iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** III **Month of publication:** March 2020

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

A Recommended System for Selection of Sample Research Areas in Computer Science

Shawni Dutta[1], Samir Kumar Bandyopadhyay[2]
[1]*Lecturer, Department of Computer Science, The Bhawanipur Education Society College, Kolkata, India*
[2]*Academic Advisor, The Bhawanipur Education Society College, Kolkata, India*

***Abstract:** This study applies text mining methods to corpus of topics related to Computer Science field for grouping them into clusters based on their similarities. By this methodology, priorities of research areas are identified and suggested as well to the Scholars or students, so that they can drive themselves into their comfortable zones. The objective of the paper is to built a recommender system that automatically suggests other research fields related to or overlapped with the particular research field. Existing clustering techniques are used for obtaining the recommendation system and later the cluster analysis results are compared.*
***Keywords:** Clustering, Tokenization, crawler, Ranker and K-means*

I. INTRODUCTION

Based on emerging technologies in the field of Computer Science, research scholars as well as teachers can carry their research in proper direction. However, due to exponential growth of research articles, people often feel bore while searching a particular Research Topic. In fact, they often fail to identify the relation between two research fields or topics. This study benefits scholars to explore research trends in certain fields or on particular topics and other alternate fields that are overlapped with the selected fields.
While identifying relationship among several topics, text mining[1] is used as tool that analyses information among the topics from a large collection of text and draw a relationship among the textual data. Unsupervised Learning method is an approach of Text mining that identifies an interesting pattern among those data. Text clustering can be in different forms such as documents, paragraphs, sentences or terms. Clustering helps organizing documents for quick retrieval and support browsing.
With the help of data mining methods, such as clustering algorithm, it is possible to discover the key characteristics from the students' interests and analysing those interests similar research fields are suggested. Clustering[1] is an unsupervised machine learning method automatically categorizes topics into groups without accepting any predefined labels. Clustering works in such a fashion that confirms that the topics in the same group will have the higher similarity and those topics will carry high dissimilarity with respect to topics in the other groups formed by the corresponding clustering algorithms. When a person is interested on a particular research topic, the corresponding cluster of that field is identified and the other members (i.e. research topics) of the clusters are recommended to that person. In this way, a person may find more interest in those recommended fields as well. The clustering is defined as [2]:
Given a dataset D, clustering aims to find out set of subsets C= C1,…,CK such that, D= C1 ∪ C2 U..U CK and Ci ∩ Cj = ∅ given that i ≠ j.
This paper provides ideas regarding topics in Computer Science field and their related branches so that anyone can choose their research topic wisely and pursue their studies. For this purpose, an application is developed that automatically gives recommendations depending on students' research interest. The main motivation of this work is to provide a structured method to the students, research scholars so they come across possible ways regarding their research. For example, when someone tries to pursue regarding Cloud Computing, he/she may also find interest in Cloud cryptography . Hence, it would be helpful for the next generation Research Scholars while choosing their research topic without getting distracted from his/her research field. Following facts are observed while developing this application-

A. There are several topics available in Computer Science for carrying out Research.
B. Students often lose their way while searching particular Research Topic.
C. Recommendations of Research topics should be inherent to Students' Research interest.

The proposed system will consider one research topic as input and will recommend other research fields related and/or overlapped with the input.

II. LITERATURE REVIEW

XIN LI et. al.[3] implemented an online Personalised Blog Reader (PBR) System that analyses user's interest and recommends their potential favourite topics over those non-favourite ones. In order to obtain this system, the following procedures are accomplished:

1) An automated *crawler* collects posts from a given user's favourite blogs in real-time and a story database is formed.
2) Clustering is applied to acquire new topics from the story database, and a main story for each topic is recommended.
3) Finally, a *ranker* is used that presents a final reading list based on personalized reading preferences. The incremental clustering algorithm, consists of both static clustering and dynamic clustering, captures temporal relationships between stories covered in the current window of observation and stories in adjacent area, and satisfies the online processing requirements. The results indicate that the proposed algorithm made stories in an optimised way.

Tian Gao et al. [4] aimed to identify the hot topics of emergencies in the Internet information sources. For this purpose, a web-crawler is used to obtain emergency related document from the internet. The features from Web document are matched and denoted by named entity approach based on the event framework. Finally, Fuzzy C-Means Clustering is applied to detect hot topics.

