Comparative Study on Context-Based Document Clustering

Soumen Swarnakar
Department of Information Technology
Netaji Subhas Engineering College,
Techno city, Garia, Kolkata-152,
West Bengal, India.

Shubhalakshmi Ray
Department of Information Technology
Netaji Subhas Engineering College,
Techno city, Garia, Kolkata-152,
West Bengal, India

Tithi Mitra Chowdhury
Department of Information Technology
Netaji Subhas Engineering College,
Techno city, Garia, Kolkata-152,
West Bengal, India.

Abstract- Clustering is an automatic learning technique aimed at grouping a set of objects into subsets or clusters. Objects in the same cluster should be as similar as possible, whereas objects in one cluster should be as dissimilar as possible from objects in the other clusters. Document clustering has become an increasingly important task in analysing huge documents. The challenging aspect to analyse the enormous documents is to organise them in such a way that facilitates better search and knowledge extraction without introducing extra cost and complexity. Document clustering has played an important role in many fields like information retrieval and data mining. In this paper, first Document Clustering has been proposed using Hierarchical Agglomerative Clustering and K-Means Clustering Algorithm.Here, the approach is purely based on the frequency count of the terms present in the documents where context of the documents are totally ignored. Therefore, the method is modified by incorporating Relatedness to measure the degree of relevance of the terms with respect to the concepts present in the documents. Thus, this Clustering is not only Term based but also understanding based (ie, Context Dependent). Next, the clustering is done by Hierarchical Agglomerative Clustering and K-Means with the Relatedness concept. Davies-Bouldin's (DB) Index, which is a well-known metric, has been used to compare the quality of clusters-as they are obtained when the concept of Relatedness is not incorporated in the above mentioned document-clustering algorithms and secondly, when relatedness is integrated into the algorithms.

Keywords – DB Index, clustering, document clustering, Relatedness, context

I.INTRODUCTION

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Thus, the goal of a good document clustering scheme is to minimize intra-cluster distances between documents, while maximizing inter-cluster distances (using a distance measure between documents).Clustering of text documents has great applications in search engines, web mining, information retrieval etc. Ontology based document clustering has been proposed by Swarnakar (2012) while Wang et al. (2001) proposed Hierarchical classification of real life documents. Steinbach et al. (2000) have done a comparative study on document clustering.

II.PROPOSED WORK

In this paper, focus has been given on document clustering by K-Means and Hierarchical Agglomerative Clustering algorithm and a comparison has been done between the mentioned algorithms with the use of DB-Index. At first all the text documents are converted in to lower case. Next, preprocessing is done by removing articles, prepositions, conjunctions and finally stemming operations on different words are applied, i.e., if word in document is *injured*, then after stemming the word would be *injure*. Next, according to concept dictionary and synonyms dictionary, occurrence matrix has been computed and next similarity matrix and then accordingly the clustering algorithms have been used on similarity matrix.

III.EXPERIMENT AND RESULT

A. Terminologies

A few terminologies are defined below for ready reference to the paper.

- *Concept:* Concepts in the Dictionary should be close to objects (physical or logical) and there exists relationships in between concepts and objects. Therefore, different concepts and related objects represent a domain. For Example:

Concepts	Objects
Medical	**Doctor, patient, nurse, operation, medicine, x-ray**
Education	**Student, teacher, class, book, library, institution, copy, Laboratory**

