm[cl==1], 50)

Determine best K value by applying distortion curve.
To find an appropriate value for k, we used the method in to generate a distortion curve for the input data by running a standard K-means operation on all k values between 1 and K_{max}. In Elbow curve (or distortion curve) x axis represents number of cluster and y axis represents within groups sum of squares in clusters. We then computed the resulting clustering distortion to find a specific range that features a minimal decrease in average diameter. This value is the best K value for the given document set[23 and 24].

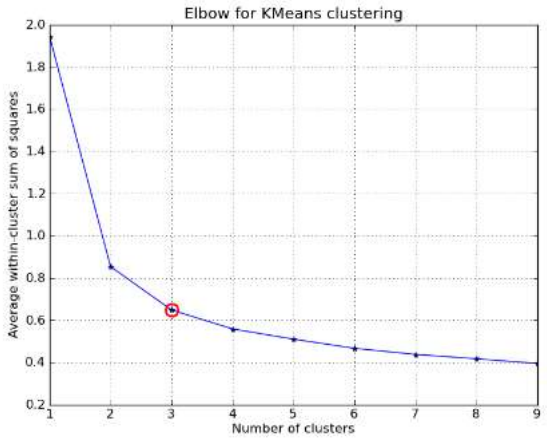

Figure 03: Distortion curve

So, Major drawback of K-means clustering is to determine initial cluster centroids. By using distortion curve method we can identify optimal number of initial centroids for given search result. By using this method we can cluster web search results in better way.

IV. DISCUSSION

Clustering text documents is different than clustering other data. Before clustering the text documents we have to represents all the web search results using vector space model. After user input the relevant search query we measure similarity between each search results and the query then we analyze the results by using cosine similarity.

Here we discussed K-means clustering method for cluster web search results from search engine. K-means clustering is the better way to segment web search results into different clusters according to their similarity. The major drawback of the K-means clustering is to determine initial centroids. Here we proposed distortion curve to find the optimal number of initial centroids. According to the above we computed the resulting clustering distortion to find a specific range that features a minimal decrease in average diameter.

Increasing number of web pages and web documents, complexity of tasks of search engine increased. Whatever user input their query search engine is responsible for give best answer for their requirements. As an example if user input keyword as "Jaguar" there may be thousands of search results may appear. It is difficult to identify what user actually wants. May be he expect Jaguar Animal may be he expect famous Jaguar car. So, search engine can't decide what he needs search engine shows the results as per the page rankings. Clustering search result is the better answer for above issue. Clustering can segment all the search results and user can easily navigate whatever he needs.

There are several problems on search result clustering using K-means. Major limitations of K-means clustering are makes hard assignments of points to clusters. A point either completely belongs to a cluster or not at all. Another major limitation is it works well only for round shaped and of roughly equal sizes/density clusters.

V. CONCLUSION

In this discussion we proposed K-means clustering method for segment web search results in better way. For answer the major drawback in K-means clustering we used distortion curve method to decide initial cluster points .For web search result clustering K-means is better but increasing number of search results means need lots of power for clustering the results. Users are expecting better results as well as better speed. They will not satisfy if the results take longer time to display. So, answer above problem now search engine designers are moving into big data concept. Which will process the result in parallel way using MapReduce functions.

REFERENCES

[1] GG. He, "Authoritative K-Means for Clustering of Web Search Results," no. June, 2010.

[2] R.M. Kapila Tharanga Rathnayaka, Wei Jianguo and D.M.K.N Seneviratne, "Geometric Brownian Motion with Ito lemma Approach to evaluate market fluctuations: A case study on Colombo Stock Exchange", 2014 International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC'2014- IEEE), Shanghai, China, 2014.

[3] M. Alam and K. Sadaf, "Labeling of Web Search Result Clusters using Heuristic Search and Frequent Itemset," Procedia - Procedia Comput. Sci., vol. 46, no. Icict 2014, pp. 216–222, 2015.

[4] M. Mahdavi, M. H. Chehreghani, H. Abolhassani, and R. Forsati, "Novel meta-heuristic algorithms for clustering web documents," vol. 201, pp. 441–451, 2008.

[5] Rathnayaka, R.M. K.T. and Seneviratne, D.M.K.N, "G M (1, 1) Analysis and Forecasting for Efficient Energy Production and Consumption", International Journal of Business, Economics and Managment Works, Kambohwell Publisher Enterprises, 1 (1), 6-11, 2014.

[6] P. P. Anchalia, "MapReduce Design of K-Means Clustering Algorithm," 2013.

[7] Jayathileke, P. M. B., and Rathnayaka, R.M. K. T. "Testing the Link between Inflation and Economic Growth: Evidence from Asia", Modern Economy,4, 87.

[8] Jones Gareth, A.M. Robertson, Santimetvirul Chawchat, P. Willett, Non-hierarchic document clustering using a genetic algorithm, Informat. Res. 1 (1) (1995).

