Recommendation System for Bangla News Article with Anaphora Resolution

Kazi Wohiduzzaman[1] and Sabir Ismail[2]
[1]Department of Electrical and Electronic Engineering
[2]Department of Computer Science and Engineering
[1]Metropolitan University, Sylhet, Bangladesh
[2]Sahjalal University of Science and Technology Sylhet, Bangladesh
ohid@metrouni.edu.bd[1] and sabir-cse@sust.edu[2]

Abstract— **This paper represents an efficient approach for Bangla News Recommendation. In traditional Bangla news recommendation system, recommend news from the different newspapers of the same day. Actually, they contain the same news just from different sources. From the user's view, it is more desired if users get to know more diverse information on the same news. In this paper, we have represented a noble approach for recommending news on the same topic with more diverse information. At first, we have done news clustering; it is an automatic learning technique aimed to create clusters that are coherent internally, but substantially different from each other. In this approach, we have used anaphora resolution to increase the keywords frequency. We build an automatic word tagger for anaphora resolution, which can tag all nouns and pronouns with five different criteria (Number, Person, Status, Gender, and POS). Next we have counted document wise unique words to calculate tf-Idf algorithm with cosine similarity to make the recommendation. Finally, we have done three different modified technique of reverse hierarchical clustering on the same cluster to identify more distinct news which is related to the same subject.**

Keywords—NLP, Bangla, News, Hierarchical Clustering, Anaphora, Machine Learning, tf-idf.

I. Introduction

The mounting amount of data or information on the internet makes harder to find what we are actually looking for. Even though the technologies like search engines and RSS readers help us, it is still hard to discover the information we really desire to get.

Recommender systems are built to help us to simply come across the most proper information on the internet. Unlike the search engines recommender systems bring the data to the user without any manual search effort. This is achieved by using the similarities between users. Recommender systems are built to help us to easily find the most proper information on the internet. Unlike the search engines recommender systems bring the information to the user without any manual search effort. This is achieved by using the similarities between users and/or items. There are many methods to build a recommender system and these methods can be applied to many specific domains. Since each application domain has its own specific needs, the method used for recommendations differs.

News clustering or document clustering is defined as determining and assigning topical labels to News or document that define their characteristics [1] [2]. News clustering is used in several contexts such as news indexing, news filtering and search engines and anywhere where News needs sorting [3]. A number of machine learning techniques have been proposed to enhance automatic news clustering.

News recommendation is a process where a cluster is made up of the different news article on the similar topic and of different dates. In conventional news clustering system, they cluster similar news. Pipilika and Google news portal clustering systems are the real examples, where they collected similar news from the different newspapers on a daily basis, those clusters do not present distinct information about the same issue and moreover they did not collect different day's news on similar topic on their cluster. From the user view, cluster needs to include miscellaneous information about the similar topic as a user can find out all distinctive messages from one point on a single topic.

Bangla has an exceptionally rich bunch of dialects and roughly 10% of world's people talk in Bangla. Bangla language is the seventh most talked language on the planet. The amount of electronic news information is available such as electronic publications, Web pages are increasing rapidly for example- digital libraries, electronic books, email messages, news articles, and however, as the volume of online text information increases, the challenge of extracting relevant knowledge increases as well. The need for tools that enhance people to find, filter, and manage these resources have grown. Thus, automatic organization of news collections and news recommendation is a very important research area.

In this paper, we develop a Hierarchical cluster for Bangla news corpus with a new idea of feature selection with the help of Anaphora resolution, with anaphora resolution to increase our keyword weighted value. We have created an automated word tagging algorithm, where we can tag our noun and pronoun with six different criteria. In the first section of our experiment we are Using Hierarchical clustering bottom-up approach (maximum to minimum) for cluster news article and the second section, we use reverse Hierarchical clustering in three different techniques to represent a cluster where viewers can find diverse information on a single topic.

II. Literature Review

Various systems have been proposed for news recommendation; some of them have used content-based news recommendation based on cosine-similarity [4], few recommendation systems are based on collaborative filtering or combination of both recommendation techniques [1] resulting in a hybrid approach to generate news recommendations for users [3]. Recommendations to users based on their interests are demonstrated by them implicitly or explicitly [2]. Another interesting news recommendation technique has been found as a Scalable Two-stage Personalized News Recommendation Approach with a two-level representation, which considers the exclusive characteristics (e.g., news content, access patterns, named

978-1-5386-8279-1/18/$31.00 ©2018 IEEE

entities, popularity, and regency) of news items while performing recommendation [5].