- *Dictionary*: For Document Clustering, a Concept Dictionary is maintained, where different related concepts and their corresponding objects are stored for each domain of interest. Naturally, various objects like computer, CPU, server, networking etc. describe the Concept Computer Science. In these Document Clustering algorithms, a Synonym dictionary is maintained, which stores synonyms of objects.
- *Relatedness vector*: The components of Relatedness vector of objects represent degree of relation of the objects with respect to a particular concept. Suppose student is an object related both to education concept and medical concept. An expert can supply Relatedness of student of education concept as 0.9 but that of medical concept as 0.4, because student is more likely related to education than medical concept. The relatedness of student in transport concept is 0.1, because it is less related.
- *Relatedness Matrix:* Relatedness vectors with respect to different concepts are combined and the Relatedness matrix is created.
- *Concept Matrix:* Each element of concept represents no. of objects associated with a particular concept present in a particular document. Suppose in document1 the total number of objects with their synonyms of medical concept, transport concept and education concept are 9, 8 and 5 respectively. Similarly, in document2 the total number of objects with synonyms of medical concept, transport concept and education concept are 1, 3 and 4 respectively. The ***Concept matrix*** is as follows:

	Medical	Transport	Education
Document 1	**9**	**8**	**5**
Document 2	**1**	**3**	**4**

B. Documents used for the Clustering process

We have considered the following ten Documents for the Clustering Process:-

- Doc1.txt: Sheila is a nurse. She works at the hospital. She specializes in providing drugs to patients. She takes the bus every day. She abides by traffic-rules to avoid accidents.

- Doc 2.txt: Ravi is a good student. His teachers are proud of his grades. He completes his homework in time, in his workbooks .He wants to be a doctor when he grows up. He wants to discover a drug for cancer.

- Doc 3.txt: The market caught on fire. Firemen rushed to the scene on trucks. They had water and hoses. They also used extinguishers to put out the flames. Burns were treated with medicine. The rest were taken to the hospital and given oxygen.

- Doc 4.txt: Dilip is a bus-conductor. He spends most of his time on the road. He tries to earn enough salary so his daughter can go to school. He often has help from the principal.

- Doc 5.txt: Many software engineers are self-employed. They make their own websites and earn profits. They acquired these skills from school and college days .They must have had very good teachers.

- Doc 6.txt: Rani loves books. She has access to her favorite books at the school library. She goes to school by schoolbus. The principal greets the students at the gate.

- Doc 7.txt: Vimal bought a motorbike. He is a reckless driver. He does not pay heed to traffic-rules. If he is not careful he will meet with an accident and end up in the ER. He did not listen to his teachers at driving school.

- Doc 8.txt: Parimal is an accountant. He handles finances for a big company. He took up commerce in school. He was the best student in his class. He is satisfied with his salary.

- Doc 9.txt: Pritha is not well. She is down with fever and headache. She is taking medicine. She has skipped work for a week. Her boss is not happy about this.

- Doc 10.txt: Mr. Lucas is the principal of Dsouza College. His work is to manage all employees. He reaches college by driving his cars.

C. Creation of Dictionary

The ***Concept Dictionary*** and ***Synonyms Dictionary*** which have been considered during the course of the paper are as follows-

Concepts	**Objects**		
Transport	**Vehicle**	**t.person**	**t.others**
Education	**e.person**	**e.place**	**knowledge**
Medicine	**Med.people**	**Disease**	**care**

Table 1: Concept Dictionary

Table 2: Synonyms Dictionary

Objects	Synonyms
vehicle	car, bus, truck, train, ambulance, motorbike
t.person	driver, traffic-police, bus-conductor, mechanic
t.others	signal, pavement, road, traffic-rule, drive
e.person	principal, dean, director, student, teacher, graduate
e.place	school, college, library, laboratory, gym
knowledge	commerce, science, grade, homework, training, subject, knowledge, book, textbook
med.people	doctor, nurse, patient, paramedic, medical, care-giver
disease	cancer, accident, burn, fever, headache, influenza, injure
care	hospital, OT, nursing-home, ER, ambulance, drugs, syrup, oxygen, burnol, medicine

