Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.8 No.2, 2019, pp. 116-121

Clustering Based Approach for Novelty Detection in Text Documents

Sushil Kumar[1] and Komal Kumar Bhatia[2]
[1]Assistant Professor, [2]Professor, [1&2]Department of Computer Engineering,
J.C. Bose University of Science and Technology, YMCA Faridabad, Haryana, India
E-Mail: panwar_sushil2k@yahoo.co.in, komalbhatia1@rediffmail.com

***Abstract* - As the information is overloaded over the internet accessing of information from the internet according to a given query provides redundant and irrelevant information. It is necessary to retrieve relevant and novel information from a given query by the user. With the result of this the user will require minimum effort to access the information need. In this work we proposed a clustering based approach for novelty detection which will provide the relevant and novel documents for the information need. Based on the user query the incoming stream of documents will be clustered using k-means algorithm. Then the cluster heads are selected from the various clusters with the minimum distance. These cluster heads are the novel documents from a collection of documents from different clusters having the large distance. The proposed technique can be further used in the field of information retrieval.**

Keywords: **Novelty Detection, Information Retrieval, Clustering, Cluster Head, Jupyter Note-Book Python**

I. INTRODUCTION

On-line information with the advent of World Wide Web has revolutionized the approach of accessing the information. As the digitized information is available on the internet and the volume of the information is also changing day by day which results in information overload. It is very difficult for the user to retrieve relevant information according to the query provided by the user. Because the search engine provides a large list of documents for a given query and the user needs to explore the whole list for relevant document, which is a very time-consuming task. Novelty detection [10] is a technique used to retrieve relevant and novel information according to the user query with less effort. Novelty detection can be viewed as going a step further than traditional document retrieval. Based on the output of a document retrieval system (i.e., a ranked list of documents), a novelty detection system will further extract documents with new information from the ranked list. The purpose of the research on novelty detection is to provide a user with a list of text documents that both are relevant and contain new information with respect to the referenced information already seen. If there are two documents both contains one sentence based on a query 'road accident in front of YMCA' as below:

D1: Road accident between car and motor bike in front of YMCA.
D2: Police came after road accident between car and motor bike in front of YMCA.

So the document D2 contains new information and is regarded as relevant and novel document

II. RELATED WORK

The Previous work done on novelty detection in text documents as below

1. Topic Detection and Tracking (TDT) research and evaluation project has been presented by Allan, Wade and Bolivar [3] that is dedicated to online event detection and tracking. This work interested in inter event novelty detection in order to determine whether the two news stories covers the same occasion and based on online story level evaluation.
2. Novelty detection can be performed at three different levels: the event level by T. Brant and Frants [2], the document level and the sentence level [8, 10]. At the event level by J. Carthy [5, 9, 11], a novel document is required to not only be relevant to a topic (i.e., a query) but also to discuss a new event.
3. At the sentence level by Le and craft [12], a novel sentence should be relevant to a topic and provide new information. This means that the novel sentence may either discuss a new event or provide new information about an old event. Sentence level novelty detection [10] is also the basis for the event level novelty detection.
4. Document to sentence level technique for novelty detection by S. Kumar and K.K Bhatia [14] in 2015.

III. PROPOSED METHODOLOGY

In this work we have proposed a clustering [1] based approach for novelty detection which will provide the relevant and novel information to the user query. Firstly the incoming stream of documents for the user query related to a domain has been clustered using k-means algorithm [4, 6, 15]. User can make a query based on specific domain using search engine and the first thirty retrieved results scarped out and store in a file on the disk. These documents are used to make ten clusters each containing of similar documents. Based on these clusters one cluster head is selected from each cluster which will provide ten documents. All of these ten documents having large distance as compared with each other and all will be relevant and novel. The architecture of proposed system is shown in the fig.1

