A Clustering Approach for Optimization of Search Result

Shruti Kohli and Shashi Mehrotra
Birla Institute of Technology, Mesra, India
Email: kohli.shruti@gmail.com, sethshashi@rediffmail.com

***Abstract*—With the massive increase in the use of internet and web data, retrieval of relevant information quickly is very important, and some efficient technique is required for data analysis. Grouping of objects may be helpful for data analysis, where clustering is useful. Clustering on-line result is a challenging technique. Search option is excessively used in almost every website. The study proposes a hybrid clustering algorithm to optimize search result of the website. The domain of the website is medical. Matrices will be used to analyze user behaviour. User trust will be measured. Clustering of search result will facilitate users to get the relevant information in a quick manner.**

***Index Terms*—clustering, analytics, cross validation, Euclidean distance**

I. Introduction

Search results are presented to users in a list, which contains title and snippet. Most of the results are not relevant to the user for many reasons such as one word can be used for various purposes. Select relevant information from the list is not easy as well as time consuming. Results are clustered in meaningful folder will facilitate user to search relevant result in a quick manner. Web Analytics is the field concerned with understanding and optimizing the Web usage. "Web Analytics is the measurement, collection, analysis and reporting of internet data for the purpose of understanding and optimizing web usage" [1]. Web Analytics can be classified in two categories: one is off-site web analytics and the other is on-site web analytics. Website's potential audience, visibility and commencing that are happening on the internet as a whole are measurements, used for off-site web analytics. On-site web analytics measures a visitor's use for a website that is owned or maintained by an organization or an individual. Clustering is the process of partitioning a set of data (or objects) in a set of meaningful sub-class called clusters. Clustering organizes data items into clusters, such that items within a cluster are more similar in nature than they are to items in the another cluster [2].

Cluster analysis is being used in various applications, such as data analysis, pattern recognition, image processing and business purpose [3].

Manuscript received March 16, 2015; revised October 30, 2015.

A. Motivation

There is a massive increase in the use of internet and web data; hence, Web Analytics plays an important role. For the web data, data analysis is required in various ways. It is needed to group these data or objects for effective analysis, where clustering is very useful. Clustering occurs in every aspect of our daily life. People encounter a large amount of information store, and analyzes it for various uses. Grouping these data into a set of clusters is one of the important tasks [1]. Clustering is an important procedure in a variety of fields, yet cluster analysis is a challenging problem, as many factors play an important role. Same algorithm with different parameters, using different presentation or using different similarity measure may generate different output.

II. Related Work

Rui Xu [4] presented survey of various clustering algorithm, its applications, and various proximity measures used for similarity checking between a pair of objects, object and cluster. They also discussed cluster validation.

A. K. Jain *et al.* [5] presented a review of data clustering methods. They also covered some applications of clustering algorithms such as informal retrieval, image segmentation and object recognition. Different approaches of clustering are explained in the paper.

P. Rai and S. Singh [6] provided survey of various clustering techniques used in data mining. Cluster can be expressed in various ways depending upon the clustering technique used such as: Any object can belong to only one cluster. An object may belong to more than one cluster. Object may belong to each cluster with a certain probability i.e. they may be probabilistic.

Ugo scaiella *et al.* [7] proposed a novel labelled-clustering algorithm, to move to graph-of-topic paradigm based on the spectral properties of a graph from big-of-words. Search engine returns clustering of short text into a list of folders, summarize the context of the searched keyword within the result pages.

M. Granitzer *et al.* [8] proposed an interactive system WebRat that is for visualizing and refining search result sets. Documents matching a query are clustered and visualized as a counter map of islands. Thematic clusters were built, analyzed and visualized in real time. It can be

©2016 Journal of Image and Graphics
doi: 10.18178/joig.4.1.63-66

used to interactively visualize and refine queries by selecting from the keywords and presented clusters.

S. Khy *et al.* [9] proposed a novelty based document clustering method. Higher weight is assigned to recent documents, and generate cluster that focus on recent topics.

T. Rui *et al.* [10] aims to find out the method of applying other nature-inspired optimization techniques such as bats, and cuckoos for clustering to use for web intelligence data. Experiments were conducted over four data sets.

N. Yang *et al.* [11] proposed a new clustering method by extending K-means. It combines the links and in-snippets together. The attached short text to in-link is valuable information and is helpful to reach high clustering quality.

Md. Ezaz Ahmed *et al.* [12] addresses the applications of data mining tools Weka by applying K-means clustering from huge data sets. They improve the quality of websites by grouping similar websites in the group for which they use clustering.