D. Formation of Occurrence Matrix

When Relatedness is not considered- The Occurrence Matrix for the given set of Documents when Relatedness is not Considered is shown:-

	vehicle	t.person	t.others	e.person	e.place	knowledge	med. people	disease	care
Doc 1	2	0	1	0	0	0	2	2	1
Doc 2	0	0	0	2	3	0	1	1	1
Doc 3	2	0	0	0	0	0	0	4	2
Doc 4	0	1	1	1	0	1	0	0	0
Doc 5	0	1	5	1	1	1	0	0	1
Doc 6	1	0	0	1	2	4	1	0	0
Doc 7	1	1	2	1	0	1	0	1	1
Doc 8	1	0	2	0	4	1	0	0	0
Doc 9	0	0	0	0	1	1	0	1	3
Doc 10	0	0	0	3	3	0	3	1	0

Table 3: Occurrence Matrix when Relatedness is not considered

When Relatedness is considered- Relatedness of objects associated with the concepts may vary from expert to expert. The Concept Matrix for the given set of Documents is shown below:-

	Transport	Education	Medicine
Doc 1	3	0	5
Doc 2	0	5	3
Doc 3	2	0	6
Doc 4	2	2	0
Doc 5	6	3	1
Doc 6	1	7	1
Doc 7	4	2	2
Doc 8	3	5	0
Doc 9	0	2	4
Doc 10	0	6	4

Table 4: Concept Matrix related to documents

The Occurrence Matrix that is obtained by incorporating Relatedness for the given set of Documents is shown below (Table 5). We consider that the Relatedness of the highest relevance of an object with respect to a document is 0.9, the Relatedness of medium relevance of an object with respect to a document as 0.125 and that of the lowest relevance as 0.01.

	vehicle	t.person	t.others	e.person	e.place	knowledge	med. people	disease	care
Doc 1	**0.25**	**0**	**0.125**	**0**	**0**	**0**	**1.8**	**1.8**	**0.9**
Doc 2	**0**	**0**	**0**	**1.8**	**2.7**	**0**	**0.125**	**0.125**	**0.125**
Doc 3	**0.25**	**0**	**0**	**0**	**0**	**0**	**0**	**3.6**	**1.8**
Doc 4	**0**	**0.9**	**0.9**	**0.125**	**0**	**0.125**	**0**	**0**	**0**
Doc 5	**0**	**0.9**	**4.5**	**0.125**	**0.125**	**0.125**	**0**	**0**	**0.01**
Doc 6	**0.125**	**0**	**0**	**0.9**	**1.8**	**3.6**	**0.125**	**0**	**0**
Doc 7	**0.9**	**0.9**	**1.8**	**0.125**	**0**	**0.125**	**0**	**0.125**	**0.125**
Doc 8	**0.125**	**0**	**0.25**	**0**	**3.6**	**0.9**	**0**	**0**	**0**
Doc 9	**0**	**0**	**0**	**0**	**0.125**	**0.125**	**0**	**0.9**	**2.7**
Doc 10	**0**	**0**	**0**	**2.7**	**2.7**	**0**	**0.375**	**0.125**	**0**

Table 5: Occurrence Matrix when Relatedness is considered: documents vs. Relatedness of objects

E. DB Index Values for the clusters formed using K-Means Clustering Algorithm when Relatedness is not considered

No. of Clusters	**Documents in each cluster**								
2 clusters	**Docs 1,3,4,5,7,9**	**Docs 2,6,8,10**							
3 clusters	**Docs 1,4,5,7**	**Docs 2,6,8,9,10**	**Doc 3**						
4 clusters	**Docs 4,5,6,7,9**	**Docs 2,8,10**	**Docs 1**	**Doc 3**					
5 clusters	**Docs 4,6,7,9**	**Docs 2,8,10**	**Doc 1**	**Doc 3**	**Doc 5**				
6 clusters	**Docs 4,7,9**	**Docs 2,,8,10**	**Doc 1**	**Doc 3**	**Doc 5**	**Doc 6**			
7 clusters	**Docs 7,9**	**Docs 2,8,10**	**Doc 3**	**Doc 4**	**Doc 1**	**Doc 5**	**Doc 6**		
8 clusters	**Doc 2,10**	**Doc 7,9**	**Doc 4**	**Doc 5**	**Doc 6**	**Doc 8**	**Doc 1**	**Doc 3**	
9 clusters	**Doc 4,7**	**Doc 2**	**Doc 3**	**Doc 4**	**Doc 5**	**Doc 6**	**Doc 8**	**Doc 9**	**Doc 10**