ISSN: 2409-2770

[9] R.M.K.T Rathnayaka and Zhong-jun Wang, "Prevalence and effect of personal hygiene on transmission of helminthes infection among primary school children living in slums", International Journal of Multidisciplinary Research Journal; ZENITH, ISSN: 2231-5780, Vol 02, April 2012.

[10] V.V. Raghavan, K. Birchand, A clustering strategy based on a formalism of the reproductive process in a natural system, in: Proceedings of the Second International Conference on Information Storage and Retrieval, 1979, pp. 10–22.

[11] R.M Kapila Tharanga Rathnayaka, D.M. Kumudu Nadeeshani Seneviratne and Zhong- jun Wang, "An Investigation of Statistical Behaviors of the Stock Market Fluctuations in the Colombo Stock Market: ARMA & PCA Approach", Journal of Scientific Research & Reports 3(1): 130-138, 2014; Article no. JSRR, www.sciencedomain.org

[12] Labroche, N. Monmarche', G. Venturini, AntClust: ant clustering and web usage mining, Genet. Evolut. Comput. Conf. (2003) 25–36.

[13] R.M. Kapila Tharanga Rathnayaka and Zhong-jun Wang, "Enhanced Greedy Optimization Algorithm with Data Warehousing for Automated Nurse Scheduling System", E-Health Telecommunication Systems and Networks, 2013, http://www.SciRP.org/journal/etsn.

[14] X. Cui, T.E. Potok, P. Palathingal, Document Clustering using Particle Swarm Optimization, IEEE Swarm Intell. Symp. (2005) 185–191.N.

[15] R.M. Kapila Tharanga Rathnayaka and Zhong-jun Wang, "Influence of Family Status on the Dietary Patterns and Nutritional Levels of Children", Food and Nutrition Sciences, 2013, http://www.SciRP.org/journal/fns .

[16] J. Mei and L. Chen, "Expert Systems with Applications Proximity-based k -partitions clustering with ranking for document categorization and analysis," Expert Syst. Appl., vol. 41, no. 16, pp. 7095–7105, 2014.

[17] R.M Kapila Tharanga Rathnayaka, "Cross-Cultural Dimensions of Business Communication: Evidence from Sri Lanka", International Review of Management and Business Research, 3(3), 1579-1587, 2014; ISSN: 2306-9007, 2014, www.irmbrjournal.com

[18] E. Fersini, E. Messina, and F. Archetti, "A probabilistic relational approach for web document clustering," Inf. Process. Manag., vol. 46, no. 2, pp. 117–130, 2010.

[19] M. Carullo, E. Binaghi, and I. Gallo, "An online document clustering technique for short web contents," Pattern Recognit. Lett., vol. 30, no. 10, pp. 870–876, 2009.

[20] Rathnayaka, R.M. K.T. and Seneviratne, D.M.K.N, "A Comparative Analysis of Stock Price Behaviors on the Colombo and Nigeria Stock Exchanges", International Journal of Business, Economics and Managment Works, Kambohwell Publisher Enterprises, 2 (2), 12-16, 2014.

[21] D. Ren, D. Zheng, G. Huang, S. Zhang, and Z. Wei, "Parallel Set Determination and K-means Clustering for Data Mining on Telecommunication Networks," 2013.

[22] Rathnayaka, R.M. K.T., Seneviratne, D.M.K.N and Jianguo,W., "Grey system based novel approach for stock market forecasting", Grey Systems: Theory and Application, Emerald Group Publishing Limited, 5 (2), 2015.

[23] Seneviratne, D.M.K.N and Long, w., "Analysis of Government Responses for the Photovoltaic Industry in China", Journal of Economics and Sustainable Development, 4 (13), 2013.

[24] Seneviratne, D.M.K.N and Jianguo, w., "Analying the causal relationship between stock prices and selected microeconomic variables: evidence from Sri Lanka", ZENITH international journal of Business economics & Management research, 3 (9), 2013.

I am Hasitha Indika Arumawadu from Sri Lanka, obtained B.Sc. in 2011 from University of Kelaniya, Sri Lanka. Currently I have been studying Masters degree at Wuhan University of Technology in China. I am reading my Masters Degree in the field of Computer Science & Technology and highly intersested in Data Mining and Big Data.

I am R.M.K.T Rathnayaka, working as a Lecturer in the Department of Physical Sciences and Technology, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka and currently pursuing my higher studies at Wuhan University of Technology in China. I am reading my Doctoral Degree in Applied statistics. I'm highly interested in research in the fields of Financial Mathematics, Time series Modeling and Graph Theory .

I am S.K.Illangarathne working as a Senior Assistant Librarian, Main Library, Rajarata University of Sri Lanka and currently pursuing my higher studies at Wuhan University of Technology in China. I am reading my Doctoral Degree in Management Science and Engineering in the field of Information Management and Information Systems in School of Management. I'm highly interested in research in the fields of Multiple criteria decision modeling, and ranking organizations specially in Libraries, cloud computing and web data management.