Rani Qumsiyeh *et al.* [13] proposed query-based cluster, which generate concise clusters of documents covering various subject area.

Hector Menendez *et al.* [14] presented a strategy to reduce the dimension of the attribute. A. K. Sharma *et al.* [15] presented a technique for search result optimization that is based on historical query logs. The technique first clusters query in query logs and after that captures the sequential patterns of clicked web pages in every cluster using an algorithm. All queries are considered to be unassigned to a custer. Each query is tested and if the similarity value is above the prespecified threshold value, then the query is placed in the cluster. This process is repeated until all the queries are placed in some cluster. Search result list is re-ranked finally.

Larry Kim [16] proposed a system that generate the database such as compilation, manipulation and segmentation analysis for search engine optimization and marketing tool.

S. Kohli *et al.* [17] presented a web analytics tool, "Keyword Similarity Measure Tool" (KSMT. The tool aims to take care of the limitations of similar keywords in the report and improves the data accuracy, thus optimizing the report. It aims to provide a consolidated view and content analysis, by combining the matrices like bounce-rate, visits for the similar content analysis. KSMT algorithm was used to measure keyword similarity and combine keyword based on factor of similarity.

S. K. Jayanti *et al.* [18] proposed and implemented an algorithm WESPACT, which use the genetic algorithm to classify the web pages as spam. Decision tree's output is the result of the algorithm. Study used WEBSPAM 2007 dataset for experiment.

Ela Kumar *et al.* [19] designed spamizer that detects spam host or page. System analyzes and improves currently given five link algorithms. It generates spamizer spamicity score by combining spamicity score of the used algorithm merge the result obtained.

Cailing Dong *et al.* [20] proposed a web browser plug-in to support online web spam detection. Spam pages are filtered on the client side, i.e., web browser. They developed an ensemble learning framework for detection of online web spam.

Ian Grout and Abu Khari Bin A'ain [21], work to extend an on-line tutorial system to analyse how user experience for tutorial for education context. Postgraduate students are taken as target audience. They used PHP to create web pages for tutorial. Paper considers four key aspects: 1) Key metrics are identified for analyzing the use of tutorial. 2) Data and user actions are identified to generate and present the metrics. 3) Implementation and integration of script into already existing system. 4) To enhance the learning experience of student, analyze and modify the tutorial system.

Jose G. Moreno and Gael Dias [22] performed analysis over frequently used algorithm for web search result clustering and evaluation metrics. They used data set OPD-239 and Moresque for experiment. Initially, paper provides the result of the algorithm by using best parameter setting. Then they showed that a simple strategy of the algorithm can lead to a scalable and real world solution. Finally some conclusion is drawn about evaluation metrics and their bias to the number of output clusters. Set of web result collected for each query, and classified manually into the disambiguation Wikipedia pages, which formed the reference clusters.

Jinxiu Yu *et al.* [23] suggested a new form of metric that measure web search results in satisfaction encompassing user behaviour. They introduced the user behaviour, click-through rate that measure the performance of search engine. Users are college students and graduates and data collected from various search engine users.

Rana Forsati *et al.* [24] proposed an algorithm to optimize K-means algorithm by integrating harmony clustering with K-means, which is less dependent on the initial parameters such as randomly chosen initial cluster centres. This method combines the speed of K-means with the power of HSCLUST.

III. Research Gaps

- The selection of distance measure
- Choosing the initial clusters
- High Dimensionality

In addition to it, following research gaps can be identified:

- Page ranking manipulation, which effect the search result and user trust as well.
- There is no clarity in discrimination of the keyword of that domain.

IV. Problem Statement

The clustering techniques developed so far need some improvement for web. It is to be developed keeping in mind the research gaps in the techniques. As noted in the previous section, the techniques proposed so far suffer from the problems of scalability, high dimensionality,

loosely structured, complex attributes, selection of initial clusters and distance measure. Research is required to come up with a technique efficient in terms of one or more of above criteria for the web such as search result as it is used excessively in almost every website.

V. Objectives

- Improve search result of a medical website. That will take care of large volume of data as well as high dimensional features.
- Clustering short text fragment returned in the result of keyword given using hybrid approach.
- Use keywords to make folder labels.
- Analyze user behaviour using some metrics such as visit characterization, engagement term and conversion terms.
- Measure user trust.

The above mentioned task will be accomplished keeping in mind the requirements of clustering for web search result. The principal behind search result clustering is to group results into distinct clusters so that the user can choose relevant document in an efficient way.