Table 6: Clusters formed using K-Means Clustering Algorithm when Relatedness is not considered

No. Of Clusters	2	3	4	5	6	7	8	9
DB Index	0.53	***0.38***	0.40	0.62	0.73	0.76	0.72	0.86

Table 7: DB Index Values for different number of clusters formed using K-Means Clustering Algorithm when Relatedness is not considered

Analysis of Results- Looking at Table 7, it is seen that minimum DB Index corresponds to 3 clusters. Documents 1,4,5,7 have been merged into one cluster, followed by Documents 2,6,8,9,10 into another and finally Document 3 which is present as the sole member in the 3rd Cluster(as evident from Table 6).However, on examining closely, it is observed that even though in Documents 4,5,7, "Transport" happens to be the dominant Concept and Document 1 is more of "Medical" nature which can be concluded by the very essence of the text in the document, Document 1 is present in the same cluster as Documents 4,5,7,which is clearly erroneous. This is due to the fact that Document 1 contains certain "Transport" terms in it which causes it to belong to a cluster where "Transport" Concept is dominant.This is an example of misclustering with respect to the considered documents. The same can be said for the 2nd cluster where even though Document 9 is more of "Medical"nature,it has been merged into a cluster along

with Documents 2,6,8,10(where "Education"Concept is dominant) simply because of the presence of certain "Education"terms in it.

F.DB Index Values for the clusters formed using Single Linkage Hierarchical Agglomerative Clustering Algorithm when Relatedness is not considered

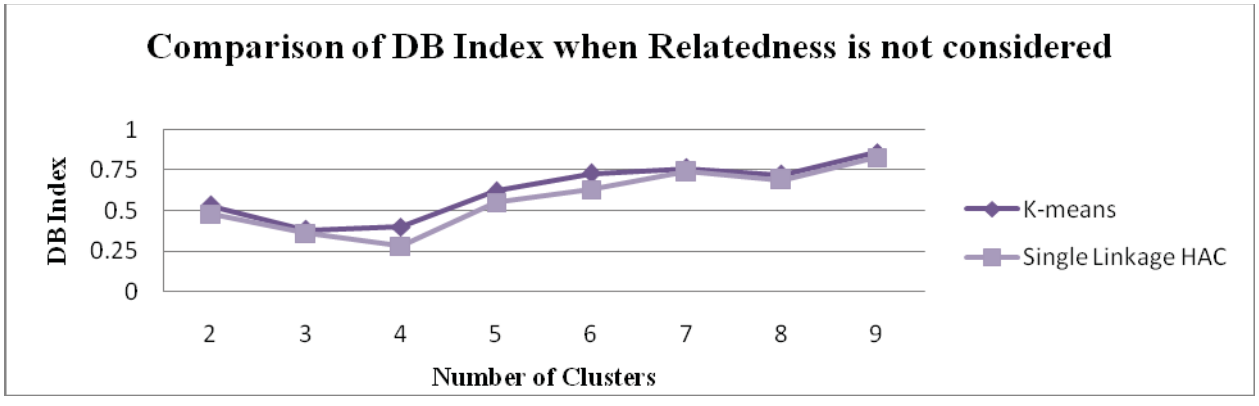
No of Clusters	**Documents in each cluster**								
2 clusters	**Docs 1,2,3,4,5,7,8,9,10**	**Doc 6**							
3 clusters	**Docs 1,3,4,5,7,9**	**Docs 2,8,10**	**Doc 6**						
4 clusters	**Docs 1,3,4,5,7,9**	**Docs 2,10**	**Doc 6**	**Doc 8**					
5 clusters	**Docs 1,3,4,5,7**	**Docs 2,10**	**Doc 6**	**Doc 8**	**Doc 9**				
6 clusters	**Docs 1,3,4,7**	**Docs 2,10**	**Doc 5**	**Doc 6**	**Doc 8**	**Doc 9**			
7 clusters	**Docs 4,7**	**Docs 2,10**	**Doc 1,3**	**Doc 9**	**Doc 1**	**Doc 5**	**Doc 6**		
8 clusters	**Docs 2,10**	**Docs 4,7**	**Doc 1**	**Doc 5**	**Doc 6**	**Doc 8**	**Doc 9**	**Doc 3**	
9 clusters	**Docs 4,7**	**Doc 2**	**Doc 3**	**Doc 4**	**Doc 5**	**Doc 6**	**Doc 8**	**Doc 9**	**Doc 10**