Bangla News recommendation has created a new era though there have been lots of works done by clustering which is fundamentally the same as news recommendation. In Bangla Document clustering, sentences are clustered on the basis of similarity ratio through sentences, N-gram (bi-gram, tri-gram) Based classifier, naïve Bayes Classifier, Locality-sensitive Hashing (LSH) [6][7][8]. Although, few more works have been found in Bangla document clustering based on keyword extraction, tf-idf algorithm and cosine similarity [9]. There are two major techniques are consistently used in document clustering, the group based (k-means) clustering and hierarchical clustering [10][11].

III. Methodology

In this recommendation system, the feature selection will be done with the help of anaphora resolution. Before anaphora resolution, we need to do some pre-processing actions on news corpora by removing all kinds of stop words. Every single step of our recommendation system is shown in figure-1. Stemming will be applied for removing all suffixes and for anaphora resolution. For that, we need to tag data (word). We tag only the noun and pronoun, as the noun is the antecedent of the pronoun. We do anaphora resolution for increasing keyword frequency. In this paper, at first step, we have done a hierarchical cluster "bottom-up" approach consider maximum to minimum of tf-idf cosine similarity value. In the second part of this paper, we have done three different reverse hierarchical clusters for news recommendation.

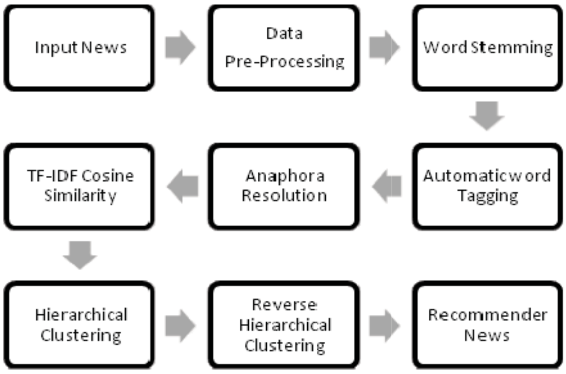


Fig. 1 Block diagram of news recommendation meetthhod

A. Word Stemming

There are a lot of inflectional words in Bangla language. Here is a table of inflectional words from where we have used a suffix list to sort out and remove all suffixes from that type of word to find out the root word.

TABLE I. Word Stemming

বইটি	বইটির	বইয়ের	বইটা	বইগুলি
বই				

B. Automatic Word Tagging

As it is very difficult to tag every word manually one by one in large news corpora, we have, therefore, we built an automatic word tagger shown in figure-2, which can tag all nouns and pronouns with five different criteria (Number-Person-Status-Gender-Parts of speech).

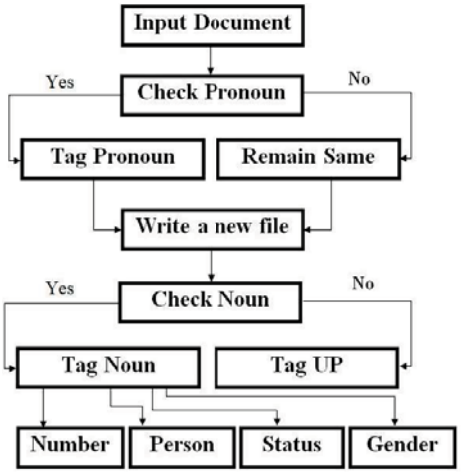


Fig. 2 Diagram of automatic word taaggging

For more understanding the automatic word tagging system, we have considered a single news corpus from news: "আখতার তাঁর পেটে ১৫ ইঞ্চি দীর্ঘ একটি ক্ষতচিহ্ন বয়ে বেড়াচ্ছেন, যেখান থেকে তাঁর একটি কিডনি কেটে নেওয়া হয়েছে।"

At first step, these corpus checks the pronoun file which already exists in this algorithm. If any match is found, it will be tagged with reference tag data, and the rest of the words remain the same.

"আখতার তাঁর (S-TP-N-U-PE-PN) পেটে ১৫ ইঞ্চি দীর্ঘ একটি ক্ষতচিহ্ন বয়ে বেড়াচ্ছেন, যেখান থেকে তাঁর (S-TP-N-U-PE-PN) একটি কিডনি কেটে নেওয়া হয়েছে।"

In the second step, the system checks out if there any noun or not. It tags itself as a noun (NO) in case a noun is found and keeps the rest of the word same.