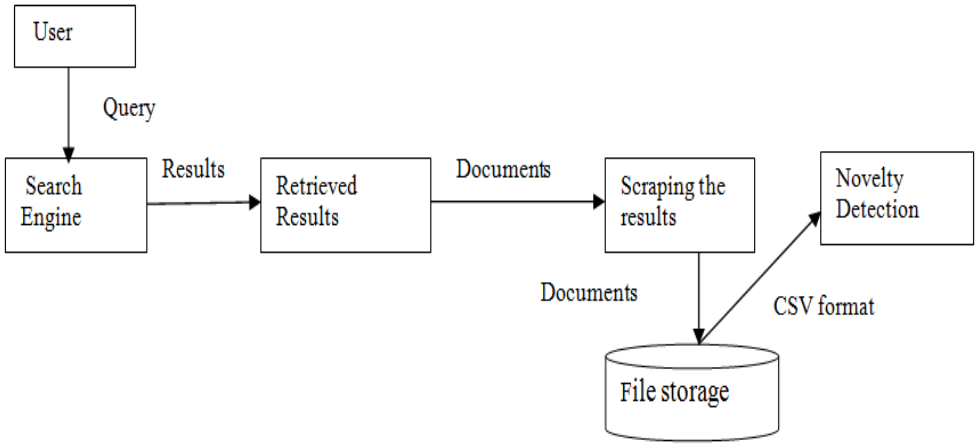

Fig. 1 Architecture of the Proposed System

A. Collection of Textual Datasets: At the beginning one hundred and fifty documents were collected which consists of thirty each of festival (f1, f2……f30), sports (s1, s2…..s30), technology (t1, t2….t30), politics (p1, p2…..p30) and education domains (e1, e2…e30). These documents undergo refinement which is fed to the algorithm to obtain clusters containing documents from similar domains.

B. Convert Documents into Vectors: A document consists of large number of words and it is not always necessary that each word is of importance. Due to which high dimensionality of the document has to be reduced by processing the document to get rid of extra words to obtain weight of each of the word to be used in the algorithm. The conversion of documents into vectors is carried in various steps.

1. Tokenization: The processes involve in information retrieval [7] require the words of documents. Tokenization is used to identify the meaningful words called tokens. The main use of tokenization is to split the sentences into individual tokens. For example 'There are readers who prefer learning' so in this sentence 'there', are, 'readers', 'who', 'prefer' and learning' are the tokens.

2. Stop-Words Removal: The collection of documents contains some unnecessary words due to which dimensionality of document will be increased. Pronoun, adverb, preposition etc. which are used throughout in the document has to be removed to get proper result. For example 'Can listening be exhausting' so after removal of stop word can and be it will results in Listening, exhausting.

3. Novelty Detection Module: This module helps in finding the novelty of the documents. The following document is used to understand the algorithm.

Education is a pillar of development. Educated citizens help in developing a nation. India is a developing nation. Indian education system will help the nation to develop. Various schemes have been launched to motivate the citizens to educate. Due to these schemes literacy rate is increasing. One day when India will reach to its goal of education it will definitely become developed nation. Firstly text documents will be break in to sentences and each sentence is act as a document like D1, D2, D3,…., etc. For preprocessing we tokenize the data, remove stop words, replace tokens by their stems and generate inverse document frequencies vectors on dynamic vector space model. Then the system will pick up relevant documents for a given query and filter out the non-relevant documents in the categorization stage. Finally based on the historical documents, the novelty detection [16, 17] determines whether the input document is novel or not.

The algorithm takes the following steps

Input: X, K where X=Set of classified instances, K= integer

Output: Set of K clusters

Require X≠ Null, K>0
1. Procedure Generate Clusters
2. Initialize K random centriods
3. Repeat
4. for all instance i in X do
5. shortest ← 0
6. membership ← null
7. for all centriod c1 do
8. dist1 ← distance(c1)
9. if dist1 < shortest then
10. shortest ← dist1
11. membership ← c1
12. end if
13. end for
14. end for
15. Recalculate Centriod (c1)
16. Until convergence
17. End of Procedure

Let us assume K=2 and the D5 and D7 are chosen for clusters, Calculate Euclidean distance using the given equation, Distance [(x,y),(a,b)] $= \sqrt{(x-a)2+(y-b)2}$