Table 8: Clusters formed using Single Linkage Hierarchical Clustering Algorithm when Relatedness is not considered

No. Of Clusters	2	3	4	5	6	7	8	9
DB Index	0.48	0.36	***0.28***	0.55	0.63	0.74	0.69	0.83

Table 9: DB Index Values for different number of clusters formed using Single Linkage Hierarchical Clustering Algorithm when Relatedness is not considered

Analysis Of Results- Similarly, looking at Table 9,it is seen that minimum DB Index corresponds to 4 clusters. Documents 1,3,4,5,7,9 have been merged into one cluster, followed by Documents 2,10 into another and finally Documents 6 and 8 which are present as the sole members in the 3rd and 4th Clusters respectively.(as evident from Table 8). However, on examining closely, it is observed that even though in Documents 4,5,7,"Transport "happens to be the dominant Concept and Documents 1,3,9 happen to be more of "Medical" nature which can be concluded by the very essence of the text in the documents, Documents 1,3,9 are present in the same cluster as Documents 4,5,7,which is clearly erroneous. This is due to the fact that Documents 1, 3, 9 contain certain "Transport" terms in them which causes them to belong to a cluster where "Transport" Concept is dominant.This is an example of misclustering with respect to the considered documents. Another interesting observation is that the DB Index Values obtained for Single Linkage HAC are comparatively lower than those for K-Means, which makes a better clustering than K-Means (on comparing Table 7 and 9). This is due to the fact that small value of DB Index indicates compact and well-separated clusters.



Graph 1: Comparison of DB Index Values for different number of clusters formed using Single Linkage Hierarchical Agglomerative Clustering and K-Means Clustering Algorithm when Relatedness is not considered

G. DB Index Values for the clusters formed using K-Means Clustering Algorithm when Relatedness is considered

Considering the highest relevance of an object with respect to a document as 0.9, that of medium relevance as 0.125 and the lowest relevance as 0.01, we obtain the following clusters:-

No of Clusters	Documents in each cluster								
2 clusters	Docs 1,3,4,5,7,9	Docs 2,6,8,10							
3 clusters	Docs 1,3,9	Docs 2,6,8,10	Docs4,5,7						
4 clusters	Docs 1,9	Docs 2,6,8,10	Docs 3	Doc 4,5,7					
5 clusters	Docs 1,9	Docs 2,6,8,10	Doc 3	Doc 4,7	Doc 5				
6 clusters	Docs 1,9	Docs 2,8,10	Doc 4,7	Doc 3	Doc 5	Doc 6			
7 clusters	Docs 1,9	Docs 2,8,10	Doc 3	Doc 4	Doc 7	Doc 5	Doc 6		
8 clusters	Doc 2,10	Doc 1,9	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8	Doc 3	
9 clusters	Doc 2,10	Doc 1	Doc 3	Doc 4	Doc 5	Doc 6	Doc 8	Doc 9	Doc 7

Table 10: Clusters formed using K-Means Clustering Algorithm when Relatedness is considered