Kumar Shubankar et. al. [5] proposed a novel approach that aims to identify topics from a corpus of research topics. They obtained topics using closed frequent keyword set and employed a time-independent modified iterative PageRank algorithm that is applied on DBLP dataset in order to obtain a ranked set of research papers within a topic cluster. Experimental results proved that this proposed algorithm outperforms the state-of-art. However, this paper also provides insights about exploring several domains like topic correlation, web-clustering, document clustering and many more.

In [6], the problem domain of separating scientific documents from the large collection of document is addressed. For this purpose, a statistical method known as topic modelling is employed as an automated method that clusters documents based on their content similarity.

Apart from focusing only in topic of interest, [7] focused on rejecting off-topics messages. They mainly focused on newsgroup messages and applied Hidden Markov Model[7] for identifying topics of interest. They also prepared a novel algorithm known as UTC that clusters topics and generates a hierarchical organization of topics (and documents) without involving any human supervision. The Unsupervised Topic Clustering (UTC) system allows different users to view the text corpus at different resolutions and find the desired information in context.

In [8], topic detection incremental clustering algorithm is used. The performance of this model can be improved by considering sub-topic points, refined feature set and pre-clustering operation using "age" features of stories. To estimate true number of topics, the method is used to see whether two topics can be merged into one cluster. This method is applied on available datasets and experimental results show that this method has a good performance within optimized execution time as compared to k-means and CMC method.

In this paper, several existing clustering methods were employed for topic clustering such as K-means[9], Hierarchical Clustering[10], Spectral Clustering[11], Density-based clustering, Affinity Propagation algorithm[12].

K-Means[9] ,popular "clustering" algorithms, belongs to the category of flat clustering algorithm. K-means randomly chooses K random objects initialize them as center of clusters from the given dataset. Each data point is assigned to its' closest central point's group by finding the shortest distance between the point and k central points, and then k new centroids are computed. This distance between each data object and cluster centroid is measured by Euclidean distance. Euclidean distance[13] is defined as follows-

$d(i,j)= \sqrt{(\sum_{i=0}^{n} (Xi - Yi)^2)}$ where Xi and Yi are two n dimensional data.

The process is continued till the closeness between the old centroids point and new centroids is less than a given threshold. The K-means algorithm depends on minimizing the sum of squared error function.

Hierarchical Clustering[10] proceeds by grouping data based on nearest data points. This clustering method follows two approaches for grouping- One is Top-down and the other one is Bottom-up. Agglomerative algorithm follows Bottom-up approach where each instance is considered to be as individual cluster and later they are combined depending on similarity factor and this process proceeds until all instances are covered.

The spectral clustering[11] technique splits a given data set into different clusters based on some specific properties. Spectral clustering partitions instances into disjoint clusters using eigen-structure of a similarity matrix. This clustering is strongly dependent on graph data and this technique clusters the nodes in the graph that are connected to each other

However, there are another two variations of the algorithm recursive spectral. Two models based on affinity 'rbf' and 'nearest neighbors' are used in the proposed system.

Density-based clustering[14] is capable of dealing with noise in the database and obtains arbitrary cluster shapes. DBSCAN clustering algorithm finds clusters by considering an arbitrary point and obtains the clusters considering two parameters Eps and Minpts. The Eps-neighborhood of an arbitrary point q of data set D can be identified as follows-

$$N_{EPS}(q) = \{p \in D \mid dist(q,p) \leq Eps\}$$

The above stated distance function (such as Euclidean distance, Manhattan distance) often determines the size of the cluster.

Affinity Propagation[12] Clustering algorithm considers all the data points as the potential cluster centers. Using negative Euclidean distance[12] the affinity between two given data points is identified; it is expressed as follows-

$$E(x_i, x_j) = -|x_i - x_j|^2 \text{ where } x_i, x_j \text{ are any two data points.}$$

This algorithm proceeds by considering two measures known as "Responsibilities" which indicates how much a data point is favoured to be a candidate exemplar point over other ones and "Availabilities" indicates how much a candidate exemplar is available for being a cluster center. Rather than taking number of clusters as input, this algorithm takes a parameter "preference" as input which in turn helps in identifying number of exemplar.