TABLE I DISTANCE MATRIX WITH TERM FREQUENCY

Terms/ documents	D1	D2	D3	D4	D5	D6	D7
Education	1	1	0	1	1	0	1
Pillar	1	0	0	0	0	0	0
Development	1	1	1	0	0	0	1
Citizen	0	1	0	0	1	0	0
Nation	0	1	1	1	0	0	1
India	0	0	1	1	0	0	1
System	0	0	0	1	0	0	0
Various	0	0	0	0	1	0	0
Schemes	0	0	0	0	1	0	0
Launch	0	0	0	0	1	0	0
Motivate	0	0	0	0	1	0	0
Literacy	0	0	0	0	0	1	0
Increase	0	0	0	0	0	1	0
Reach	0	0	0	0	0	0	1
Goal	0	0	0	0	0	0	1
Definitely	0	0	0	0	0	0	1

Table I represents the distance matrix with term frequency in the given documents.

Euclidean Distance between D1and D5 is calculated as:

$\sqrt{(1-1)^2+(1-0)^2+(1-0)^2+(0-1)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-1)^2+(0-1)^2+(0-1)^2+(0-1)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2}$

$=\sqrt{0+1+1+1+0+0+0+1+1+1+1+0+0+0+0+0} = \sqrt{7} = 2.64$

Similarly, we can calculate the distance with other documents.

Table II represents the movement of documents D1, D2, D3, D4, D7 to cluster D7 and D5, D6 move to cluster D5 based upon the minimum Euclidean distance.

TABLE II DOCUMENTS MOVEMENT TO CLUSTERS

Document	D5 Cluster	D7 Cluster	Minimum distance	Movement to cluster
D1	2.64	2.44	2.44	D7
D2	2.64	2.23	2.23	D7
D3	3	2	2	D7
D4	2.82	2	2	D7
D5	0	3.31	0	D5
D6	2.82	3	2.82	D5
D7	3.31	0	0	D7

Table II shows that how clusters are formed when the set of documents have trained by Jupyter K-means model. The clustering will be dynamic for similarity calculation in the documents. We trained the CSV clean data file by using K-means model and make clusters of documents. Then we find out the cluster head from K-means model in document format from each cluster.

These clusters heads from each cluster having large distance with other cluster head. So the collection of these cluster heads yields the novel documents from a collection of thirty documents.

IV. COMPARISON WITH OTHER APPROACHES

1. Search Engine provides the list of documents that can be relevant and redundant.
2. This approach takes the list provided by the Search Engine after scraping the results.
3. The proposed approach filters out the redundant documents and provides the relevant and novel documents.

V. IMPLEMENTATION AND DISCUSSION

Initially the experiment has been performed on a set of thirty documents related to query from different domains. Thirty documents read from the CSV file which is a clean data. Then we used a TFIDF vectorizer to convert in to TFIDF vectors. We trained this by using K-means model and make clusters based on the similarity between the documents. Then we find out the cluster head from K-means model in document format. . These clusters heads from each cluster having large distance with other cluster head. Thus, the cluster head from each cluster are the novel documents from a collection of thirty documents.

The implementation included Anaconda (Python distribution) Jupyter Notebook Python 3.6 [13] is a free open source distribution of the Python and R-programming languages. We have used this because it provides support for scientific computing i.e data science, machine learning applications, large scale data processing and predictive analytics that aims to simplify package management and deployment. It also includes the necessary library Numpy for numerical calculation, Pandas for opening large file in hard disk, Sklearn for importing the K-means model and nltk library to clean data. The steps that are used to execute the algorithms are below:

Step 1: The screen shot shown below is used to make a query for Bing search engine.