No. Of Clusters	2	3	4	5	6	7	8	9
DB Index	0.5	***0.33***	0.39	0.58	0.72	0.71	0.69	0.82

Table 11: DB Index Values for different number of clusters formed using K-Means Clustering Algorithm when Relatedness is considered

Analysis Of Results- Looking at Table 11, it is seen that minimum DB Index corresponds to 3 clusters. Documents 1,4,5,7 have been merged into one cluster, followed by Documents 2,6,8,10 into another and finally Documents 1,3,9 in the 3rd Cluster(as evident from Table 10).Here, it is observed that perfect Clustering has taken place. Documents which are of "Transport", "Education" and "Medical" nature have been grouped accordingly into 3 separate Clusters. Thus, it is seen that by incorporating Relatedness into the above Algorithm, misclustering has been prevented. Again, on comparing Tables 7and 11, it is observed that the DB Index Values for the respective number of Clusters have decreased when Relatedness has been considered, which indicates better quality of Clusters in this case.

H. DB Index Values for the clusters formed using Single Linkage Hierarchical Agglomerative Clustering Algorithm when Relatedness is considered

Considering the highest relevance of an object with respect to a document as 0.9, that of medium relevance as 0.125 and the lowest relevance as 0.01, we obtain the following clusters:-

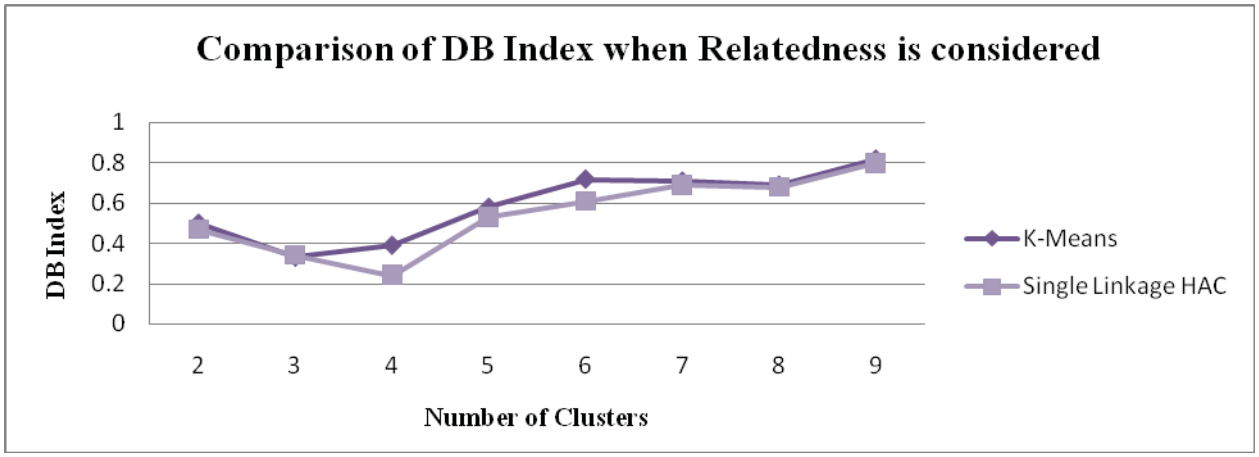
No of Clusters	Documents in each cluster								
2 clusters	Docs 1,3,4,5,7,9	Doc 2, 6,8,10							
3 clusters	Docs 1,3, 9	Docs 2,6,8,10	Doc 4,7,5						
4 clusters	Docs 1,3, 9	Docs 2,8,10	Doc 4,5,7	Doc 6					
5 clusters	Docs 1,3,9	Docs 2,8,10	Doc 4,7	Doc 5	Doc 6				
6 clusters	Docs 1,3	Docs 2,8,10	Doc 4,7	Doc 6	Doc 5	Doc 9			
7 clusters	Docs 4,7	Docs 2,8,10	Doc 1	Doc 9	Doc 3	Doc 5	Doc 6		
8 clusters	Docs 2,10	Docs 4,7	Doc 1	Doc 5	Doc 6	Doc 8	Doc 9	Doc 3	
9 clusters	Docs 2,10	Doc 4	Doc 3	Doc 4	Doc 5	Doc 6	Doc 8	Doc 9	Doc 7