III. PROPOSED METHODOLOGY

The proposed method incorporates the concept of topic clustering. The following series of steps are required-

A. Corpus of topics related to Computer Science is prepared.
B. Text pre-processing is applied to that corpus.
C. A weighing vector [15] matrix is prepared with respect to topics within the corpus.
D. After obtaining TF-IDF matrix, several clustering i.e., K-means, DBSCAN, Agglomerative Hierarchical, Spectral clustering, Affinity Propagation algorithms are applied to the matrix.
E. The cluster analysis results are compared with respect to cluster analysis metric.

Algorithm

1) Step 1: Corpus of topics from several domains of Computer Science is build.
2) Step 2: Corpus are then pre-processed by a means of *tokenization*[16], *stopwords removal*[16], and *stemming*[16].
a) Tokenization: It is the process of breaking text into words, phrases, or other meaningful elements called tokens. Consider the following example-

Input- Dog is barking at night Output-[Dog] [is] [barking] [at] [night]

b) Stopwords Removal: It denotes the process of removing common words. Consider the following example-

Input- Dog is barking at night Output- [Dog] [barking] [night]

c) Stemming: It is the process of identifying root word of a given word, for example, 'connect' is the root word of 'connection'.
3) Step 3: Now pre-processed text document is represented as a binary vector, i.e. considering the presence or absence of word in the document or superior representations which encompass weighting methods such as TF-IDF can be used. TF-IDF identifies impact of a word of a document with respect to the collection of documents. It constitutes of two terms- Term frequency (TF) and Inverse Document Frequency (IDF). Term frequency[15] measures number of times a given word w_i appears in a document and can be expressed as follows-

$$TF(w_i) = Count(w_i) / \sum Count(W)$$

where W denotes collection of all words in a document. (2)

Inverse Document Frequency[15] measures the ratio between total number of documents and number of documents containing the word w_i and can be expressed as follows-

$$IDF(w_i) = \log (N / Count(doc(w_i))$$

where N is the total number of documents and $doc(w_i)$ denotes documents containing the word w_i.

(3)

Now, to calculate TF-IDF following formula is used-

$$TF\text{-}IDF(w_i) = TF(w_i) * IDF(w_i)$$

(4)

Using TF-IDF we obtain a TF-IDF matrix for all the words in the corpus in order to apply any clustering methods.

4) *Step 4:* Apply clustering methods such as k-means clustering, Agglomerative Hierarchical Clustering, Spectral Clustering, DBSCAN clustering, Affinity Propagation algorithms on the TF-IDF vectors in order to group them depending on their content similarity.
5) *Step 5:* Next clustering results are analysed and compared with respect to some cluster analysis metric. For analysing the clustering algorithms several metrics are available.

Silhouette Score[17] and Calinski-Harabasz score[17] are chosen as clustering algorithms in the proposed system.
Silhouette Score[17] is a similarity metric that measure how much a data point is close to its neighbour cluster. It proceeds by evaluating correct allocation of data point to a cluster instead of another cluster. This score ranges from -1 to +1, where a value close to +1 denotes data points are associated to the right cluster, whereas -1 denotes data points are wrongly assigned to cluster.
Calinski-Harabasz score [17], or the Variance Ratio Criterion, another evaluation metric, is defined as the ratio between the intra-cluster dispersion and the inter-cluster dispersion.
Finally, the clustering method that has better performance is identified and the recommender system that uses that clustering method is used for providing list of recommendations those are inherent to input choice.

IV. RESULTS

The analysis of result of the Recommender System is based on user's choice and provides other topics inherent to input choice. The following is the result of proposed Recommender System that uses the Spectral Clustering method (that uses affinity as 'rbf') for topic clustering process because it has a better performance than its peer clustering algorithm as indicated in performance comparison chart. Table 1 shows Comparison Chart.

Clustering Method	Number of Clusters	Silhouette Score
K-means	9	0.12093783498979237
Agglomerative Hierarchical	9	0.16136460898744928
Spectral	7	0.16136460898744928
DBSCAN	9	0.11208056630208443

Table 1 shows Comparison Chart

```
Enter Your Subject ::Cloud Computing
Other Recommended Fields is/are-
================================
        Recommendation   1   ::  Green Cloud Computing

        Recommendation   2   ::  Cloud Security

        Recommendation   3   ::  Cloud Analytics

        Recommendation   4   ::  Mobile Cloud Computing

        Recommendation   5   ::  Cloud cryptography

        Recommendation   6   ::  Edge Computing
```

The following is a comparison chart for different cluster techniques such as K-means, Agglomerative Hierarchical, Spectral(RBF), Spectral(NN), DBSCAN, Affinity Propagation, with respect to Silhouette Score and Calinski-Harabasz score for evaluating clustering performances. Figure 1 and Figure 2 show the results.