VI. Research Methodology

A. Literature Survey

A comprehensive review will be conducted to understand different aspects of the clustering algorithm used for web elements.

B. Pre-Processing

Data is to be pre-processed such as missing values are to be filled with common values.

C. Design and Implementation

For clustering K-means, algorithm combined with a genetic algorithm will be used. As K-mean algorithm works well for large datasets but lack global perspective, need to define k initially and sensitive to outliers. It seems a hybrid algorithm that combines the features of both techniques can result in an algorithm that outperforms either one individually.

The proposed algorithm will be implemented using MATLAB language. Spam detection to improve search result accuracy. Web spam can significantly deteriorate the quality of search result.

D. Experimental Focus for Evaluating the Technique

- Cross-Validation technique is to be used for conducting the experiments.
- Experiments will be conducted.
- Evaluation parameters: Relevancy and speed are to be taken into consideration.

Following test will be performed:

- Precision
- Recall
- F-Measure item True negative item True negative
- User evaluation to see the satisfaction

VII. Architecture of Search Result Optimizer System

- Similarity Analyzer will measure the closeness between objects using Euclidian distance.
- Clustering Tool will group the search result, to improve visualization and fast access.
- Analysis tool will be developed to measure user behaviour and user trust.

VIII. Conclusion

Use of electronic data and internet is increasing day by day. Thus, Web Analytics plays an important role. Clustering of web element is useful for grouping of data and objects in various ways for efficient analysis. Clustering is an important unsupervised technique, which is being used for research work in various fields. We feel proposed system will improve search result visualization and length time keeping in mind the relevancy of the keyword. User trust is also an important factor, which is taken into consideration.

References

[1] D. Waisberg and A. Kaushik, "Web analytics 2.0: Empowering customer centricity," *The Original Search Engine Marketing Journal*, vol. 2, no. 1, pp. 5-11, 2009.

[2] N. Grira, M. Crucianu, and N. Boujemma, "Unsupervised and semi-supervised clustering: A brief survey," *A Review of Machine Learning Techniques for Processing Multimedia Content*, Report of the MUSCLE European Network of Excellence (6th Framework Programme).

[3] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 2nd ed., 2006.

[4] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transaction on Neural Networks*, vol. 16, no. 3, pp. 645-678, May 2005.

[5] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Computing Survey*, vol. 31, no. 3, pp. 264-323, September 1999.

[6] P. Rai and S. Singh, "A survey of clustering techniques," *International Journal of Computer Application*, vol. 7, no. 12, pp. 1-5, October 2010.

[7] U. Scaiella, P. Ferragina, A. Marino, and M. Ciaramita, "Topical clustering of search results," in *Proc. Fifth ACM International Conference on Web Search and Data Mining*, Seattle, Washington, USA, February 2012, pp. 223-232.

[8] M. Granitzer, W. Kienreich, V. Sabol, and G. Doiinger, "WebRat: Supporting agile knowledge retrieval through dynamic, incremental clustering and automatic labelling of web search result sets," in *Proc. Twelfth IEEE International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises*, June 2003.

[9] S. Khy, Y. Ishikawa, and H. Kitagawa, "A novelty-based clustering method for on-line documents," *World Wide Web*, vol. 11, no. 1, pp. 1-37, 2008.

[10] T. Rui, S. Fong, X. Yang, and S. Deb, "Nature-Inspired clustering algorithms for web intelligence data," in *Proc. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Dec. 2012.

[11] N. Yang, Y. Liu, and G. Yang, "Clustering of web search results based on a combination of links and in-snippets," in *Proc. Eighth Web Information System and Application Conference*, 2011.

[12] M. E. Ahmed and P. Bansal, "Clustering technique on search engine dataset using data mining tool," in *Proc. Third International Conference on Advanced Computing & Communication Technologies*, 2013.

[13] R. Qumsiyeh and Y. K. Ng, "Enhancing web search using query-based clusters and labels," in *Proc. International Conference on Web Intelligence and Intelligent Agent Technology*, 2013.