"আখতার (- - - -NO) তাঁর (S-TP-N-U-PE-PN) পেটে ১৫ ইঞ্চি দীর্ঘ একটি ক্ষতচিহ্ন বয়ে বেড়াচ্ছেন, যেখান থেকে তাঁর (S-TP-N-U-PE-PN) একটি কিডনি (- - - -NO) কেটে নেওয়া হয়েছে।"

Next, tagging a word as noun, it is then checked through a series of criteria to find out the type of that noun, such as: whether the noun is in singular or plural form; whether the noun, as subject, is first, second or third person; whether, as a subject in a sentence, the noun is respectable or normal; and the gender of the noun is masculine, feminine or other.

Applying number, person, status, and gender file, the output comes like-

"আখতার (S-TP-N-M-NO) তাঁর (S-TP-N-U-PE-PN) পেটে ১৫ ইঞ্চি দীর্ঘ একটি ক্ষতচিহ্ন বয়ে বেড়াচ্ছেন, যেখান থেকে তাঁর (S-TP-N-U-PE-PN) একটি কিডনি (S-TP-N-U-NO) কেটে নেওয়া হয়েছে।"

Finally, if any word does not match with the condition of being noun and pronoun, then it is remarked as undefined (UP). Here is the example of our final word tagging output.

"আখতার (S-TP-N-U-PE-PN) তাঁর (S-TP-N-U-PE-PN) পেটে ১৫ (UP) ইঞ্চি (UP) দীর্ঘ (UP) একটি (UP) ক্ষতচিহ্ন (UP) বয়ে (UP) বেড়াচ্ছেন (UP),যেখান (UP) থেকে (UP) তাঁর (S-TP-N-U-PE-PN) একটি (UP) কিডনি (S-TP-N-U-PE-PN) কেটে (UP) নেওয়া (UP) হয়েছে (UP)।

We have collected nouns together, for this experiment, in a single file from various sources and we also make it updated on regular basis, which contains approximately thirty-two thousand (32000) words and those are basically nouns. In the same way, we have found fifty-two (54) different pronouns in Bangla.

In noun classified section, we have created the file containing some conditions for the number, person, and status and also for gender. Since it is quite complicated to find a name and to tag from the noun file, we have made a pre-processing action to reduce this problem and we frequently update the file.

C. *Anaphora Resolution*

Anaphora resolution is one of the most productive zones of research in the natural language processing. For feature selection, we have done anaphora resolution since we can get the maximum keyword weighted value with the help of this method. Following example will make it clear.

"রনি এর দশ বছর, সবার মত রনি ও স্কুলে যায়। সে খুব ভাল ছাত্র না এবং তার পড়ালেখা করতে ও ভাল লাগে না। তার সাইকেল চালাতে ও খেলাধুলা করতে ইচ্ছে করে। সে ভাল ফুটবল খেলতে পারে।"

TABLE II. EXAMPLE OF KEYWORD CHANGING

Keyword	*Normal TF*	*TF with anaphora resolution*
রনি	2	6
তার	2	0
সে	2	0

In this sample document the term frequency of "রনি" is two; it is going to be six when anaphora resolution is done. "রনি" is the antecedent of pronoun "তার" and "সে". Those frequencies are two which are going to be zero.

In anaphora resolution, we consider noun as an antecedent of pronoun [12]. We have constructed an automatic word tagger for anaphora resolution. In this word tagging system, we have used five different criteria (Number, Person, Status, Gender, and Parts of speech). By tagging only noun and pronoun, we have done anaphora resolution, and the rest of the words are denoted by undefined (UP) [16].

D. *Term Frequency-Inverse Document frequency and Cosine Similarity*

We have found out the similarity among the news corpora after calculating term frequency-inverse document frequency (tf-idf) from the equation one and two (1&2) and finally, using equation three (3) calculate the cosine similarity of each document followed by anaphora resolution.

$$TF = How\ many\ word\ in\ a\ single\ news\ article \qquad 1$$

$$IDF = \frac{N}{DF} \qquad 2$$

Where,

TF= How many word are in a single corpus.

N= Total number of document in the corpus.

DF= Document Frequency.

Calculating cosine similarity by using

$$\cos\theta = \frac{\mathrm{A.B}}{\|\mathrm{A}\|\|\mathrm{B}\|} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{A_i B_i}}{\sqrt{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{A_i^2}}\sqrt{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{B_i^2}}} \qquad 3$$

Where, Ai and Bi are components of document A and B respectively.