Table 12: Clusters formed using Single Linkage Hierarchical Clustering Algorithm when Relatedness is considered

No. Of Clusters	2	3	4	5	6	7	8	9
DB Index	0.47	0.34	***0.24***	0.53	0.61	0.69	0.68	0.80

Table 13: DB Index Values for different number of clusters formed using Single Linkage Hierarchical Clustering Algorithm when Relatedness is considered

Analysis Of Results- Looking at Table 13, it is seen that minimum DB Index corresponds to 4clusters. Documents 1,3,9 have been merged into one cluster, followed by Documents 2,8,10 into another, Documents 4,5,7 in the 3rd Cluster and finally Document 6 as the sole member in the 4th Cluster(as evident from Table 12).Here, it is observed that better Clustering has taken place. Documents which are of "Transport", "Education" and "Medical" nature have been grouped accordingly into 3 separate Clusters while Document 6 has been grouped into a separate one since it is a document having a mixture of "Transport", "Medical" and "Education" terms. Thus,it is seen that by incorporating relatedness into the above algorithm , misclustering has been prevented upto a certain extent. Again, on comparing Tables 9 and 13, it is observed that the DB Index Values for the respective number of Clusters have decreased when Relatedness has been considered, which indicates better quality of Clusters in this case. Another interesting observation is that the DB Index Values obtained for Single Linkage HAC (considering Relatedness) are comparatively lower than those for K-Means (considering Relatedness) which reinforces the fact that gives better clustering than K-Means algorithm (on comparing Tables 11 and 13). This is due to the fact that small values of DB Indices are indicative of the presence of compact and well-separated clusters.



Graph 2: Comparison of DB Index Values for different number of clusters formed using Single Linkage Hierarchical Agglomerative Clustering and K-Means Clustering Algorithm when Relatedness is considered

IV. CONCLUSION

From the results, it can be concluded that without Relatedness Vector, misclustering is happening because several documents are clustered along with documents of a different concept just because of the presence of a few words in the document which is responsible for misclustering .When Relatedness Vector is used, the relatedness assigned to each word with relevance to each concept of the document allows each document to be clustered into a set with documents of highest similarity.

From the results obtained, it can also be concluded that Single Linkage Hierarchical Clustering is giving better clustering results than K-Means algorithm. This is evident from the fact that the cluster validation DB Index values are less in case of Single Linkage HAC in both cases where Relatedness Vector is and is not considered .

REFERENCES

[1] S.Swarnakar (2012) "Ontology-based context dependent document clustering method", Int. J. Knowledge Engineering and Data Mining, Vol. 2, No. 1, pp.35–59.
[2] F. Murtagh, "A Survey of Recent Advances in Hierarchical Clustering Algorithms," The Computer Journal, 26(4): 354-359, 1983.
[3] G. Salton, and M. J. McGill, "Introduction to Modern Information Retrieval," McGraw-Hill Inc., 1983.
[4] K. Wang, S. Zhou and Y. He, "Hierarchical classification of real life documents," SIAM International Conference on Data Mining, SDM'01, Chicago, United States, April, 2001.

[5] Michael Steinbach, George Karypis, and Vipin Kumar "A Comparison of Document Clustering Techniques" Department of Computer Science and Engineering, University of Minnesota., 2000.
[6] Anton V. Leouski and W. Bruce Croft "An Evaluation of techniques for clustering search results" Computer Science Department, University of Massachusetts at Amherst. (1996)
[7] Brian S. Everitt, Sabine Landau, and Morven Leese "Cluster Analysis" Oxford University Press, fourth edition, 2001.