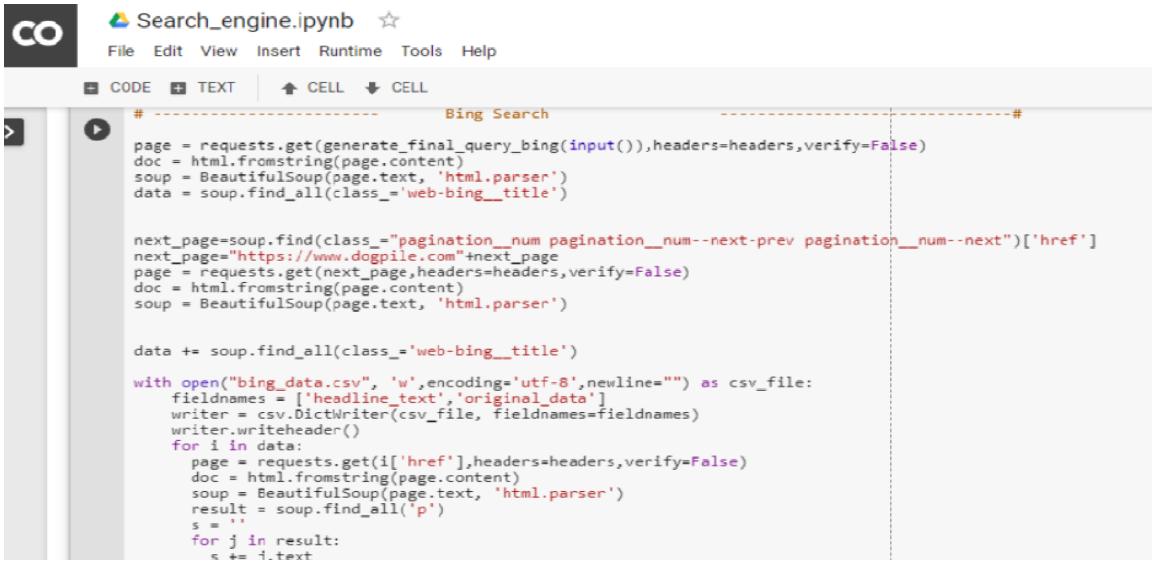

Fig. 2 Bing Search Engine Query Interface

To run this module by pressing on the arrow sign on top left side of the module it will provide the interface to make a query. We can make a query in any of domains i.e festival, politics, entertainment, sports and education. After making the query for '*holi'* related to domain festival the module first scrap the thirty results from Bing search engine by using BeautifulSoup method of Jupyter notebook. Then these thirty documents have been written in the bing_data.csv file with header text data and original data. This file is automatically downloadable on the system.

Step 2: To open the clustering based novelty detection file and perform the necessary refinements with available libraries

Fig. 3 Cleaning of the data and Trained the K-means model

The above screen shot show that this module is trained with k-means algorithm. Library Sklearn is used to train the k-means algorithm. We used Tfidfvetorizer to get the word features and after tokenization the meaningful words called tokens are identified. The main use of tokenization is to split the sentences into individual tokens. For example 'There are readers who prefer learning' so in this sentence 'there', are, 'readers', 'who', 'prefer' and learning' are the tokens.

Then using sklearn library k-means algorithm has imported to make the cluster based on the results on the bing_data CSV file. The different cluster heads from various clusters have been stored in array. The novel results have been generated by selecting the centriod with minimum distance from each cluster. The centriod from each cluster has large distance from the other cluster centriod. So these documents are the novel documents.

VI. RESULT ANALYSIS

The Table III shows that different queries have fired in different domains on the Bing Search Engine Interface. As in the table thirty documents retrieved corresponding to the query 'holi', IPL, Narender Modi and Pulwama attack. Result analysis represents that when we manually compare the novel documents based on result retrieved by Bing search engine and our approach.

TABLE III COMPARISON OF BING SEARCH ENGINE RESULTS WITH PROPOSED APPROACH

Domain for Query	Bing Search Engine Retrieved Documents	Bing Search Engine		Proposed Approach	
		Novel documents out of 30 documents	Novel documents in first ten documents	Novel documents out of 30 documents	Novel documents in first ten documents
Festival (Holi)	30	09	05	09	09
Sports (IPL)	30	07	02	08	08
Politics (Narendra Modi)	30	08	05	09	09
Police force (Pulwama Attack)	30	09	05	09	09