E. *Hierarchical Clustering*

Hierarchical clustering involves creating clusters that have a predetermined ordering from bottom to top. This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up at the hierarchy. Here, we have done HR cluster to get maximum similarities. After tf-idf cosine similarity calculation, we can find the similarity percentage and make the cluster. Gradually, it goes bottom to up. For our experiment, we consider forty different news corpuses from different online newspapers, news blogs, and news portals.

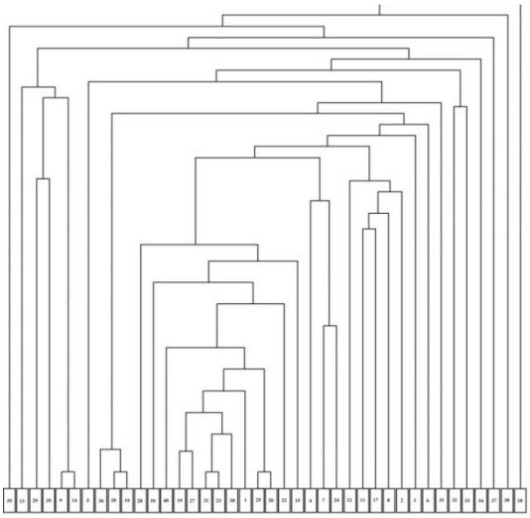
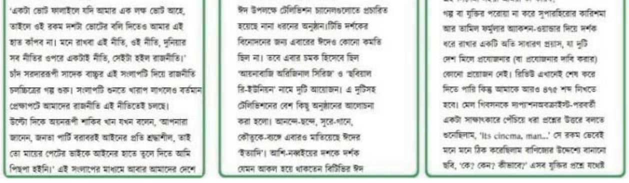


Fig. 3 Hierarchical cluster of Bangla news

This HR cluster represents news corpora with numerical number In figure-3. In this figure corpus number 25 and 26 are most similar, and similarly, 9,14 and 11,17 number corpus are most analogous to each other, very next corpus number 8 is most related to corpus 20 and 34. In figure-4 represent the prototype of the news corpus we have mentioned above so far of the cluster with the pieces of news.



হলে গিয়ে 'রাজনীতি' দেখুন

'একটা ভোট ফালাইলে যদি আমার এক লক্ষ ভোট আছে, তাইলে ওই রকম দশটা ভোটের বলি দিতেও আমার এই হাত কাঁপব না। মনে রাখবা এই নীতি, ওই নীতি, দুনিয়ার সব নীতির ওপরে একটাই নীতি, সেইটা হইল রাজনীতি।' চাঁদ সরদাররূপী সাদেক বাচ্চুর এই সংলাপটি দিয়ে রাজনীতি চলচ্চিত্রের গল্প শুরু। সংলাপটি শুনতে খারাপ লাগলেও বর্তমান প্রেক্ষাপটে আমাদের রাজনীতি এই নীতিতেই চলছে। উল্টো দিকে অয়নরূপী শাকিব খান যখন বলেন, 'আপনারা জানেন, জনতা পার্টি বরাবরই আইনের প্রতি শ্রদ্ধাশীল, তাই তো মায়ের পেটের ভাইকে আইনের হাতে তুলে দিতে আমি পিছপা হইনি।' এই সংলাপের মাধ্যমে আবার আমাদের দেশে

ঈদ আয়োজনে ভরপুর বিনোদন।

ঈদ উপলক্ষে টেলিভিশন চ্যানেলগুলোতে প্রচারিত হয়েছে নানা ধরনের অনুষ্ঠান।টিভি দর্শকের বিনোদনের জন্য এবারের ঈদেও কোনো কমতি ছিল না। তবে এবার চমক হিসেবে ছিল 'আয়নাবাজি অরিজিনাল সিরিজ' ও 'ছবিয়াল রি-ইউনিয়ন' নামে দুটি আয়োজন। এ দুটিসহ টেলিভিশনের বেশ কিছু অনুষ্ঠানের আলোচনা করা হলো। আনন্দে-ছন্দে, সুরে-গানে, কৌতুকে-ব্যঙ্গে এবারও মাতিয়েছে ঈদের 'ইত্যাদি'। আশি-নব্বইয়ের দশকে দর্শক যেমন আকুল হয়ে থাকতেন বিটিভির ঈদ

এই সিনেমা লইয়া আমরা কী করিব?