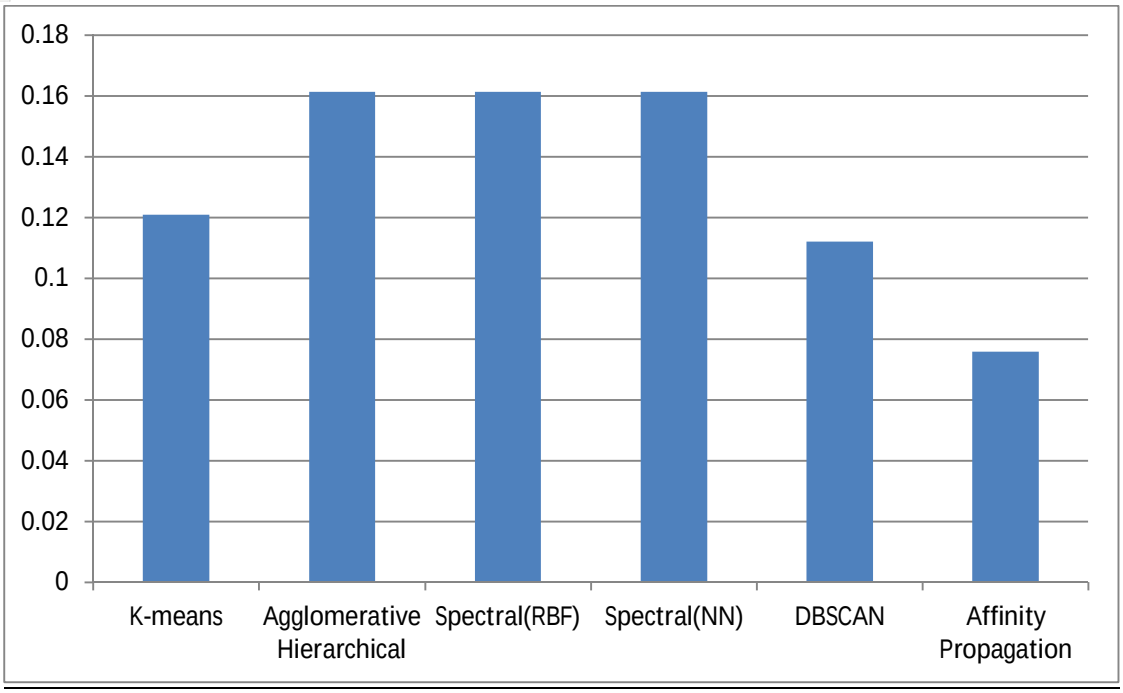


Fig.1. Comparison of clustering results with respect to Silhouette Score

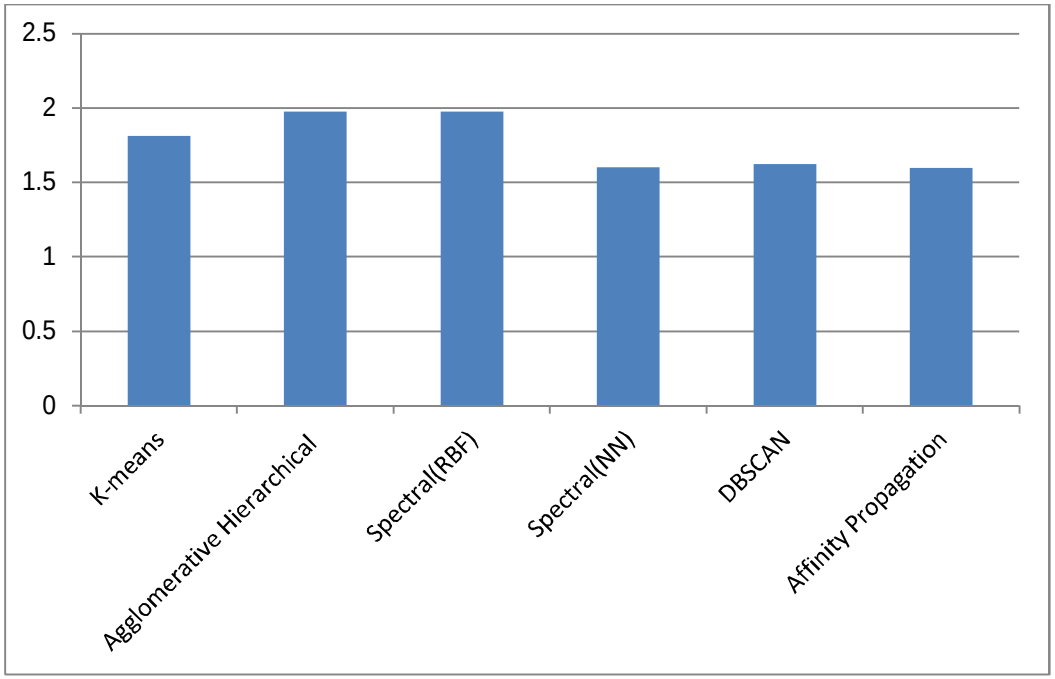


