Computers, Materials & Continua
DOI:10.32604/cmc.2022.019451
Article

Tech Science Press

Hierarchical Stream Clustering Based NEWS Summarization System

M. Arun Manicka Raja[1,*] and S. Swamynathan[2]

[1]Department of Computer Science and Engineering, RMK College of Engineering and Technology, Chennai, 602106, India
[2]Department of Information Science and Technology, College of Engineering Guindy, Anna University, Chennai, 600025, India
[*]Corresponding Author: M. Arun Manicka Raja. Email: arunmcse@rmkcet.ac.in

Received: 14 April 2021; Accepted: 16 May 2021

Abstract: News feed is one of the potential information providing sources which give updates on various topics of different domains. These updates on various topics need to be collected since the domain specific interested users are in need of important updates in their domains with organized data from various sources. In this paper, the news summarization system is proposed for the news data streams from RSS feeds and Google news. Since news stream analysis requires live content, the news data are continuously collected for our experimentation. The major contributions of this work involve domain corpus based news collection, news content extraction, hierarchical clustering of the news and summarization of news. Many of the existing news summarization systems lack in providing dynamic content with domain wise representation. This is alleviated in our proposed system by tagging the news feed with domain corpuses and organizing the news streams with the hierarchical structure with topic wise representation. Further, the news streams are summarized for the users with a novel summarization algorithm. The proposed summarization system generates topic wise summaries effectively for the user and no system in the literature has handled the news summarization by collecting the data dynamically and organizing the content hierarchically. The proposed system is compared with existing systems and achieves better results in generating news summaries. The Online news content editors are highly benefitted by this system for instantly getting the news summaries of their domain interest.

Keywords: News feed; content similarity; parallel crawler; collaborative filtering; hierarchical clustering; news summarization

1 Introduction

Knowledge identification from online news articles have received keen attention among the news readers, especially from the Really Simple Syndication (RSS) feed-based news updates and Google news [1]. The knowledge extracted from various news sources are mapped into many day-to-day applications. Various events are identified from news articles and the summaries are generated about a particular event with respect to different timelines [2]. The news events are extracted by identifying the named entities present in the news content. The abstractive and extractive summaries are generated using summarization techniques such as abstractive and extractive summarizations [3]. The semantic relevance is estimated using the wordnet and the hierarchical structure is represented for news articles [4]. Single news article contains many keywords related to a particular topic. It is necessary to identify the domain of the keywords by tagging the keywords present in the news. Though the keywords are tagged in the news content, it is important to organize the content in a hierarchical structure for retrieving the similar news content for summarizing to the users.

In this work, a news clustering based summarization system is proposed to cluster various category of news content from multiple news sources and to generate news summaries on user interested topic. The proposed system is distinctive in handling the news updates for effectively organizing the news content to retrieve it later. Further, the extractive summary of the specific topic is generated from the clustered news contents. The proposed system has been evaluated for news crawling, news content retrieval and news summarization. The evaluation results shown that the proposed system performs better in summarizing the news contents to the end users.

The paper is organized as following sections. In Section 2, the related works of the clustering and news summarization mechanisms are discussed. In Section 3, the architecture of the news retrieval system is explained. In Section 4, the experimental results of the proposed system are discussed. In Section 5, the performance evaluation of the parallel crawler, hierarchical clustering and news summarization method are explained. In Section 6, the conclusion of the work is given.

2 Related Works

In recent years, there are lot of online recommendation systems available for assisting online shopping to various users depending upon their knowledge level. Here, we have discussed various methods related to the data collection, domain corpus, hierarchical clustering and summarization. RSS new feeds are the important sources of information from different online websites. The users are subscribing to only the required feed updates [5]. In addition to the RSS feeds, Google news is providing news updates on various domains. In addition, the news contents are extracted for building corpuses which help domain oriented news analysis [6]. Wordnet [7] is the prominently used Synset generator along with the tagger. The feed updates contain titles which has keywords that are used to identify the domains. They have used wordnet for tagging the keywords and identify the domain wise data.

Multi granularity hierarchical representation [8] is the content representation of the data for easy access of the fine grain level data. The authors have employed this method for the systematic organization of the content and its retrieval. Further, RSS news feeds are represented in Extensible Mark-up Language (XML) formats [9]. This method is effective if the similar news items are merged together to gather the news from various sources. The relatedness between the RSS elements is also identified to merge the contents effectively. The RSS news articles are collected from various sources. In many cases, the news articles are redundant [10] in content wise. These redundant articles may be eliminated and the distinctive news articles may be clustered for later access. News content clustering and recommendation requires the categorization of the users in the web and their web browsing behavior needs to be analysed. The authors [11] have used user behavior data along with collaborative filtering for recommending the specific user interested content. Latent semantic analysis use mapping of high dimensional and sparse words into a semantic space with the correlation among the words [12]. Text analysis model [13] uses deep learning techniques for effective product recommendation to the users.

In addition, it is essential to summarize the categorized news contents to the respective users. Extractive summarization [14] is one of the summarization techniques. It captures the sentences from the documents and generates the summary from the captured contents. Contextual information is used with the captured contents to generate effective summaries. The social media contents are summarized [15] topic wise and given to the users. Further, some of the semantic based clustering [16] is helpful in generating summaries from non-conclusive short texts. External knowledge resources are used for establishing the semantic relations among the text contents. Multi-document summarization [17] system is used for generating summaries from multiple document collections. The best summary is generated by estimating the information distance among the document collections. Many works have been carried out in the literature for incorporating the credible features of few existing mechanisms for developing a system with better performance. Few such works are used in building prediction models [18] and creating fake news detection system [19]. In some works, the deep learning-based algorithms are used as risk analysis models and building mechanisms for defensing from denial-of-service attacks [20,21]. The comparison of various methodologies related to the proposed system has been tabulated in Tab. 1.