গল্প বা যুক্তির পরোয়া না করে সুপারহিরোর কারিশমা আর তামিল ফর্মুলার অ্যাকশন-ওয়াজার দিয়ে দর্শক ধরে রাখার একটি অতি সাধারণ প্রয়াস, যা দুটি দেশ মিলে প্রযোজনার (বা প্রযোজনার দাবি করার) কোনো প্রয়োজন নেই। রিভিউ এখানেই শেষ করে দিতে পারি কিন্তু আমাকে আরও ৪৭৫ শব্দ লিখতে হবে। মেল গিবসনকে প্যাশনঅবক্রাইস্ট-পরবর্তী একটা সাক্ষাৎকারে পেঁচিয়ে ধরা প্রশ্নের উত্তরে বলতে শুনেছিলাম, 'Its cinema, man...' সে রকম ভেবেই মনে মনে ঠিক করেছিলাম বাণিজ্যের উদ্দেশ্যে বানানো ছবি, 'কে? কেন? কীভাবে?' এসব যুক্তির প্রশ্নে যথেষ্ট

Fig. 4 Cluster with news article

F. Reverse Hierarchical Clustering

In the second section of this paper, we have applied reverse hierarchical cluster which has been represented in the form of three new techniques for checking the efficiency of the algorithm. For this experimentation, we have taken only thirteen different documents from our cluster having diversified information but the same topic from different news portal of three consecutive days.

First Method

At first, we, for the experiment, put the first method which takes the samples of the news, and consider their similarity. Then, it connects with the document which has less similarity. In figure-5 corpus number 3 and 6 are far indifferent from each other and those two corpora are connected first and then corpus 5 is most divergent, so corpus 5 are connected with cluster 3 and 6. Thus it continues to the end of maintaining the same process.

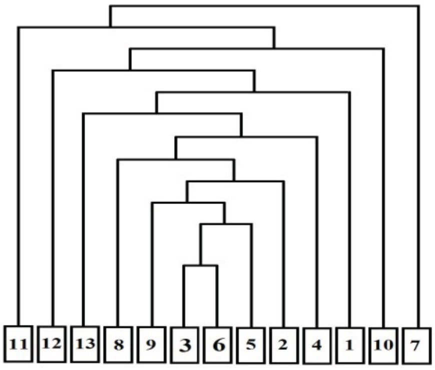


Fig. 5 HR cluster for first method

Second Method

Despite having a slight resemblance to the first method, the second method functions in another way. Like the first one, the second method connects two documents having the least similarity, but, unlike the first one, it continues connecting the documents making individual pairs or cluster. Finally, it starts connecting the cluster as the first method does. At the time of connecting, if there is a single document left, it will be connected later when coupled documents get connected.

From the figure-6, news corpora are made up couple cluster like (1,5),(2,9),(4,10),(7,11),(8,13),(3,6), then very next step those couple clusters are connected each other. As we discussed earlier if any single news corpus failed to make a couple cluster and left single then it will be connected with any of couple cluster on the second step.

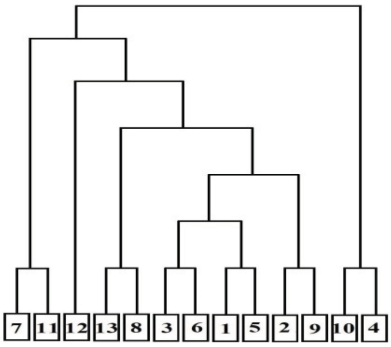


Fig. 6 HR cluster for second method

Third Method

Finally, in the third method taking the minimum value of tf-idf cosine similarity as with earlier method we did, we have made them connected to have a cluster (coupled), Next, a single document will be created with those two documents (3&6 makes a new document X), which is shown in figure-7. The same system will run again and find the most dissimilar document with our combined document (minimum value of tf-idf cosine similarity) and connect this document (9&X makes new document Y) with earlier coupled (cluster). Again a single file with those two documents will emerge and reduce another document consequently. We will have a continuation of this loop until the end of the cluster. As a final point, we will get our hierarchical cluster for different thematically connected news content for the same news topic.