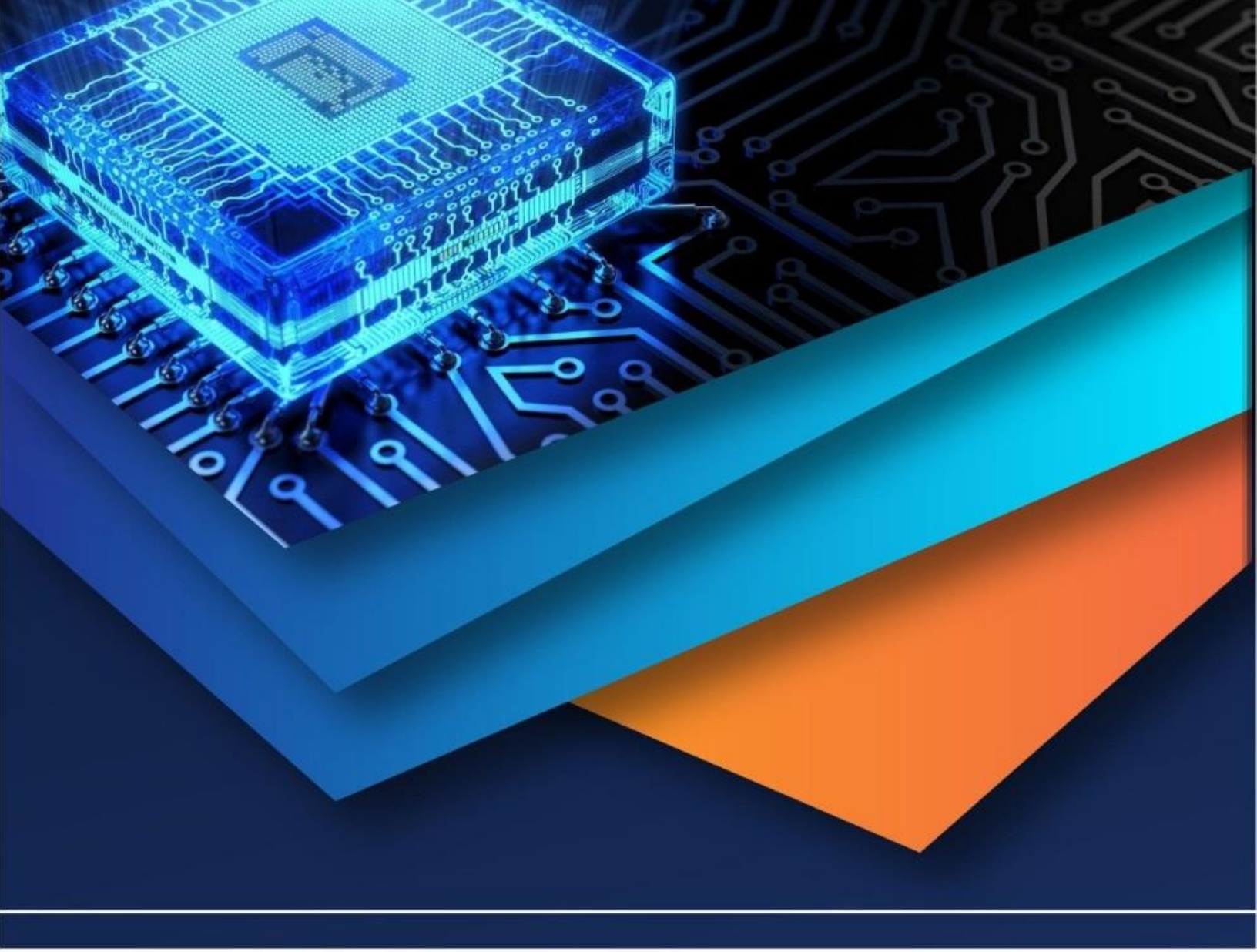
Fig.2. Comparison of clustering results with respect to Calinski-Harabasz score