Table 1: Comparison of related works with merits and demerits of the methodologies

Authors	Title	Methodology/Algorithm	Merits	Demerits
Taddesse et al. [8]	Semantic-based merging of RSS items	Multi granularity hierarchical representation for summarization	easy access of the fine grain level data	not evaluated for multiple domains. Higher retrieval time.
Xu et al. [10]	Research on topic discovery technology for web news	Collaborative filtering-based content retrieval for summary generation	Web browsing behaviour of users is used	User categorization performed but content organization not done.
Diao et al. [11]	CRHASum: extractive text summarization with contextualized-representation hierarchical-attention summarization network	Latent semantic analysis-based summary generation	Semantic correlation among words with mapping of high dimensional words	Evaluation done on only limited corpuses.
Katarya et al. [12]	Capsmf: A novel product recommender system using deep learning based text analysis model	Text analysis model for summarization	Apply deep learning mechanism for content recommendation	Only content analysis done, not evaluated for user query collaborated recommendation.
Balahur et al. [13]	Challenges and solutions in the opinion summarization of user-generated content	Extractive summarization	Contextual information is used for summary generation.	Extracted sentences make contextual overload in summary generation.
Long et al. [15]	A new approach for multi-document update summarization	Semantic based clustering for summary creation	Semantically identified short texts help for better summary generation	Limited semantic relation established among contents.

This research paper work is motivated and inspired by the related works discussed in this section. Our proposed system provides an improvement to the news summarization methods for news data streams and content retrieval is simplified with hierarchical news content clustering and user collaborative filtering. The quality of summary generation has significantly improved.

3 Collaborative Filtering Based NEWS Retrieval System

Hierarchical clustering is applied in many of the content retrieval system. Since hierarchical structure provides topic wise categorical representation elegantly, it is widely encouraged in most of the content structuring works. The retrieval time is considerably less in hierarchical structured content retrieval system [22]. It performs well in processing the user given query and recommend better results from the hierarchically arranged contents for summary generation. Hence, we have proposed hierarchically clustering based news summarization system for generating effective news summaries in less time. The flow chart of the proposed system is illustrated in Fig 1.

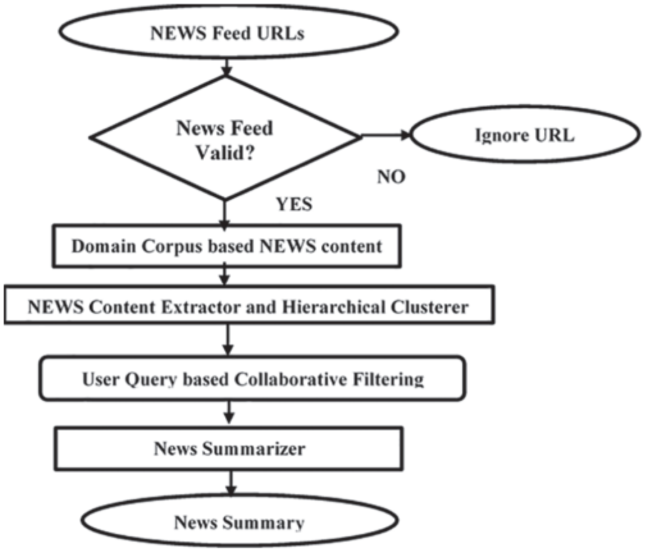
NEWS Feed URLs
News Feed Valid?
Ignore URL
NO
YES
Domain Corpus based NEWS content
NEWS Content Extractor and Hierarchical Clusterer
User Query based Collaborative Filtering
News Summarizer
News Summary

Figure 1: Flow chart of the collaborative filtering based news retrieval system

The architecture of the proposed system is shown in Fig 2. The feed collector helps to collect the news feeds from various news Uniform Resource Locator (URL). Further, the collected feeds are checked for the domain specification in the title content available in the feed and the domain of the feed is identified. Various domain corpuses along with wordnet are used for checking the domain of the feeds. Hierarchical clustering is used for clustering the news articles category wise. It performs the categorization of the news contents and organizes the content topic wise. The user queries are obtained and the summaries are given as a result to the users. The user queries are natural language-based keywords. The summarizer generates the summary both topic wise and magazine wise. In addition, the user given keyword specific news contents are also retrieved from the repository.