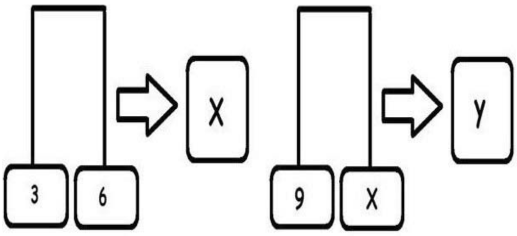


Fig. 7 First two step of HR cluster for third method

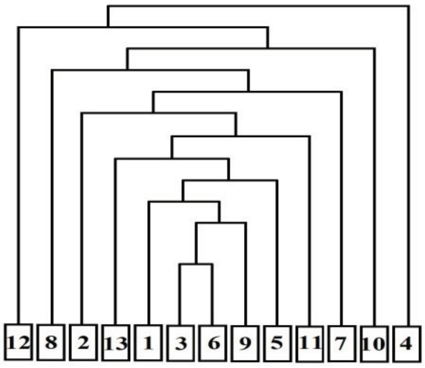


Fig. 8 HR cluster for third method

IV. Result Analysis

A. Hierarchical Clustering

Our first hierarchical cluster methods with most similarity news corpora worked better. For testing our first HR method, we consider another three HR cluster; the first one is normal keyword tf-idf weighted, the second one is normal keyword if-idf cosine similarity and the third one is keyword weight after anaphora resolution, and we find our method works better. For result analysis, we have made a small amount of news corpora cluster so that we can compare with each cluster and we got the better result with our proposed methods.

As we mentioned earlier that, to know our system accuracy we give manually input in our system, so we know that our news corpus 01, 02, 08, 11, 12, is the news of Bangla movie related, those are connected each other before being connected with other news's. Similarly, news corpus 09, 14, 13 are sports news, and also document 04, 07, 10 is the news of cricket and they are connected each other at first and eventually, they get connected with all other news.

B. *Reverse Hierarchical Clustering*

In the final section, we have proposed three shorts of reverse hierarchical clustering algorithms to find more distinct information on a single news topic. What pipilika and Google news portal do is to cluster similar or most relevant news on a single topic; however, our proposed methods work differently. Unlike Pipilika and Google, Reverse hierarchical clustering finds out more different or more diverse news on a similar topic. In addition, Pipilika and Google show the news of last one day (24 hour) showing the latest news first while our processes show news more period three days or more. The finding result are shown on the basis of date, i.e. the news happened earlier is shown first and chronologically the latest news is shown.

As mentioned above, we have proposed three different types of reverse hierarchical cluster and as expected we have got slightly different outcomes distinctly although all those are efficient enough to get our expected result.

Output of First Method

If we organize our first cluster with news corpuses with their date of publication then we find the arrangement of news corpuses in the table III.

TABLE III. NEWS RECOMMENDATION FOR FIRST METHOD

Date	News No.	Newspaper name	News Cluster
30-7-17	**3**	যুগান্তর	প্রধানমন্ত্রীর বাসভবন ছাড়লেন নওয়াজ
29-7-17	**6**	ইত্তেফাক	পাকিস্তানের পরবর্তী প্রধানমন্ত্রী শাহবাজ, অন্তর্বর্তীকালীন আব্বাসী
29-7-17	**5**	যুগান্তর	নির্বাচিত প্রধানমন্ত্রী হবেন শাহবাজ, পাকিস্তানের অন্তর্বর্তী প্রধানমন্ত্রী শহীদ খাকান আব্বাসি
29-7-17	**9**	প্রথম আলো	নওয়াজ না ইমরান: দম কার বেশি লম্বা
30-7-17	**2**	ইত্তেফাক	পাকিস্তানে প্রধানমন্ত্রী নির্বাচনে ভোট মঙ্গলবার
29-7-17	**8**	প্রথম আলো	শহীদ খাকান আব্বাসী অন্তর্বর্তীকালীন প্রধানমন্ত্রী, ছোট ভাইকেই বেছে নিলেন নওয়াজ
30-7-17	**4**	প্রথম আলো	নওয়াজ সরে যাওয়ায় লাভ কার?
28-7-17	**13**	যুগান্তর	নওয়াজ শরীফের পতন: বিরোধীদের উল্লাস
30-7-17	**1**	ইত্তেফাক	প্রধানমন্ত্রীর বাসভবন ত্যাগ করলেন নওয়াজ
28-7-17	**12**	প্রথম আলো	এখন নওয়াজের কী হবে?
28-7-17	**10**	ইত্তেফাক	পদত্যাগ করলেন নওয়াজ শরিফ
28-7-17	**11**	ইত্তেফাক	নওয়াজ শরিফকে অযোগ্য ঘোষণা পাকিস্তান সুপ্রিমকোর্টের
29-7-17	**7**	ইত্তেফাক	নওয়াজের পদত্যাগে ইমরান খানের সাবেক স্ত্রীর স্বস্তি

Output of second Method

This is the output arrangement of our second reverse hierarchical method in table IV. the table represents news headline with their publication date and also newspaper name.