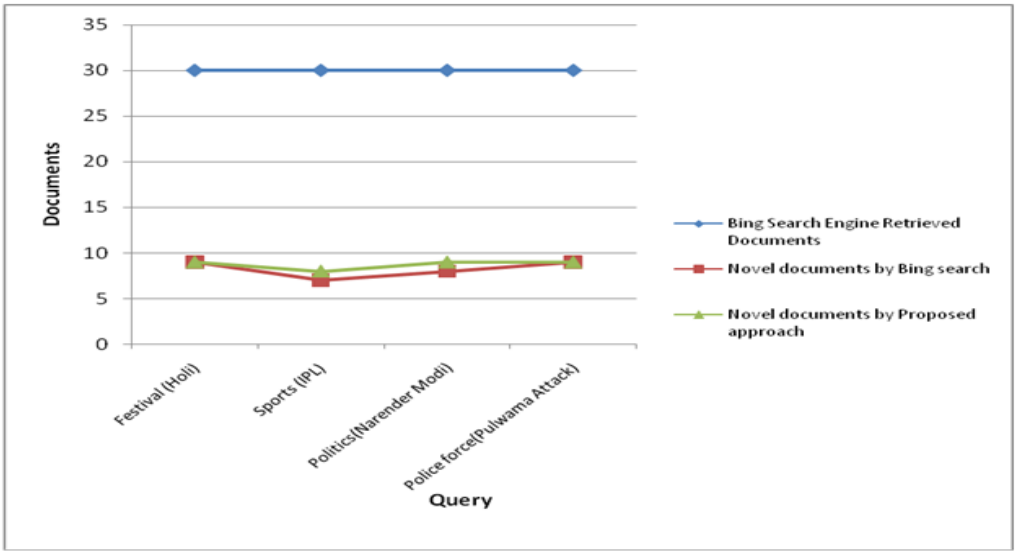

Fig. 4 Comparison of Bing search engine results with proposed approach with in 30 documents

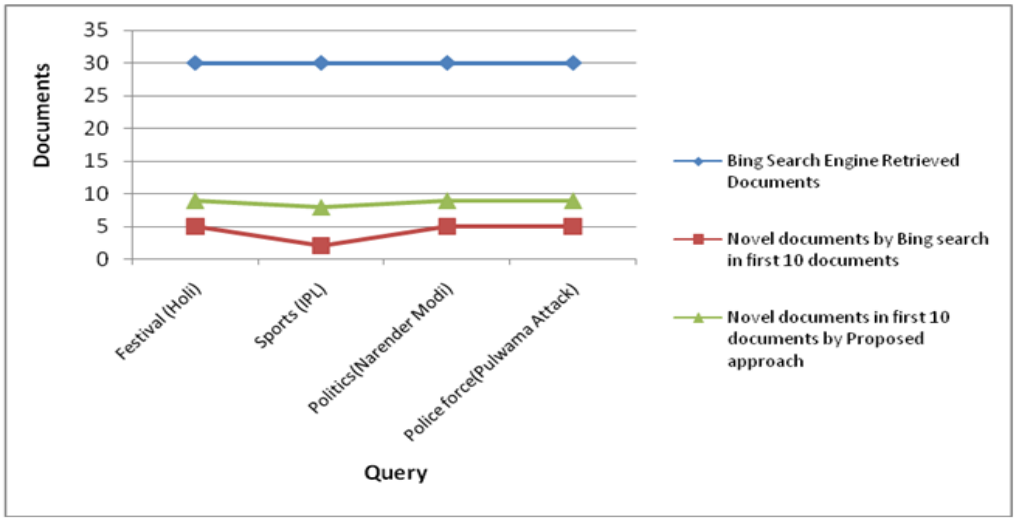

Fig. 5 Comparison of Bing search engine results with proposed approach with in first 10 documents

Result Analysis 1: As shown in the table III, Fig 4 and Fig 5 above that Bing search engine give 09 novel documents from 30 documents for query*' holi'* and only 05 novel documents from the first 10 documents. On the other hand proposed approach give 09 documents which all are novel and filter out the remaining 21 documents out of 30.

Result Analysis 2: As shown in the table III, Fig 4 and Fig 5 above that Bing search engine give 07 novel documents from 30 documents for query*' IPL'* and only 02 novel documents from the first 10 documents. On the other hand proposed approach give 08 documents which all are novel and filter out the remaining 22 documents out of 30.

Result Analysis 3: As shown in the table III, Fig 4 and Fig 5 above that Bing search engine give 08 novel documents from 30 documents for query 'narendra modi' and only 05 novel documents from the first 10 documents. On the other hand proposed approach give 09 documents which all are novel and filter out the remaining 21 documents out of 30.