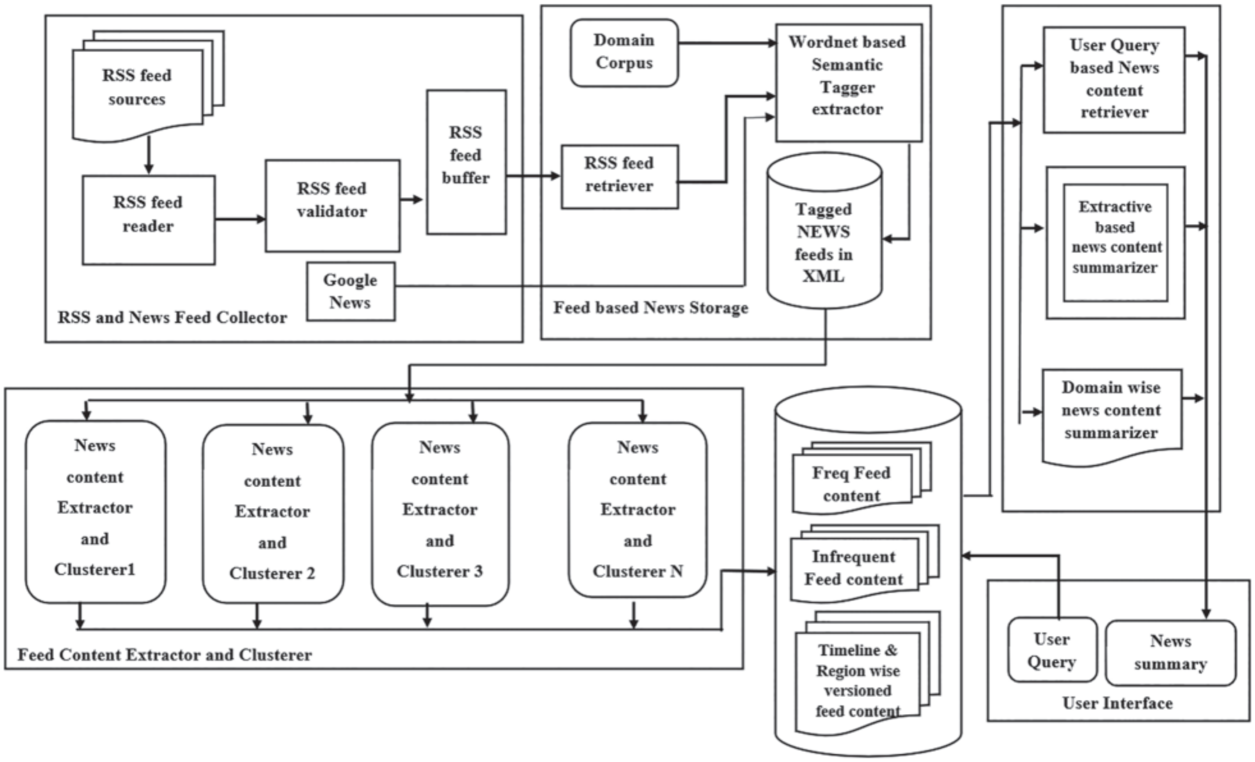
RSS feed sources
RSS feed reader
RSS feed validator
RSS feed buffer
Google News
RSS and News Feed Collector
Domain Corpus
Wordnet based Semantic Tagger extractor
RSS feed retriever
Tagged NEWS feeds in XML
Feed based News Storage
User Query based News content retriever
Extractive based news content summarizer
Domain wise news content summarizer
News content Extractor and Clusterer1
News content Extractor and Clusterer 2
News content Extractor and Clusterer 3
News content Extractor and Clusterer N
Feed Content Extractor and Clusterer
Freq Feed content
Infrequent Feed content
Timeline & Region wise versioned feed content
User Query
News summary
User Interface

Figure 2: Collaborative filtering based NEWS retrieval system

4 Experimental Results and Discussion

The significance of this research work focusses on collecting the news data dynamically and organizing the news data hierarchically. Further, the news contents are summarized effectively based on the user given query by processing with the collaborative filtering method.

4.1 Dataset

The dataset used in this work, is collected from the news sources using the news crawler program which we implemented in our system as part of news summarization system. The RSS feed news and news data streams are monitored and collected from google news [23].

4.2 News Data Collection

In this work, the news summaries are generated based on the user interest using the news updates received from numerous sources. To perform this, the first stage of work considered in this paper, is the data collection from various sources. The hierarchical structure is created for various domains. For example, Sports news are categorized with different types like cricket, football, basketball, etc. In addition, the region wise hierarchy is also represented to easily identify the location of the news such as country, state, district, city, etc. The consolidated summary of the news data collection is shown in Tab. 2 wherein the news source and the topics and its news updates count are tabulated. The news content collection is observed for 1-day, 1-week, 1-month and 3-month period. The news articles collected during these periods is illustrated in Tab. 3.

Table 2: Summary of various news sources, news topics and news updates count

News source	Top news topics	News articles count
Google news	Hand sanitizer	36
	Damage to the lungs of COVID 19 patients	27
	Coronavirus symptoms	46
	Coronavirus live in patients for 5 weeks	34
	Vegetarian diet prevents a stroke	17
Times of India	Asymptomatic patient	36
	Children less vulnerable to coronavirus infections	25
	Low dose aspirin mitigates liver cancer risk	32
	List of testing centres in India	28
	Drugs to tackle coronavirus	38
Hindu	West Bengal government closes all educational institutions	16
	On IPL and coronavirus	47
	Coronavirus cases increase in the country	42
	Coronavirus treatment	22
	Countries winning corona battle	19
	Youngest corona virus victim	18
	India and coronavirus	11
	Coronavirus—who all cant travel to India	12
	Avian flu: culling of birds begins in Malappuram	17

Table 3: Weekly, monthly statistics of the news feed updates

Domain	1-day (10.04.2021) news count	1-week news count (05.04.2021 to 12.04.2021)	1-month news count (11.03.2021 to 12.04.2021)	3-month news count (11.01.2021 to 12.04.2021)
Health	12	136	542	1654
Sports	15	147	448	1428
Business	10	124	492	1574
Technology	18	159	656	1952
Entertainment	22	173	752	2159
Science	26	98	398	1027

4.3 Domain Corpuses

Around 97000 words are available in political domain corpus [24] and it has been applied in tagging the keywords from news updates. Healthcare domain consists of around 60000 words [25] and it has been incorporated to identify similar terms in the news contents. The business corpus contains around 600000 words [26] and it is used to know the business keywords present in it. Sports domain consists of keywords from various sports events. Around 32000 sports related words are available in sports corpus [27]. Education domain contains the terms prominently used in education related activities. There are around 84000 words available in education domain [28]. Electronics, nature, software and travel domain corpuses are taken from Wikipedia corpus collection [29].