TABLE IV. NEWS RECOMMENDATION FOR SECOND METHOD

Date	News No.	Newspaper	News Cluster
30-7-17	3	যুগান্তর	প্রধানমন্ত্রীর বাসভবন ছাড়লেন নওয়াজ
29-7-17	6	ইত্তেফাক	পাকিস্তানের পরবর্তী প্রধানমন্ত্রী শাহবাজ, অন্তর্বর্তীকালীন আব্বাসী
30-7-17	1	ইত্তেফাক	প্রধানমন্ত্রীর বাসভবন ত্যাগ করলেন নওয়াজ
29-7-17	5	যুগান্তর	নির্বাচিত প্রধানমন্ত্রী হবেন শাহবাজ, পাকিস্তানের অন্তর্বর্তী প্রধানমন্ত্রী শহীদ খাকান আব্বাসি
29-7-17	9	প্রথম আলো	নওয়াজ না ইমরান: দম কার বেশি লম্বা
30-7-17	2	ইত্তেফাক	পাকিস্তানে প্রধানমন্ত্রী নির্বাচনে ভোট মঙ্গলবার
29-7-17	8	প্রথম আলো	শহীদ খাকান আব্বাসী অন্তর্বর্তীকালীন প্রধানমন্ত্রী, ছোট ভাইকেই বেছে নিলেন নওয়াজ
28-7-17	13	যুগান্তর	নওয়াজ শরীফের পতন: বিরোধীদের উল্লাস
28-7-17	12	প্রথম আলো	এখন নওয়াজের কী হবে?
30-7-17	4	প্রথম আলো	নওয়াজ সরে যাওয়ায় লাভ কার?
28-7-17	10	ইত্তেফাক	পদত্যাগ করলেন নওয়াজ শরিফ
28-7-17	11	ইত্তেফাক	নওয়াজ শরিফকে অযোগ্য ঘোষণা পাকিস্তান সুপ্রিমকোর্টের
29-7-17	7	ইত্তেফাক	নওয়াজের পদত্যাগে ইমরান খানের সাবেক স্ত্রীর স্বস্তি

Output of Third Method

This is the output arrangement of our third reverse hierarchical method with news corpuses and it is shown in table V.

TABLE V. NEWS RECOMMENDATION FOR THIRD METHOD

Date	News No.	Newspaper	News Cluster
30-7-17	3	যুগান্তর	প্রধানমন্ত্রীর বাসভবন ছাড়লেন নওয়াজ
29-7-17	6	ইত্তেফাক	পাকিস্তানের পরবর্তী প্রধানমন্ত্রী শাহবাজ, অন্তর্বর্তীকালীন আব্বাসী"
29-7-17	9	প্রথম আলো	নওয়াজ না ইমরান: দম কার বেশি লম্বা
30-7-17	1	ইত্তেফাক	প্রধানমন্ত্রীর বাসভবন ত্যাগ করলেন নওয়াজ
29-7-17	5	যুগান্তর	নির্বাচিত প্রধানমন্ত্রী হবেন শাহবাজ, পাকিস্তানের অন্তর্বর্তী প্রধানমন্ত্রী শহীদ খাকান আব্বাসি
28-7-17	13	যুগান্তর	নওয়াজ শরীফের পতন: বিরোধীদের উল্লাস
28-7-17	11	ইত্তেফাক	নওয়াজ শরিফকে অযোগ্য ঘোষণা পাকিস্তান সুপ্রিমকোর্টের
30-7-17	2	ইত্তেফাক	পাকিস্তানে প্রধানমন্ত্রী নির্বাচনে ভোট মঙ্গলবার
29-7-17	7	ইত্তেফাক	নওয়াজের পদত্যাগে ইমরান খানের সাবেক স্ত্রীর ‘স্বস্তি
29-7-17	8	প্রথম আলো	শহীদ খাকান আব্বাসী অন্তর্বর্তীকালীন প্রধানমন্ত্রী, ছোট ভাইকেই বেছে নিলেন

			নওয়াজ
28-7-17	**10**	ইত্তেফাক	পদত্যাগ করলেন নওয়াজ শরিফ
28-7-17	**12**	প্রথম আলো	এখন নওয়াজের কী হবে?
30-7-17	**4**	প্রথম আলো	নওয়াজ সরে যাওয়ায় লাভ কার?