Result Analysis 4: As shown in the table III, Fig 4 and Fig 5 above that Bing search engine give 09 novel documents from 30 documents for query 'pulwama attack' and only 05 novel documents from the first 10 documents. On the other hand proposed approach give 09 documents which all are novel and filter out the remaining 21 documents out of 30.
From the above results it has been cleared the our proposed approach provided the novel documents for the given query and filter out the redundant documents.

VII. CONCLUSION

In this paper, clustering based approach for novelty detection has been investigated and tested on the set of documents. The incoming stream of documents based on the query have clustered using k-means clustering algorithm and then the clusters head are calculated. The cluster heads selected form different clusters retrieved novel documents and filter out the redundant documents. We have been compared proposed approach with the results given by Bing Search Engine based on the query in different domains. Our proposed approach provided the novel documents based on the given query and filter out the redundant documents. This proposed approach can be further used in the field of information retrieval. Future work further can be enhanced if the architecture can be extended to increase the efficiency of novelty detection method. Some other techniques such a text summarization and semantic similarity can be used which may increase the novelty identification for text documents.

REFERENCES

[1] R.C Balabntaray, C Sharma, and M. Jha, "Document Clustering using K-means and K-medoid", Vol. 1, No. 1, June, 2013.

[2] T. Brants, F. Chen, and A. Farahat, "A System for New Event Detection", *in Proc. SIGIR-03*, pp. 330-337, 2003.

[3] J. Allan, C. wade, and A. Bolivar, "Retrieval and novelty detection at sentence level", *in Proceeding of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, Toronto, Canada* , pp. 314-321, 2003.

[4] G. Satheelaxmi, M.R. Murty, J.V.R. Murty, and P. Reddy, "Cluster analysis on complex structured and high dimensional data objects using K-means and EM algorithm", Vol. 1, No. 1, 2012.

[5] H. Song, L. Wang, B. Li, and X. Liu, "New Trending Events Detection", based on the multi representation index tree clustering", *Int. J. Intell. Syst Appl.,* Vol 3, No. 3, pp. 26-23, 2011.

[6] G. Hu, S. Zhou, J. Guan, and X. Hu, Towards effective document clustering: "A constrained K-means based approach", *Information, Processing and Management,* Vol. 44, No. 4, pp. 1397-1409, 2008.

[7] E. Greengrass: Information Retrieval: A Survey, *DOD Technical Report TR-R52-008-001*, November, 2000.

[8] X. Li and W. B. Croft, "Sentence level information patterns for novelty detection", *Ph.D. dissertation, University of Massachusetts Amherst,* 2006.

[9] J. Carthy, "First Story Detection using a Composite Document Representation," *in Proc. HLT01*, 2001.

[10] L. Zhao, M. Zheng, and S. Ma, "The nature of novelty detection", *information retrieval*, Vol. 9, pp. 527-541, 2006 .

[11] J. Allan, R. Paka and V. Lavrenko, "Online new event detection and tracking", *in Proc of SIGIR-98*, pp. 32-45, 1998 .

[12] X. Li, and W. B. Croft, "Novelty detection based on sentence level patterns", *In: CIKM 2005*, pp. 744–751, 2005.

[13] The anaconda website. [Online] Available: https://www.anaconda. com/distribution/

[14] S. Kumar and K. K Bhatia, "Document to sentence level technique for novelty detection", *in Proc. of CSI Dec, XIV 215 illus., softcover*. pp.104, 2015.

[15] C. Ding, and X, He, "K-means Clustering via Principal Component Analysis", pp. 225-232, 2004.

[16] E.J. Spinosa, A.C.P.L.F. Carvalho, and J.Gama, "Novelty detection with application to data streams" *Intell. Data Anal.,* Vol. 13, No. 3, pp. 405–422, 2009.

[17] EE. R. Faria, J Gama, and A.C.P.L.F Carvalho, "Novelty detection algorithm for data streams multi class problems", *in Proc. 28th Symp. Appl. Comoute.,* pp. 795-800, 2013.