4.4 Hierarchical Clustering of News Articles

Cosine similarity is determined to find the similar content existing in the news updates. Hierarchical clustering algorithm is used to detect the hierarchical structures among the news articles. The algorithm is shown as follows.

Algorithm 1: Hierarchical Clustering Algorithm

```
Input Data: $D = \{d_1, d_2, d_3, \ldots, d_n\}$ data set containing 'n' news articles
Result: hierarchy of clusters
begin
    Initialize clusters
    Assign $C_i = x_i$ wherein $x_i$ represents data points and $i = 1$ to $n$
    Create cluster for each news article
    loop
     for i ← 1 to n do
      for j ← 1 to n do
       $d(i, j) \leftarrow$ compute_similarity($x_i$, $x_j$)
         calculate similarity measure
         identify appropriate cluster
         for each $x_i$ in C do
             $\alpha_i = \text{score}(x_i)$        //assigns similarity score value to alpha
         $\text{Cluster}_d = \text{highest}(\alpha_i)$    // similarity score makes cluster distance among clusters
         $(C_i, C_j) \leftarrow C_d$
       $(C_{min1}, C_{min2})$ = minimum_dist $(C_i, C_j)$ for all $C_i$, $C_j$ in C
       remove $C_{min1}$ and $C_{min2}$ from C
       add $\{C_{min1}, C_{min2}\}$ to C
       clusters = clusters + 1
         assign data to closest clusters $C_i$ or $C_j$
         remove from current clusters
    until all clusters generated
    return clusters
end
```

The domain of the cluster is also identified with the clustering process. The cluster formation from various news articles is tabulated in Tab. 4. The top 3 clusters formed out of the news articles received in a particular interval is mentioned in Tab. 5

The clustered articles with its corresponding clusters and domain, is shown in Tab. 6. It contains the cluster category, cluster topic, number of feeds and number of news articles.

4.5 Collaborative Filtering Based NEWS Content Retrieval

The collaborative filtering algorithm is used to filter the similar news content among the interested users. The similar news content is added to the recommendation set. The collaborative filtering based score is calculated for every similar news content and the news with maximum score is recommended to the user. The collaborative filtering algorithm is shown as follows.

Table 4: Cluster formation from news articles

Number of feeds	Number of articles	Number of clusters
1026	3245	12
942	2287	16
1124	3695	24
1089	3578	21
846	1896	13

Table 5: Top 3 cluster information

Cluster	Number of feeds	Number of news articles
Health	70	749
Science	32	198
Technology	70	387

Table 6: Cluster categories, Topics with its feeds and article statistics

Category of the cluster	Cluster topics	Number of news feeds	Number of news articles
Health	Coronavirus	32	356
	Alzheimers disease	12	69
	Menopause	4	98
	Cereals for kids	4	95
	WHO	2	32
	Ebola outbreak	8	48
	Dry skin	5	35
	midlife sex slump	3	16
Science	NASA's Mars Lander	12	82
	Arctic ocean	10	56
	Cosmic fire	2	14
Technology	Windows 10	8	33
	Motorolo	9	47
	OnePlus 8	2	14
	Xiaomi	5	36
	Samsung	8	47
	Sony	4	18
	Microsoft	3	17

Algorithm 2: User based Collaborative Filtering

```
Input: user set ‘S’, user query data ‘U’
Output: Suggested results ‘R’
R ← null
foreach u in S do
    C(u) ← null, C represents collaborativeness
    neighbour(u) ← null,
    c ← n, wherein ‘n’ represents ‘news sequence’
foreach v in S do
      if Sim (u, v) >= T, T represents filtering threshold
          neighbour(u) ← neighbour(u) U v
          if $v_i$ + 1 exists then
              C(u) ← C(u) U $v_i$ + 1
      for $n_j$ in C(u)
          compute $r_j$          //computes the recommendation
          $r_j$ ← $n_j$          //mapping user to recommendation
          R ← R U $r_j$
      return R          //returns the recommended news to users
```

The results of the collaborative filtering algorithm are shown in Tab. 7.

Table 7: Collaborative filtering accuracy using recommended articles

User query	No. of news articles related to query	No. of correctly recommended news articles to the specific user	Collaborative filtering accuracy
Badminton	124	96	77.41%
NASA	142	107	75.35%
Pandemic	136	103	75.73%
Delhi metro	118	93	78.81
COVID19	374	271	72.45

4.6 News Content Summarization

We have applied Extraction based summarization algorithm as a baseline method for performing document summarization using multiple document contents. Further, we have computed the probability distribution of the news for summary generation. The sentence with maximum score is taken for summary generation. The summarization steps are represented as follows.

Step 1. Calculate the probability distribution of the words in the news content.
Let w_i be the words and $p(w_i)$ be the probability of the words, where i represents the words.

Step 2. Calculate the average probability of words in the news content.

$$weight\ (S_j) = \sum_{wi \in Sj} \frac{p(wi)}{|\{wi \in Sj\}|},\ \mathrm{S_j}\text{—sentences}$$

Step 3. Choose the top score sentence that has the high probability word.