V. Conclusion

In this paper, we have done reverse hierarchical clustering of the same cluster to classify more distinct news. Unlike other search engines, we have applied three different methods of reverse Hierarchical clustering to find different news or information related to the same subject. What we found in pipilika or Google news portal that they assemble daily news and formulate cluster by collecting analogous news from the different sources of the newspaper. In this regard, they take the news cluster into account on daily news basis and do not cover news of previous days on the same issue. It, from a user view, is more desired provided that they get to know more diversified information of the same news since it creates monotony. And what we have done in this paper is that we have put an efficient approach into operation for recommending news of the same topic along with more diversified information where viewers can come across almost all of the unlike news on the same agenda in a single cluster which also ensures last few days' news. Three different recommendation approaches are based on reverse hierarchical clustering we have employed shows three different results although all these three methods have successfully shown our desired output, finding different news on a single news agenda in a single cluster. However, it should be mentioned that our third method is proved more effective, in terms of showing result, than the first and second one.

References

[1] R. Burke, P. Brusilovsky, A. Kobsa, & W. Nejdl, "Hybrid web recommender systems", The Adaptive Web, LNCS 4321, pp. 377 – 408, 2007.

[2] G. Adomavicius & A. Tuzhilin, "Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions", IEEE Transactions on Knowledge and Data Engineering, vol 17, no 6, pp: 734 – 749, June 2005.

[3] M. Sood, & H. Kaur, "Survey on news recommendation", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, vol-3, issue-6, 2014. ISSN ONLINE (2278-8875) PRINT (2320-3765).

[4] M. Kompan & M. Bielikova, "Content-Based News Recommendation", 11th International Conference on Electronic Commerce and Web Technologies, , EC-Web 2010, Bilbao, Spain, September 1-3, 2010, DOI: 10.1007/978-3-642-15208-5_6.

[5] L. Li, D. Wang, T. Li, D. Knox & B. Padmanabhan, "A Scalable Two-Stage Personalized News Recommendation System", Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval Beijing, China — July 24 - 28, 2011, ISBN: 978-1-4503-0757-4 .

[6] M. M. Haque, S. Pervin, & Z. Begum, "Automatic Bengali news documents summarization by introducing sentence frequency and clustering", 18th International Conference on Computer and Information Technology (ICCIT), 21-23 Dec. 2015, Electronic ISBN: 978-1-4673-9930-2.

[7] A. N. Chy, M. H. Seddiqui & S. Das, "Bangla news classification using naïve Bayes Classifier" 16th International Conference Computer and Information Technology, 8-10 March 2014, Khulna, Bangladesh.

[8] M. Monsur, N. Uzzaman & M. Khan, "Analysis of N-gram based text categorization for bangla in a newspaper corpus" Center for Research on Bangla Language Processing (CRBLP), BRAC University, 2006.

[9] K. Sarkar, "An approach to summarizing Bengali news documents" Proceeding of the International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2012. Pages: 857 – 862.

[10] C. Lee, C. Hsu, & D. Chen, "A Hierarchical document clustering approach with frequent item sets", International journal of engineering and technology, Volume 9, Number 2, April 2017.

[11] J. Graovac, J. J. Kovacevic, & G. M. Pavlovic-Lazetic, "Hierarchical vs. flat n-gram-based text categorization: can we do better?", Computer science and information system 14(1):103-121, January 2017.

[12] T. Tazakka, Md. Asifuzzaman, & S. Ismail, "Anaphora Resolution in Bangla Language", International Journal of Computer Applications (0975 – 8887) Volume 154 – No.9, November 2016.

[13] Chatterji, & Sanjay, "Anaphora resolution for bengali, hindi, and tamil using random tree algorithm in weka." In Proceedings of the ICON-2011 (2011).

[14] Sikdar, U. Kumar, A. Ekbal, S. Saha, O. Uryupina, & M. Poesio, "Anaphora Resolution for Bengali: An Experiment with Domain Adaptation." Computación y Sistemas 17, no. 2 (2013): 137-146.

[15] S. Ismail, M. & S. Rahman, "Bangla Word Clustering Based on N-gram Language Model" International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), 10-12 April 2014, Dhaka.

[16] Md. M. Haque, S. Pervin, Z. Begum, "Rule Based Replacement of Pronoun by Corresponding Noun for Bangla News Documents", International Journal of Technology Diffusion (IJTD), Volume-8, Issue-2, Pages 26-42, March, 2017.