©2016 Journal of Image and Graphics

[14] H. Menéndez, G. Bello-Orgaz, and D. Camacho, "Features selection from high dimensional web data using clustering analysis," in *Proc. 2nd International Conference on Web Intelligence, Mining and Semantics*, Craiova, Romania, June 2012.
[15] A. K. Sharma, N. Aggarwal, N. Duhan, and R. Gupta, "Web search result optimization by mining the search engine query logs," IEEE Explorer, 2010.
[16] L. Kim, "Integrated web analytics and actionable workbench tool for search engine optimization and marketing," United States Patent Application Publication, US 2009/0292677 A1, Nov. 26, 2009.
[17] S. Kohli, S. Kaur, and G. Singh, "A website content analysis approach based on keyword similarity analysis," in *Proc. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Macau, China, Dec. 2012.
[18] S. K. Jayanti and S. Sasiksla, "WESPACT: Detection of web spamdexing with decision tree in GA perspective," in *Proc. International Conference on Pattern Recognition, Informatics and Medical Engineering*, 2012.
[19] E. Kumar and S. Kohli, "Improving link spam detection: Design and development of spamizer," in *Proc. World Congress on Engineering and Computer Science*, San Francisco, USA, Oct. 2011.
[20] C. Dong and B. Zhou, "An ensemble learning framework for online web spam detection," in *Proc. IEEE 12th International Conference on Machine Learning and Applications*, 2012.
[21] I. Grout and A. K. B. A'Ain, "Adapting an on-line tutorial tool with web analytics to incorporate analysis of tutorial use," in *Proc. IEEE 15th International Conference on Interactive Collaborative Learning*, Sep. 2012.
[22] J. Moreno and G. Dias, "Easy web search result clustering: When baselines can reach state-of-the-art algorithm," in *Proc. 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, April 2014, pp. 1-5.
[23] J. Yu, Y. Lu, F. Zhang, and S. Sun, "A metric for measuring web search results satisfaction incorporating user behavior," in *Proc. IEEE 2nd International Conference on Cloud Computing and Intelligent Systems*, Oct. 2012.
[24] R. Forsati, M. Meybodi, M. Mahdavi, and A. Neiat, "Hybridization of k-means and harmony search methods for web page clustering," in *Proc. IEEE/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Canada, Dec. 2008.

Dr Shruti Kohli is working as an Assistant Professor in the Department of Computer Science in Birla Institute of Technology, Mesra, Noida Centre. She did her Master Degree in Operational Research from the University of Delhi in 2001. She obtained Master's degree in Computer Application from IGNOU in 2002. She did her M.Phil in Operational Research from University of Delhi in year 2004 and obtained the Ph.D. in (Technology in the year 2012) from Birla Institute of Technology. The area of her doctoral research work was web intelligence. She has also cleared Digital Fundamental Course from Google Analytics Academy.
Her areas of interest include Information retrieval, Operational Research, Data Mining, Web Analytics. At present she is guiding three PhD students in mainly in area of web intelligence and one in the area of Mobile Ad-Hoc Networks. She has already mentored two M.Tech Scholars and currently mentoring 1 M.Tech Scholar in the area of Semantic web.
She has presented papers in many international and national conferences and had been a resource person in DST sponsored FDPs. She is an active blogger and has great interest Mobile Apps development. She had been conducting Mobile App workshop in college and is currently running Mobile Incubator Cell in her college. She is teaching subjects like web technology, Simulation & Modeling, E-commerce and had written 2 course book on Web technology for Maharishi Dayanand University (MDU) and KKHSOU (Kirshna Kanta Handiqui State Open University). She has also co-authored a book on banking in which she highlighted role of IT in banking. She had been a member of program committee and review committee of many national international conferences and had been session chair and speaker in few conferences. She was presented best paper award in 2014 by South Asian University for presenting her research work in international workshop titled "Machine Learning and Text Analytics". She is in the advisory word of WARSE (Word Academic Research in Science and Engineering). She is active member of IEEE, IAENG International Society for Engineers and Soft Computing Research Society.
Currently she is working on a UGC Major Research Project titled "Smart use of web analytics and data mining techniques for improving online Information Retrieval" and have appointed project fellow for the same.

Shashi Mehrotra Seth is employed as an Assistant Professor in Information Technology Department, Tecnia Institute of advance Studies, Delhi, India.
She is currently pursing Ph.D. (CSE) from Birla Institute of Technology, Mesra, India. She did an M.Tech in Computer Science Engineering in 2010. She obtained M.Phil degree in Computer Science in 2005 from Madurai Kamraj University, India. Her area of research is Data Mining, and Data Analytics.
She has presented papers in national & international conferences and has received awards for her work. She has also published papers in national and international journals and conference proceedings.
She taught and mentored B.Tech, M.Tech and MCA students in her teaching career.
She is member of IACSIT and life time member of Computer Society of India (CSI).