Step 4. Update probability for word chosen
$p_{new}(w_i)=pold(w_i).\ p_{old}(w_i)$

Step 5. Generate summary based the probability estimation of words in the sentences

The information about the user submitted query and the summary generation details from the news feeds is shown in Tab. 8.

Table 8: Statistics about Query *vs.* Summary

Query	Feeds count	Article count	Total words present in article contents	Total words present in the summary
Corona virus	2	6	192	102
Avian flu	3	8	300	158
Delhi violence	2	7	184	97
Liver cancer	5	9	176	95
Vegetarian diet	3	5	142	84

Tab. 9 shows the generated summary with the news feed count and article count for the user given query. The summaries generated for various user given queries are shown in Tab. 10.

Table 9: Query based summary with news feed and article count

Query	Feeds count	Article count	Summary
Corona virus	2	6	The Indian government has defended its handling of the coronavirus outbreak after a strict lockdown—introduced with little warning—left millions stranded and without food. India has been put in lockdown to halt the spread of the coronavirus outbreak. India has been criticised for its poor record of testing people in the battle against coronavirus. A 68-year-old woman from Delhi has been confirmed as the second Indian to die from the coronavirus. With India now in a 21-day lockdown to prevent the spread of the coronavirus, there's been plenty of advice shared on how to prevent or cure the disease.
Avian flu	3	8	Highly pathogenic avian influenza has been reported in new regions of Germany and Hungary. Two children have been confirmed with flu infections of avian origin. Highly pathogenic avian influenza has returned to the Philippines, after an absence of two years, as well as other Asian and European countries. They include 464 chicken, 326 and 173 domestic birds, 20 pets and six turkeys found within the one-kilometre radius of the bird flu epicentre. The Philippines has detected an outbreak of avian flu in a northern province after tests showed presence of the highly infectious H5N6 subtype of the influenza A virus at a quail farm, the country's agriculture secretary said on Monday. Germany has confirmed a case of H5N8 avian flu on a small poultry farm in Saxony – a state that borders Poland and Czech Republic.

Table 10: Summary for the user given query

Query	Original summary	Summary generated by the summarizer system used in this work
Corona virus	The Indian government has defended its handling of the coronavirus outbreak after a strict lockdown—introduced with little warning—left millions stranded and without food. India has been put in lockdown to halt the spread of the coronavirus outbreak. People have been told to stay indoors, but for many daily-wage earners this is not an option. The BBC's Vikas Pandey finds out how they were coping in the days leading up to Tuesday's announcement. India has been criticised for its poor record of testing people in the battle against coronavirus. That, however, is set to change, thanks in large part to the efforts of one virologist, who delivered on a working test kit, just hours before delivering her baby. A 68-year-old woman from Delhi has been confirmed as the second Indian to die from the coronavirus. With India now in a 21-day lockdown to prevent the spread of the coronavirus, there's been plenty of advice shared on how to prevent or cure the disease. "We have a simple message to all countries—test, test, test," World Health Organisation (WHO) head Tedros Adhanom Ghebreyesus told reporters in Geneva earlier this week.	The Indian government has defended its handling of the coronavirus outbreak after a strict lockdown—introduced with little warning—left millions stranded and without food. India has been put in lockdown to halt the spread of the coronavirus outbreak. India has been criticised for its poor record of testing people in the battle against coronavirus. A 68-year-old woman from Delhi has been confirmed as the second Indian to die from the coronavirus. With India now in a 21-day lockdown to prevent the spread of the coronavirus, there's been plenty of advice shared on how to prevent or cure the disease.

The summary generated for the actual google news is shown in Tab. 11. Here, the summary is generated from 2 different news article contents.

5 Performance Evaluation

5.1 News Crawler

The news collection time for different number of URLs using various crawlers is tabulated in Tab. 12.

The news collector is compared with different news crawler and is shown in Fig. 3. The news collector results indicate that the news collector is performing faster than other news collecting crawlers for any number of feed URLs. This is achieved with the parallel crawler which performs the news collection by sharing the URLs to multiple thread program to run parallelly.

Table 11: Summarization results

Actual Google news		News summary generated by summarizer system used in this work
News 1	News 2	
India Business News: Two employees working with IT companies Dell and Mindtree have been tested positive for coronavirus, according to company statements. The total number of novel coronavirus cases in the country touched 60 today, he health ministry said.	Two fresh cases were reported from Delhi and Rajasthan today. An 85-year-old man in Jaipur tested positive for the disease, a state government official said. Talking about the deadly outbreak of coronavirus, Kerala Health Minister KK Shailaja informed that those who are not revealing their travel history of coming from affected areas will be considered a crime.	The total number of novel coronavirus cases in the country touched 60 today, he health ministry said. Talking about the deadly outbreak of coronavirus, Kerala Health Minister KK Shailaja informed that those who are not revealing their travel history of coming from affected areas will be considered a crime.

Table 12: News crawling time for different set of feed URLs using various crawlers

Crawler	News collection time (s)			
	100 URLs	200 URLs	300 URLs	400 URLs
NewsTracker (Proposed system)	10	15	22	28
Mercator [30]	12	25	36	48
Focused news crawler [31]	14	24	38	54
Semantic web crawler [32]	18	28	41	62

5.2 News Retrieval Efficiency

The similar relevant keywords of the user given input are generated and the retrieval performance is evaluated. The news retrieval performance for direct user queries and relevance keywords is shown in Figs. 4 and 5 respectively.

5.3 Query Evaluation

The user queries are evaluated on pre-processed keyword indexing, non-pre-processed keyword indexing and non-indexing news contents. The query processing time is tabulated in Tab. 13. The comparison of query processing time for different indexing based retrieval is shown in Fig. 6.

5.4 Evaluation of Summarization

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics used for evaluating the summarization. It compares the summary against a set of references summary generated by human [33]. This quantitative of overlapping words is measured using the precision.

$$\text{Precision of reference} - \text{summary} = \frac{no.\ of\ overlapping\ words}{total\ words\ in\ reference\ summary}$$

$$\text{Precision of system} - \text{summary} = \frac{no.\ of\ overlapping\ words}{total\ words\ in\ system\ summary}$$

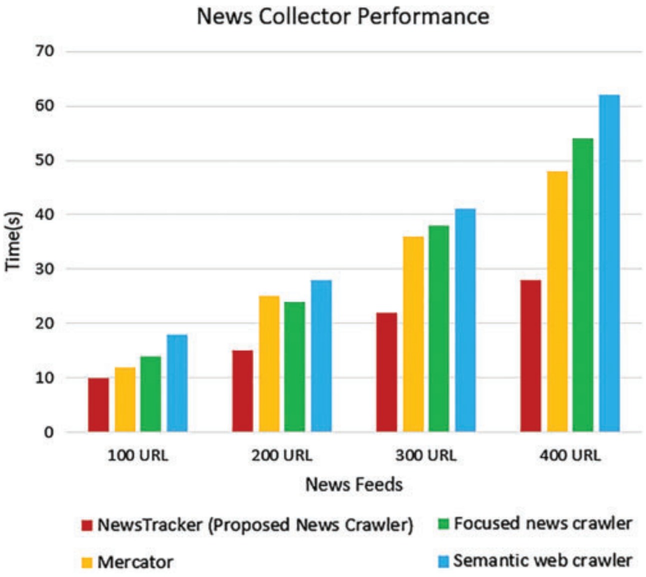

Figure 3: Comparison of different news crawlers

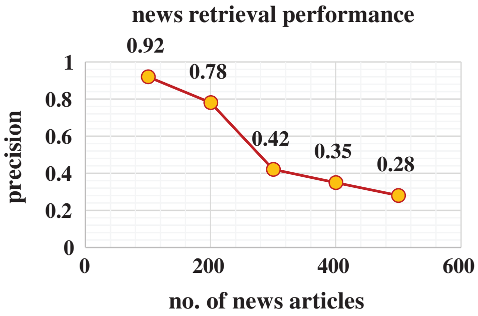

Figure 4: News retrieval performance for direct user queries

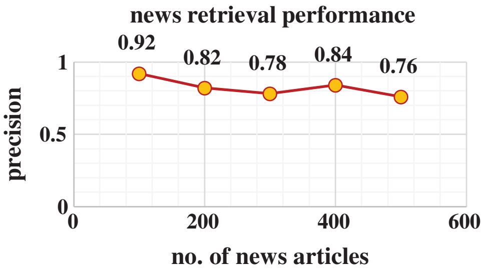

Figure 5: News retrieval performance for relevance keywords of user queries

Table 13: Query processing time using various indexed news contents

Queries	Pre-processed keyword) indexing (s	Non-pre-processed keyword indexing (s)	Non-indexing (s)
Corona virus	5	12	16
Avian flu	7	13	18
Delhi violence	4	9	15
Liver cancer	8	11	21

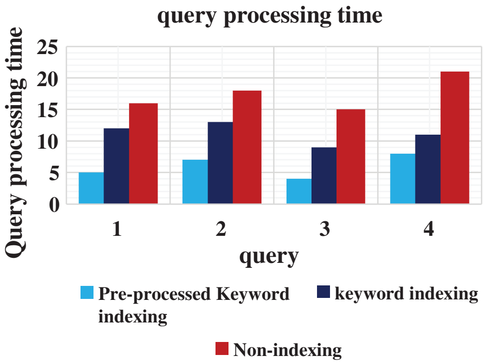

Figure 6: Query processing time for various indexing based retrieval

The precision of the automatic summarization is shown in Tab. 14. It means that the precision is 1.0 that is all the words in the reference summary is available in the automatic system summary. The precision calculated using the system summary is 0.88.

Table 14: Precision of automatic summarization

Original text	Automatic summary generator	Reference summary by human	Precision using reference summary	Precision using system summary	ROUGE-1	ROUGE-L
In the wake of novel coronavirus spread in India, the Delhi Metro services will remain completely closed, the Delhi Metro Rail Corporation (DMRC) declared.	Delhi Metro rail service completely closed till 31 March	Delhi Metro rail service closed till 31	7/7 = **1.0**	7/8 = **0.88**	8/9 = **0.88**	**0.78**

Further, we applied ROUGE specific metrics for effectively measuring the summary generation. The measures are ROUGE-N, ROUGE-S, ROUGE-L. These refers the size of the texts compared among the system summary and reference summary. ROUGE-1 refers the overlap of unigrams among the reference and system summaries. ROUGE-2 refers the overlap of bigrams among the reference and system summaries. ROUGE-1 and ROUGE-2 are the ROUGE-N type measures. It is referenced in the literature that ROUGE-1 and ROUOGE-L are appropriate for extractive summarization [34].

$$\text{ROUGE - N} = \frac{\sum_{S\in\{Reference\ Summaries\}}\sum_{gram_n\in S} Count\_match(gram_n)}{\sum_{S\in\{Reference\ Summaries\}}\sum_{gram_n} Count(gram_n)}$$

We have observed from the summarization evaluation that the ROUGE-N and ROUGE-L measures indicated that 88.88% and 77.77% of the actual news content is covered by the news summary generated. Since ROUGE-L needs to measure the longest sentence covered in the summary, the received value is a good measure that it has generated a summary covering the required sentences. The summarization performance of the proposed system is compared with other methodologies used in the literature for the summarization of document contents. The comparison result has been ensured with the ROUGE-1 metric which is the appropriate measure for news text summarization. The comparative results are tabulated in Tab. 15. It is observed from the result that the proposed system is highly useful for effectively summarizing the dynamically collected news data.

Table 15: Comparison of summarization performance

Summarization model	ROUGE-1
Variational auto encoder model [35]	0.608
Latent semantic analysis based topic summarization model [36]	0.540
LexRank based automatic summarization model [37]	0.484
NEWS summarization model (Proposed system)	0.880

5.5 Computational Complexity

The news data streams are received and the similarity needs to be estimated. The similarity computation involves the use of similarity matrix. It requires little large memory than other clustering algorithms since it needs to keep the data elements to store the matrix values.

$$\text{Space complexity} = O(n^2)$$

Even hierarchical clustering takes more space, it is widely used in many of content organization systems. The hierarchical clustering algorithms satisfy reducibility property. The increased computational time required for generating the clusters help in providing the hierarchy of cluster set with exact and unique structure with this reducibility property.

5.6 Scope and Application

Mainly, in this work, the automatic news summarization system for the dynamic news articles with timeframes from google news. The scope of the proposed collaborative filtering based news retrieval system includes concise information from various news articles. It helps to eliminate the difficulty of going through huge news articles and provides 20% to 30%from the original news content. The scope is limited to generate the summary for the user interested keyword using the news articles in a time frame. This news retrieval system helps in a better way for the online news content editors who are in need of accessing the interested domain content immediately.

6 Conclusion

In this paper, the hierarchical clustering based news summarization system has been proposed to apply on RSS feed based news and google news. The news crawler used thread based news crawling to collect the news articles effectively with better collection efficiency which has been compared with various state of the art news crawlers. This work used various recent domain corpuses to tag and extract the topic wise news efficiently. The hierarchical clustering handled the news contents by estimating the similarity and produced the hierarchical clusters of the various domains appropriately. The evaluation of the automatic summary with the human generated summary models proved that it performed maximum for the hierarchically clustered news article contents. Hence, proposed news summarization system is suitable and useful for the content readers who are keen in knowing recent domain specific news with the generated summary from various news sources.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] G. News, “Google news sources,” *News Portal*, 2002. [Online]. Available: http://news.google.com.

[2] C. Wang, X. He and A. Zhou, “Event phase-oriented news summarization,” *World Wide Web: Internet and Web Information Systems*, vol. 21, no. 4, pp. 1069–1092, 2018.

[3] D. Nagalavi and M. Hanumanthappa, “The NLP techniques for automatic multi-article news summarization based on abstract meaning representation,” *Proc. Emerging Trends in Expert Applications and Security*, vol. 841, pp. 253–260, 2019.

[4] J. Cha and P. K. Kim, “The automatic text summarization using semantic relevance and hierarchical structure of wordnet,” *Proc. Broad-Band Wireless Computing, Communication and Applications*, vol. 2, pp. 215–222, 2016.

[5] N. Araibi, E. B. Ahmed and W. K. B. Abdessalem, “IRORS: Intelligent recommendation of RSS feeds,” *Vietnam Journal of Computer Science*, vol. 3, no. 1, pp. 47–56, 2016.

[6] S. M. Alzahrani, “Building profiling analysing and publishing an arabic news corpus based on google news rss feeds,” *Asia Information Retrieval Symp. on Information Retrieval Technology*, vol. 8281, pp. 488–499, 2013.

[7] H. Alharthi and D. Inkpen, “Content-based recommender system enriched with wordnet synsets,” *Proc. Intelligent Text Processing and Computational Linguistics*, vol. 9042, pp. 295–308, 2015.

[8] F. G. Taddesse, J. Tekli, R. Chbeir, M. Viviani and K. Yetongnon, “Semantic-based merging of RSS items,” *World Wide Web: Internet and Web Information Systems*, vol. 13, no. 1, pp. 169–207, 2010.

[9] C. Bouras and V. Tsogkas, “Improving news articles recommendations via user clustering,” *International Journal of Machine Learning and Cybernetics*, vol. 8, no. 1, pp. 223–237, 2017.

[10] G. Xu, Z. Yu, C. Wang and A. Wang, “Research on topic discovery technology for web news,” *Neural Computing and Applications*, vol. 32, no. 1, pp. 73–83, 2020.

[11] Y. Diao, H. Lin, L. Yang, X. Fan, Y. Chu *et al.*, "CRHASum: Extractive text summarization with contextualized-representation hierarchical-attention summarization network," *Neural Computing and Applications*, vol. 32, no. 15, pp. 11491–11503, 2020.

[12] R. Katarya and Y. Arora, "Capsmf: A novel product recommender system using deep learning based text analysis model," *Multimedia Tools and Applications*, vol. 79, no. 48, pp. 35927–35948, 2020.

[13] A. Balahur, M. Kabadjov, J. Steinberger, R. Steinberger and A. Montoyo, "Challenges and solutions in the opinion summarization of user-generated content," *Journal of Intelligent Information Systems*, vol. 39, no. 2, pp. 375–398, 2012.

[14] M. Kozlowski and H. Rybinski, "Clustering of semantically enriched short texts," *Journal of Intelligent Information Systems*, vol. 53, no. 1, pp. 69–92, 2019.

[15] C. Long, M. L. Huang and X. Y. Zhu, "A new approach for multi-document update summarization," *Journal of Computer Science and Technology*, vol. 25, no. 4, pp. 739–749, 2010.

[16] R. Willis, "Taming the climate? corpus analysis of politicians' speech on climate change," *Environmental Politics Journal*, vol. 26, no. 2, pp. 212–23, 2017.

[17] F. Huang, S. Zhang, M. He and X. Wu, "Clustering web documents using hierarchical representation with multi-granularity," *World Wide Web: Internet and Web Information Systems*, vol. 17, no. 1, pp. 105–126, 2014.

[18] N. Kumar, V. Poonia, B. B. Gupta and M. K. Goyal, "A novel framework for risk assessment and resilience of critical infrastructure towards climate change," *Technological Forecasting and Social Change*, vol. 165, pp. 1–12, 2021. [Online]. Available: https://doi.org/10.1016/j.techfore.2020.120532.

[19] S. R. Sahoo and B. B. Gupta, "Multiple features based approach for automatic fake news detection on social networks using deep learning," *Applied Soft Computing*, vol. 100, pp. 1–16, 2021. [Online]. Available: https://doi.org/10.1016/j.asoc.2020.106983.

[20] S. Jha, M. K. Goyal, B. Gupta and A. K. Gupta, "A novel analysis of COVID 19 risk in India incorporating climatic and socioeconomic factors," *Technological Forecasting and Social Change*, vol. 167, pp. 1–11, 2021. [Online]. Available: https://doi.org/10.1016/j.techfore.2021.120679.

[21] A. Mishra, N. Gupta and B. B. Gupta, "Defense mechanisms against DDoS attack based on entropy in SDN-cloud using POX controller," *Telecommunication Systems*, vol. 77, no. 1, pp. 47–62, 2021.

[22] M. S. Pera and Y. K. D. Ng, "Using maximal spanning trees and word similarity to generate hierarchical clusters of non-redundant RSS news articles," *Journal of Intelligent Information Systems*, vol. 39, no. 2, pp. 513–534, 2012.

[23] D. Hunt and K. Harvey, "Health communication and corpus linguistics: Using corpus tools to analyse eating disorder discourse online," *Corpora and Discourse Studie, Palgrave Advances in Language and Linguistics*, vol. 2, pp. 134–154, 2015. [Online]. Available: https://doi.org/10.1057/9781137431738_7.

[24] Google News Data Source, "Topics, locations and sources," *Google News*, 2002. [Online]. Available: https://news.google.com/topstories?hl=en-IN&gl=IN&ceid=IN:en.

[25] A. T. Patanasorn, "Constructing an academic word list of business English: A corpus-based approach," *Humanities and Social Sciences Journal*, vol. 34, no. 2, pp. 1–31, 2017.

[26] M. Callies, "Corpora of sports commentaries," *Text Mining and Applications*, 2015. [Online]. Available: https://mailman.uib.no/public/corpora/2015-February/022099.html.

[27] A. Mozaffari and R. Moini, "Academic words in education research articles: A corpus study," *Proc.-Social and Behavioral Sciences*, vol. 98, no. 6, pp. 1290–1296, 2014.

[28] T. Wikipedia corpus, "English corpus from wikipedia," *Wikipedia*, 2014. [Online]. Available: https://www.english-corpora.org/wiki/.

[29] A. Nenkova and L. Vanderwende, *The Impact of Frequency on Summarization*, Microsoft Research, Redmond, 2005. [Online]. Available: https://www.cs.bgu.ac.il/~elhadad/nlp09/sumbasic.pdf.

[30] K. Ahmad, "A new mercator web crawler," in *Proc. Recent Trends in Engineering & Technology*, pp. 111–117, 2012. [Online]. Available: https://www.researchgate.net/publication/262840418_A_New_Mercator_Web_Crawler.

[31] S. Y. Yang, "A focused crawler with ontology-supported website models for information agents," in *Proc. Grid and Pervasive Computing*, vol. 6104, pp. 522–532, 2010.

[32] M. Kumar and R. Vig, "Term-frequency inverse-document frequency definition semantic (TIDS) based focused web crawler," in *Proc. Computing and Communication Systems*, vol. 270, pp. 31–36, 2020.

[33] Lin and C. Yew, "Automatic evaluation of summaries using n-gram co occurrence statistics," *Language Technology Conf*., 2003. [Online]. Available: http://aclweb.org/anthology/N/N03/N03-1020.pdf.

[34] Lin and C. Yew, "ROUGE: A package for automatic evaluation of summaries," *Workshop on Text Summarization*, 2004. [Online]. Available: http://www.aclweb.org/anthology/W/W04/W04-1013.pdf.

[35] N. Alami, N. En-nahnahi, S. A. Ouatik and M. Meknassi, "Using unsupervised deep learning for automatic summarization of arabic documents," *Arabian Journal for Science and Engineering*, vol. 43, pp. 7803–7815, 2018.

[36] I. V. Mashechkin, M. I. Petrovskiy, D. S. Popov and D. V. Tsarev, "Automatic text summarization using latent semantic analysis," *Programming Computer Software*, vol. 37, no. 6, pp. 299–305, 2011.

[37] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, vol. 22, no. 1, pp. 457–479, 2004.