V. CONCLUSIONS

This paper attempts to facilitate research scholars so that they are guided in a structured way while pursuing their research in the field of Computer Science. To fulfil the purpose, a recommender system is proposed that incorporates the concept of topic clustering which is applied to the corpus of topics on the Computer Science field. For performing topic clustering, this paper has chosen several existing clustering methods and the results are analysed and compared in order to indicate which clustering has the better impact on this topic clustering. After evaluation, the recommender system employs the best clustering method in order to provide recommendations those are inherent to the input choice.

REFERENCES

[1] M. Allahyari et al., "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," 2017.
[2] L. Rokach, "Data Mining and Knowledge Discovery Handbook," Data Min. Knowl. Discov. Handb., 2010.
[3] X. Li, J. Yan, W. Fan, N. Liu, S. Yan, and Z. Chen, "An online blog reading system by topic clustering and personalized ranking," ACM Trans. Internet Technol., vol. 9, no. 3, 2009.
[4] T. Gao, J. Du, S. Wang, and L. Chen, "Topic detection for emergency events based on FCM document clustering," Proc. - 2010 3rd IEEE Int. Conf. Broadband Netw. Multimed. Technol. IC-BNMT2010, pp. 1181–1185, 2010.
[5] Institute of Electrical and Electronics Engineers., IEEE Computer Society. Malaysia Chapter., IEEE Malaysia Section., Universiti Kebangsaan Malaysia. Center for Artificial Intelligence Technology., and Universiti Kebangsaan Malaysia. Data Mining and Optimization Research Group., 2011 3rd Conference on Data Mining and Optimization (DMO) : 28-29 June 2011, Putrajaya, Malaysia. IEEE, 2011.
[6] C. K. Yau, A. Porter, N. Newman, and A. Suominen, "Clustering scientific documents with topic modeling," Scientometrics, vol. 100, no. 3, pp. 767–786, 2014.
[7] P. Natarajan, R. Prasad, K. Subramanian, S. Saleem, F. Choi, and R. Schwartz, "Finding structure in noisy text: Topic classification and unsupervised clustering," in International Journal on Document Analysis and Recognition, 2007, vol. 10, no. 3–4, pp. 187–198.
[8] X. Zhang and Z. Li, "Automatic topic detection with an incremental clustering algorithm," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6318 LNCS, no. M4D, pp. 344–351, 2010.
[9] Y. Li and H. Wu, "A Clustering Method Based on K-Means Algorithm," Phys. Procedia, vol. 25, pp. 1104–1109, 2012, doi: 10.1016/j.phpro.2012.03.206.
[10] F. Murtagh and P. Contreras, "International Encyclopedia of Statistical Science," Int. Encycl. Stat. Sci., no. May 2014, 2011.
[11] M. Meilă, "Spectral Clustering: a Tutorial for the 2010's," Handb. Clust. Anal., p. 753, 2015.
[12] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," Science (80-.)., vol. 315, no. 5814, pp. 972–976, 2007.
[13] N. Krislock and H. Wolkowicz, "Euclidean distance matrices and applications," Int. Ser. Oper. Res. Manag. Sci., vol. 166, pp. 879–914, 2012, doi: 10.1007/978-1-4614-0769-0-30.
[14] M. Ester, H. Kriegel, X. Xu, and D.- Miinchen, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," 1996.
[15] J. Chen, C. Chen, and Y. Liang, "Optimized TF-IDF Algorithm with the Adaptive Weight of Position of Word," vol. 133, pp. 114–117, 2016.
[16] C. Paper, "Preprocessing Techniques for Text Mining Preprocessing Techniques for Text Mining," vol. 5, no. October 2014, pp. 7–16, 2016.
[17] N. Tomašev and M. Radovanović, "Clustering evaluation in high-dimensional data," Unsupervised Learn. Algorithms, pp. 71–107, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY