IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 March 2024, accepted 12 April 2024, date of publication 22 April 2024, date of current version 24 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3392252

RESEARCH ARTICLE

A Probabilistic Approach for Extractive Summarization Based on Clustering Cum Graph Ranking Method

AMREEN AHMAD[1], TANVIR AHMAD[2], SARFARAZ MASOOD[2], MOHD. KHIZIR SIDDIQUI[3], BASMA ABD EL-RAHIEM[4], (Member, IEEE), PAWEL PLAWIAK[5,6], AND FAHAD ALBLEHAI[7], (Member, IEEE)

[1]Galgotia College of Engineering and Technology, Greater Noida 201310, India
[2]Department of Computer Engineering, Jamia Millia Islamia, Jamia Nagar, New Delhi 110025, India
[3]Department of Computer Engineering, BITS Pilani, Rajasthan 333031, India
[4]Department of Mathematics and Computer Science, Faculty of Science, Menoufia University, Shebin El-Koom 32511, Egypt
[5]Department of Computer Science, Faculty of Computer Science and Telecommunications, Cracow University of Technology, 31-155 Krakow, Poland
[6]Institute of Theoreticl and Applied Informatics, Polish Academy of Sciences, 44-100 Gliwice, Poland
[7]Computer Science Department, Community College, King Saud University, Riyadh 11437, Saudi Arabia

Corresponding authors: Amreen Ahmad (amreen.ahmad10@gmail.com) and Pawel Plawiak (pawel.plawiak@pk.edu.pl)

This work was supported by King Saud University, Riyadh, Saudi Arabia, through the Researchers Supporting Project RSPD2024R564.

ABSTRACT Online information has increased tremendously in today's age of the Internet. As a result, the need has arisen to extract relevant content from the plethora of available information. Researchers are widely using automatic text summarization techniques for extracting useful and relevant information from voluminous available information. The summary obtained from the automatic text summarization often faces the issues of diversity and information coverage. Earlier researchers have used graph-based approaches for ranking and optimization. This research work introduces a probabilistic approach named as ClusRank for summary extraction, comprising of a two-stage sentence selection model involving clustering and then ranking of sentences. The initial stage involves clustering of sentences using a proposed overlapping clustering algorithm on the weighted network, and later selection of salient sentences using the introduced probabilistic approach. In the analysis of real-world networks, community structure development is essential because it provides strategic insights that help decision-makers make well-informed choices. Furthermore, methodologically strict community detection algorithms are required due to the occurrence of discontinuous, overlapping, and nested community patterns in such networks.This research work, an algorithm is presented for detecting overlapping communities based on the concept of rough set and granular information on links. The sentence selection algorithm based on budget maximum coverage approach supports the assumption that larger sub-topics in a document are of more importance than smaller subtopics. The performance of the proposed probabilistic ClusRank is validated on DUC2001, DUC 2002, DUC2004, and DUC 2006 data sets.

INDEX TERMS Automatic text summarization, clustering, graph ranking, diversity, information coverage.

I. INTRODUCTION

Recently, multi-document summarization and document clustering have gained a lot of attention for analyzing textual information. The aim of document clustering is to divide a set of documents into distinct classes called clusters, where similar documents are occurring in the same cluster and dissimilar documents [1] occur in different clusters. Another efficient technique for extracting relevant content from huge voluminous data is multi-document summarization, which aims to create a reduced summary while retaining the gist of the documents [2], [3]. Both these techniques find a range of applications in information retrieval and management, apart from retrieving relevant content from documents. Results

The associate editor coordinating the review of this manuscript and approving it for publication was Yilun Shang.

of web search [4] can be organized and presented in an efficient way using document clustering. Apart from it, multi-document summarization finds its use in the creation of snippets on the Web, that can be further used for future purposes [5].

Usually, researchers while performing document clustering consider the set of documents as a document term matrix where documents are represented by rows and terms are represented by columns. Several conventional clustering algorithms exist for grouping similar documents. But these algorithms were unable to correctly interpret the document cluster. In recent years, some works [6] have focused on capturing the dual knowledge between documents and terms and performing clustering at the same time. But this framework has limitations, since every document cluster is represented by a set of representative words, and these words cannot give a true interpretation of clusters as they lack contextual and semantic information. Another way could be a selection of salient sentences from the cluster.

Multi-document summarization(MDS) aims at producing a condensed version of the document while retaining the girth of the document. Based on the output type, MDS can be categorized into abstractive or extractive. In extractive summarization, representative sentences are selected as summaries from the original source document. This approach uses some pre-defined methodology to compute sentence scores and based on that salient sentences are selected. Whereas in abstractive summarization, the summary is composed of salient sentences that represent the gist of the document but are not part of the original document. Some techniques such as reformulation, sentence compression, and information fusion are involved in abstractive summarization. The summary obtained from abstractive summarization is compact since it uses deep learning techniques.

Since abstractive summarization requires extensive NLP techniques, this has diverted the attention of the research community towards extractive summarization. Different techniques for extraction are:

- Elimination of Redundancy - length limitation is one of the constraints for an effective summary. The summary generated from extractive summarization contains similar information, duplicacy can be eliminated using some similarity measure.
- Coherency - the biggest challenge is determining the best sequence for recovered sentences to construct a cohesive context flow. There are two tasks in MDS - (a) using corpus information for learning sentence natural sequence, and (b) usage of the chronological ordering of sentences.
- Coverage - a critical role is played by coverage in text summarization. It mainly focuses on extracting information that covers diverse topics from the source document. Different algorithms have been proposed for coverage at different levels such as text, phrase, word, paragraph, and sentence.

This research work concentrates on extractive summarization since it is more practical and feasible. In MDS, the set of documents are represented as a sentence-term matrix where rows and columns are represented as sentences and terms. Numerous clustering approaches [7], [8] have been proposed for extractive summarization where the first step is cluster generation and then identification of salient sentences from these clusters. The limitation of such an approach is that they consider sentences as independent units and contextual dependency among sentences are ignored. However, the mutual influence of sentences occurring within same cluster should be considered for correct interpretation of cluster.

A. MOTIVATION

Two interpretations of communities inside complex networks are possible: a node-centric cluster of nodes that are strongly coupled or a collection of intricately connected linkages (link-centric).In the past decades researchers have focused on node centric community detection but less work has been done on link centric community detection. Furthermore, there are still a number of issues with modern community detection methods that limit their practical uses. Most of the current techniques allocate every node to a single community, resulting in disconnected communities. Certain algorithms are specialized to a given area, while others necessitate prior knowledge regarding the count of communities within a complicated network. Certain algorithms exhibit scalability issues when applied to big networks, while others remain impervious to fluctuations in community size. Finding overlapping communities—which are more representative of the interconnected world of today—is one of the main problems with community recognition in complicated networks nowadays. Generating summary within required words from these overlapping, nested and disjoint cluster of sentences will be more informative.

B. RESEARCH GAP

- In addition to having overlapping communities, a network's community structure may also feature nested communities, in which one community is enclosed within another. For instance, a location-based community may contain a number of ethnic communities. Still, work needs to be done.
- Since links are more unique than nodes, it is preferable to cluster links rather than nodes to find persistent overlaps in community structure. Research work incorporating link concepts are less.

C. RESEARCH QUESTIONS

- Real-world network is overlapping, disjoint, and nested. Develop an approach that can detect all such types of communities.
- The idiosyncratic nature is observed more in links in comparison to nodes. Hence, link clustering can be better approach for developing community detection

algorithm. How to use the concepts of link in community detection?
- How to identify salient sentences within limit from the overlapping, disjoint and nested cluster of sentences?

The essential distinction between nodes and links' properties inspired us to create a novel link-centric community detection technique, that can adapt to unique aspects of complex networks' community structures while having the ability to identify overlapping, nested, and discontinuous groups that coexist.

The focus of the proposed methodology in this paper is to convert a document into an appropriate graph structure, cluster it into overlapping, distinct, and nested communities, and extract out the most important sentences of the document. The uniqueness of the proposed research work is highlighted below:

D. CONTRIBUTIONS

- The given textual document is converted into a graph. An algorithm is developed to detect overlapping, disjoint and nested communities in a weighted network with feasible computational requirements(in terms of time complexity).
- Till date, no research work has been conducted to identify communities within complex network on the basis of link connectivity using rough set concept. The idea of mutual link reciprocity is introduced in order to determine link similarity inside a complex network. Mutual link reciprocity is used to execute the restricting and merging of link subgroups in each replication.
- In accordance with the underlying theory that links—rather than nodes—are more distinctive than relationships. Communities are defined as sets of strongly connected links, and an approach is suggested for detecting communities that is based on a rough grouping of links.
- It is possible to identify discontinuous, nested, and overlapping communities when link-based rough clustering and mutual link reciprocity are used together. The suggested algorithm is proven to be feasible by experimental and comparative assessment of real-world networks.
- Summaries are generated within limit using budgeted maximum coverage problem. It has been found that larger sub-topics in a document carry more importance than the smaller subtopics. The proposed probabilistic ClusRank method picks sentences depending on the weightage of subtopics.
- We conduct experiments on the standard DUC-2001, DUC-2002, DUC-2004, and DUC-2006 data sets to validate the performance of the proposed ClusRank method.

The further sections are divided as follows. Section II introduces an overview of the work done in this area and connected relevant fields followed by a brief introduction of the problem statement given in Section III. Section IV discusses a step-by-step procedure for the proposed methodology. Section V discusses the explanation of the proposed methodology using a toy network example. Section VI describes in detail the analysis of the experiments followed by the conclusion in Section VII.

II. RELATED WORK

Extractive text summarization has been in active developments in recent years, numerous methods have been proposed to solve the problem. The key idea lies in developing an efficient scoring method for the sentences in the document. Many methods apply topic-wise clustering on the document and identify key individual sentences with respect to the topics. Some other approaches revolve around using evolutionary algorithms [9] or machine learning techniques [10], [11].

A. GRAPH BASED APPROACHES

The method proposed in [12] simultaneously clusters the sentences and scores them for a ranking. The focus of work in [13] is to cluster the sentences and perform their selection as a solution to an optimization problem. Reference [14] targets both diversity and coverage of the summary using an integrated clustering-based technique. The idea in [15] and [16] uses the co-clustering method on words and sentences individually to perform a topic-based summarization. The framework introduced allows words to have an explicit decision in sentence selection to squeeze out better performance. Reference [17] proposes a fuzzy c-medoid-based clustering approach to produce a cluster of sentences similar to a sub-topic of the topic. A tool named Compendium is proposed in [18], which combines textual entailment, statistical and cognition-based techniques to remove redundant information and find relevant content in summary. The work in [19] focuses on probabilistic modelling topic relevance and coverage in summarization. In [20], the authors use fractal theory to infer the interplay of sentences and perform the summarization.

The work in [21] uses a graph-based approach to cluster the sentences. The document is modelled as a graph and then different methods are used to rank the sentences (modelled as nodes). Reference [22] propose a semi-supervised clustering method on the graphs combined with topic modelling. In [23], they use the ideas of graph matching to improve upon the results. Reference [24] proposes *Collabsum*, which exploits information from multiple documents by clustering them and extracting the mutual influence to summarise a single document. This methodology incorporates both intra-document and inter-document relationships. Reference [25] models summarization as a modified p-median problem. The work in [26], uses external knowledge from Wikipedia to enhance performance on the existing graph-based methods. In [27], *LexRank* method is introduced, it determines the salience using eigenvector centrality in graphical representation. The work in [28] models the documents as graphs and

exploits mutual information between documents to generate a summary.

A separate line of work has evolved in community detection and influential node identification. Reference [29] introduces *TOPSIS* to identify influential nodes by considering it as a multi-attribute decision-making problem. Reference [30] proposes *NDOCD* where links are iteratively removed to reduce the graph into clusters. In [31], network embedding is used to decompose the network into communities and then nodes are chosen to maximize influence. Reference [32] propose *HWSMCB*, where various degrees are considered to choose influential nodes in a network as a influence maximization problem. The authors in [33] introduce an algorithm for overlapping community detection based on granular information of links and concepts of rough set theory.

Reference [34] used knowledge of events for creating a summary for MDS. In this, a document representation technique was created to extract and sieve the information about the events mentioned in the text. A combined approach incorporating machine learning and rule-based models was proposed that used event information at the sentence level and assessed the temporal relationships between them. Event graphs were used to estimate graph kernels, that was further used for measuring the similarity between queries and documents. Reference [21] developed a MDS approach for generating a summary based on statistics and linguistic measures. A sentence selection approach was presented that was efficient in removing redundancy and maximizing coverage. A domain-independent framework was developed for extracting summary by [35]. A set of rules was designed for categorizing the text of the source document. This approach was applicable for both extractive and abstractive summarization.

Typically, overlapping communities are detected using the Clique Percolation Method (CPM) [30], which makes the assumption that communities are made up of overlapping full sub graphs and it searches these sub graphs to identify the community structure. Nevertheless, when it comes to large-scale networks, it has been observe that CPM is only effective in networks with highly connected subgroups and is unable to identify the community structure. A straightforward yet effective technique for community detection that may identify community structure in almost linear time is the label propagation algorithm [11]. The very inconsistent nature of community detection results is a major flaw in this technique, though. The overlapping community detection is expanded upon by researchers, who suggest SLPA, BMLPA, and COPRA. By introducing additional label expressions, these techniques enable a node label to have several community identifiers. In order to assess the quality of community structure, Newman [34] proposed modularity (Q), which is also regarded as an optimization goal in the context of heuristic community detection techniques. Shen et al. developed an approach called EAGLE based on both extended modularity (EQ) and maximal cliques for overlapping community discovery. To be more precise, EAGLE creates initial communities by determining the maximal cliques, then grows these communities by combining communities that are comparable in order to optimize extended modularity. An approach based on coalition formation games is proposed by Zhou et al. [9]. In these games, participants cooperate with one another to establish coalitions in an effort to increase the group's score. Avrachenkov and colleagues [36] present a pair of cooperative game-theoretic models for the purpose of community detection. Some authors [16], [55], [56] have employed graph based neural approaches for community detection but they were capable of detecting overlapping communities only. While some have [55], [56] used concept of graph attention neural network that was capable of detecting only overlapping communities.

B. OTHER APPROACHES

In [9], a mematic algorithm MA-Single DocSum is introduced, the method uses evolutionary algorithm to solve the extractive summarization as a binary optimization problem. Reference [36] propose a hierarchical selective encoding network for both sentence-level and document-level representations and data containing important information is extracted. The method introduced in [37] improves upon the cohesiveness of the summaries generated by the extractive summarization systems. It is based on a post-processing step that binds dangling co-reference to the most important entity in a given co-reference chain. In [38], the approach selectively removes unimportant sentences until a desired compression score is achieved. The work by [39] models the document as a semi-graph to extract both linear and non-linear relationships between the features. In [40], a weighted graphical representation of the document is formed and coherence, non-redundance, and importance are optimized using ILP (Inductive logic programming). Reference [10] uses algorithms based on Latent Semantic Analysis to summarize Turkish and English text. Reference [11] proposes a deep-learning method to perform unsupervised summarization. The researchers in [41] use Langragian relaxation to solve summarization as a combinatorial problem.

Several unsupervised algorithms have also been introduced to cluster the sentences and rank them. The methods use k-means [42] or fuzzy c-means [43] due to their good generalization performance in other tasks also. Fuzzy c-means is not robust to noises and is sensitive to outliers in Euclidean distance. In [44], the method uses a support vector machine (SVM) to train a summarizer using features like sentence position, sentence centrality, sentence similarity, and several more. Reference [45] uses sentence regression to score and greedily selects them to form the summary. In [46], the authors train an ensemble of SVMs over gram overlap, LCS, WLCS, skip-bigram, gloss overlap,BE overlap, length of sentence, the position of the sentence, NE, cue word match, title match to approach the problem.

C. DIFFERENCE BETWEEN PREVIOUS APPROACHES AND THE PROPOSED APPROACH

There exists difference between the proposed community detection algorithm and previous modularity based community detection algorithms such as Louvain, Girvan-Neuman etc. The differences are:

- Till date, no weighted clustering algorithm was developed for detection of overlapping, disjoint, and nested communities. Louvain was able to detect only non-overlapping communities whereas in Girvan-neuman entire community can be left out if edge split occurs at early stage.
- Apart from clustering documents, it is simultaneously assigning weights to sub-topics, that played a crucial role in determining salient sentences.
- A probabilistic approach is introduced named as ClusRank that ranks the salient sentences within a cluster using bmcp concept.

III. PROBLEM STATEMENT

Given a document D consisting of set of sentences $D = \{s_1, s_2, s_3, \ldots s_n\}$, where n denotes the number of sentences in the document and s_j is the *j-th* sentence, $1 \leq j \leq n$. The aim of of extractive summarization is to find a subset $D_s \subset D$ which contains different important topics mentioned in the complete document. It is expected to have $|D_s| \ll |D|$ where $|.|$ represent the number of sentences in the set.

A document consists of vast information covering various sub-topics and a common main theme connecting them. Coverage means that the summary extracted by the algorithm should cover most of the subtopics. Poor coverage of sub-topics is indicated by the absence of some relevant sentences. While extracting the sentences, deciding the importance based on relevance alone can be misleading and ignoring lesser covered but important sub-topics. Therefore, focus of the algorithm on both relevance and coverage is necessary.

IV. METHODOLOGY

A. GRAPH CONSTRUCTION

The document is split into sentences. Let the graph formed be $G = (V, E)$ where elements in set V are the sentences and E represents the set of edges between a pair of sentences. The presence of an edge between a pair of sentence is decided by a weighted sum of their statistical and semantic similarities. A hyper parameter, δ_e is chosen and if a pair has similarity lesser than δ_e, no edge exists between them. Clearly, a high value of δ_e encourages lesser number of edges and a lower value will include all the $\binom{V}{2}$ edges in the graph. An appropriate threshold will retain relations between important sentences and discard edges between insignificant sentences. This workflow is represented in the figure 1.

B. COMMUNITY DETECTION

The graph $G = (V, E)$ formed, is a weighted graph, let $w_{i,j}$ represent weight of edge $e_{i,j}$ between node v_i and node v_j.

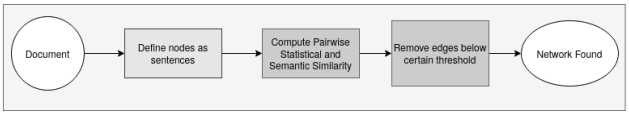

FIGURE 1. Flow diagram showing construction of graph-network from a given document.

Additionally, it satisfies the condition: $w_{i,j} \geq \delta_e$ The figure 2 shows the flow diagram for this section of algorithm. The algorithm introduces the terminologies mentioned below:

1) **FOAN: First Order Approximate Neighbors -** First Order Approximate Neighbors of a link $e_{i,j}$ are defined by:

$$F(e_{i,j}) = \{e_{i,k} : k \in N_i\} \cup \{e_{k,j} : k \in N_j\} \quad (1)$$

where N_i and N_j represent the nodes connected to v_i and v_j.

2) **SOAN: Second Order Approximate Neighbors -** Second Order Approximate Neighbors of a link $e_{i,j}$ are defined by:

$$S(e_{i,j}) = \bigcup \{F(e_{m,n}) : (m, n) \in F(e_{i,j})\} \quad (2)$$

3) **JS: Jaccard Similarity -** Jaccard Similarity between two vectors **x** and **y** is given by:

$$J_{\mathcal{W}}(x, y) = \frac{\sum_i min(x_i, y_i)}{\sum_i max(x_i, y_i)} \quad (3)$$

4) **CSOAN: Constrained Second Order Approximate Neighbors -** Constrained Second Order Approximate Neighbors of a link $e_{i,j}$ are defined by:

$$C(e_{i,j}) = \{e_{a,b} : e_{a,b} \in S(e_{i,j}) | J_{\mathcal{W}}(\overline{e}_{i,j}, \overline{e}_{a,b}) \geq \delta_{csoan}\} \quad (4)$$

5) **LNS: Link Node Set -** Link Node Set of a link $e_{i,j}$ is defined by:

$$L(e_{i,j}) = \{v_m, v_n : e_{m,n} \in C(e_{i,j})\} \quad (5)$$

So, $L(e_{i,j}) \subset V$ whereas $C(e_{i,j}) \subset E$, $S(e_{i,j}) \subset E$, and $F(e_{i,j}) \subset E$.

6) **Conductance -** Conductance of a graph $G = (V, E)$ is given by:

$$\phi(G) = \min_{S \in V; 0 \leq a(S) \leq a(V)/2} \frac{\sum_{i \in S; j \in \overline{S}} a_{i,j}}{a(S)} \quad (6)$$

The algorithm processes the graph of a document through several steps iteratively unless a stable set of edges is obtained. A set of first-order approximate neighbors is formed for every edge in the graph, using the equation 1. Next, a set of second-order approximate neighbors is formed by the union of first-order approximate neighbors of every edge in the first-order approximate neighbors of the target edge (equation 2). The cardinality of this set determines the number of nodes that share strong similarities with the nodes of a given edge. The sum of the weights of the elements

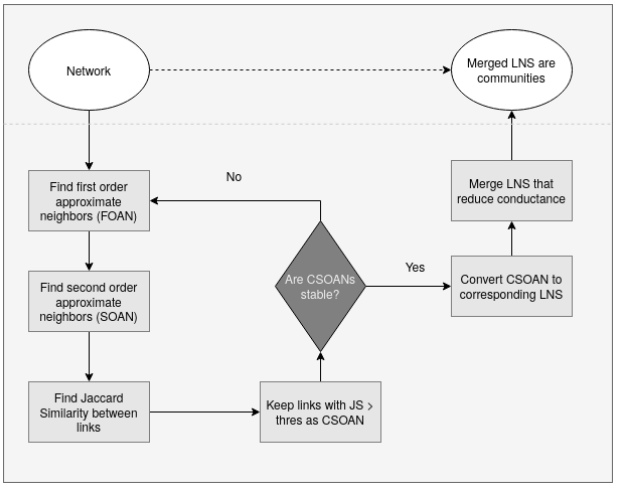

FIGURE 2. Flow diagram showing detection of communities by the algorithm.

in $S(e_{i,j})$ is higher for an edge with greater importance in the graph. Every set $F(e_{i,j})$ can be represented as vector with orthogonal components corresponding to the weight of edge elements. So every edge $e_{i,j}$ has a $S(e_{i,j})$ which can be expressed as a vector $\overline{e}_{i,j}$, given by equation 7.

$$\overline{e}_{i,j} = \sum_{t \in S(e_{i,j})} w_t \cdot \hat{t} \tag{7}$$

For every pair of edges in the graph $e_{i,j}$ and $e_{m,n}$, Jaccard Similarity is calculated between their corresponding vectors $\overline{e}_{i,j}$ and $\overline{e}_{m,n}$ using equation 3. A higher coefficients for an edge indicates greater similarity and higher importance than the other edges. The coefficients found are used to filter out edges with low similarity in $S(e_{i,j})$. A threshold δ_{cson} is used to calculate the constrained second order approximate neighbors $C(e_{i,j})$ using equation 4. The set $C(e_{i,j})$ is used as $F(e_{i,j})$ in the next iterative step (if need be). The loop stops processing an edge $e_{i,j}$ when the $C(e_{i,j})$ for a step is same as $C(e_{i,j})$ in previous iteration, and the set $C(e_{i,j})$ is deemed stable.

When stable sets of $C(e_{i,j})$ for every edge $e_{i,j}$ are found out, loop completely terminates. Every set $C(e_{i,j})$ is used to make the corresponding link node set $L(e_{i,j})$ using equation 5. Next, conductance of every $L(e_{i,j})$ is calculated using equation 6. A pair of $L(e_{i,j})$ and $L(e_{m,n})$ are merged if the resultant set $L(e_{i,j}) \cup L(e_{m,n})$ has a conductance lower than the individual sets. $L(e_{i,j})$ are merged (union) until conductance can no longer be reduced. The resultant set of node sets corresponds to the communities detected.

The computed pair-wise compatibility between connections is aided by the use of mutual link reciprocity following each iteration of the suggested procedure. The intuitive premise that any two entities in the real world are regarded as comparable is the basis for the idea of mutual link reciprocity, if there are more entities that are frequently connected to them than there are entities that are just connected to each of them. Prior research has employed a metric similar to mutual link reciprocity to assess the degree of connectivity among nodes within a complex network. Given that the set difference operates in a non-associative manner and that reciprocity between two connections should be symmetric. In the denominator, at least two set differences are considered. First linkage upper estimates are limited on each cycle to remove weakly related links and the confined link upper approximations exhibit substantial association linkages remaining. Until each CSOAN reaches stability and no further FOAN filtering is feasible, this process is repeated. The last stage of the suggested algorithm is to fine-tune the convergent CSOAN. In order to identify accurate and significant communities, redundancy in the node subsets is eliminated during the fine-tuning phase. Lowest conductance of subset of nodes are selected from given subset of redundant nodes. In order to reduce the conductance of the matching subset, the remaining nodes are inserted one at a time. In the input network, community subsets are found during the fine tuning process. Depending on the network structure, these subsets may be discontinuous, nested, or overlap.

1) ANALYSIS OF PROPOSED COMMUNITY DETECTION

For assessing the quality of overlapping communities, this research work uses Extended Modularity [45] metric for comparison with these baseline community detection algorithms:

- Community detection based on modularity and fuzzy logic [45] - This approach is based on a combination of modularity and fuzzy logic. It is efficient in determining overlapping communities in considerably lesser time.
- Edge betweenness [45] - Uses edge betweenness concept for detecting communities.
- Label propagation [45] - This approach assigns a node to a particular community based on the count of its neighbors occurring in that community.

Data sets

- Zachary's karate club [45] - Depicts a friendship network of karate club involving 34 members.
- Les Miserables [45] - It represents characters' network occurring in the novel Les Miserables.
- US politics [45] - A network of books concerned with US politics.

Analysis of proposed community detection algorithm

The proposed community detection algorithm is compared with some other methods such as label propagation, edge betweenness, and community detection algorithm based on modularity and fuzzy logic on benchmark data sets mentioned above. As can be seen from Table 1, the extended modularity value obtained for the proposed community detection algorithm is highest in comparison to other baseline methods on three standard data sets.This is due to the fact that in real-world, communities are often either disjoint, overlapping or nested. The proposed method is able to detect all such types of communities due to which it is showing

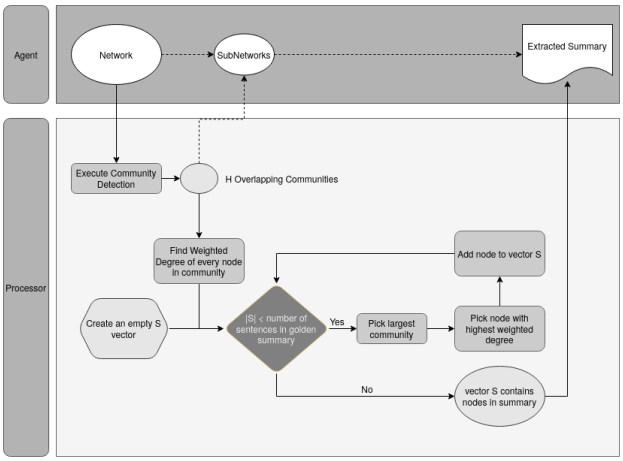

FIGURE 3. Flow diagram showing extraction of important sentences in the document using the proposed probabilistic ClusRank method.

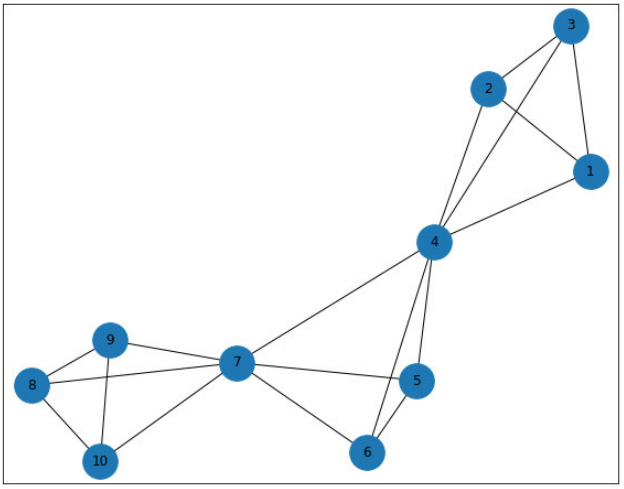

FIGURE 4. Flow diagram showing detection of communities by the algorithm.

good performance. This validates the efficacy of the proposed community detection algorithm.

C. FINDING IMPORTANT SENTENCES

Finding most significant sentences in document network is equivalent to finding most influential nodes in a network. Let the communities detected by the following algorithm in section IV-B are $\mathcal{H}$. The sub-graphs formed using $G = (V, E)$ and nodes in $H_i \in H$ are overlapping in nature, hence the influence of a node in graph G is determined by the influence in H_i and also by $H_j \in \mathcal{H}, j \neq i$.

For every sub-graph using nodes of $H_i \in \mathcal{H}$, the weighted degree of each node is calculated. A larger weighted degree of a node signifies a larger influence in the community. Additionally, a larger community is responsible for a larger influence in the graph. So, the algorithm picks the largest community and the node with largest weighted degree. This node is removed from the community and again the algorithm picks. The step mentioned is iteratively done until a desired number of nodes are extracted. Figure 3 shows the flow diagram for this section of the algorithm. where N_i and N_j represent the nodes connected to v_i and v_j.

V. EXPLANATION OF THE PROPOSED METHODOLOGY USING EXAMPLE NETWORK

To further illustrate the process undertaken by the algorithm, a graph $G = (V, E)$ is taken as given in figure 4. To assign the weights to this example network, degree of the nodes is calculated. A parameter $\theta = \frac{E}{V}$ is used as threshold. For every link, if either of nodes have a degree greater than θ, weight is drawn from a random number generator (between 2 and 10), available through the *random* package in python 3, otherwise the degree is set to unity. The weight of each link for the example network thus found is mentioned in the table 2. Table 3 presents the index mapping of the edges of the example network given in figure 4.

A. PART A: COMMUNITY DETECTION

1) STEP 1: FIND FIRST ORDER APPROXIMATE NEIGHBORS (FOAN)

FOAN of every edge in the network is calculated using the equation 1. The table 4 shows the calculated FOAN for the example network given in Fig.4.

2) STEP 2: FIND SECOND ORDER APPROXIMATE NEIGHBORS (SOAN)

SOAN of every edge in the network is calculated using the equation 2. The table 5 shows the calculated FOAN for the example network(given in Fig.4).

3) STEP 3: FIND JACCARD SIMILARITY

For every pair of $F(e_{i,j})$ corresponding vectors $\bar{e}_{i,j}$ are calculated and Jaccard similarity is found out using the equation 3. The Jaccard Similarity for the given example graph(shown in Fig.4) is given in table 6.

4) STEP 4: FIND CONSTRAINED SECOND ORDER APPROXIMATE NEIGHBORS

The algorithm uses the Jaccard Similarity found in the previous step to eliminate weaker relations in SOAN. Using a threshold $\delta_{cson} = 0.5$ and $\alpha_{decay}^{t=1} = 0.8$, CSON (table 7 is found. This completes the first iteration. The CSON acts as FOAN for the next iteration. The loop stops when $CSON_{new}$ is same as $CSON_{prev}$.

5) STEP 5: MERGING LINK NODE SETS (LNS)

After 4 iterations, a stable set of CSON is formed. The CSOAN of each edge does not change in any further iteration. CSON are converted into LNS - Link Node Set using equation 5. For every set of LNS, conductance is calculated and two LNS are merged (union) if and only if the resultant set has a lower conductance than the individual values. The

TABLE 1. Comparative analysis of proposed community detection algorithm with other methods on the basis of extended modularity.

Method	Zachary Karate	US Politics book	Les Miserables
Proposed Community detection	0.71	0.59	0.60
Modularity and Fuzzy logic	0.68	0.56	0.59
Edge betweenness	0.40	0.52	0.52
Label Propagation	0.42	0.51	0.51
Louvain	0.54	0.51	0.56
Girvan Neuman	0.65	0.53	0.59

TABLE 2. Weights of the edges in the example graph(given in Fig.4).

Edge	Weight	Edge	Weight	Edge	Weight	Edge	Weight
(1, 2)	9	(1, 3)	6	(1, 4)	5	(2, 3)	7
(2, 4)	2	(3, 4)	5	(4, 5)	2	(4, 6)	2
(4, 7)	5	(5, 6)	2	(5, 7)	8	(6, 7)	5
(7, 8)	9	(7, 9)	3	(7, 10)	9	(8, 9)	2
(8, 10)	4	(9, 10)	6	-	-	-	-

values of stable CSON, LNS and their conductance is given in table 7.

After merging the LNS, the communities thus found are given in table 9.

By using proposed community detection algorithm, five communities are identified A, B, C, D,and E. Community A and B are disjoint with community E, Community A is contained within community B and D, while community B and C have overlapping nodes such as 4,5, and 6. Thus the proposed approach is capable of accurately identifying disjoint, overlapping, and nested communities.

B. COMPLEXITY OF THE PROPOSED PROBABILISTIC CLUSRANK

For a graph $G = (V, E)$ where V is the set of vertices and E is the set of edges, complexity is computed of the proposed probabilistic ClusRank algorithm.

1) **Community Detection -**
   a) Computing FOAN: $O(|E|)$
   b) Computing SOAN: $O(|E|^2)$
   c) Finding Jaccard Similarity: $O(|E|^2 \cdot log(\overline{E}))$ where $\overline{E}$ represent average number of edges in FOAN.
   d) Computing CSON: $O(|E|^2)$
   e) Merging LNS: $O(|E| \cdot log(m))$ where m is the average number of LNS merged.

   This results in complexity of $O(|E|)+N_{iter} \times (O(|E|^2)+ O(|E|^2 \cdot log(\overline{E})) + O(|E| \cdot log(m))$ where N_{iter} is the number of iterations performed to obtain a stable CSOAN, equivalent to $N_{iter}O(|E|^2 \times log(\overline{E}))$. The maximum number of iterations the algorithm will be at most $|E|$. This gives a upper bound on the complexity of this step as $O(|E|^3)$.
2) **Influential Nodes** - For g number of sentences in the golden summary, this step takes $g \times O(|V|^2)$. This is a poor upper bound as the number of edges formed during graph construction are far fewer than $\binom{|V|}{2}$ and this step takes very few seconds on real test cases.
3) Thus, the complexity of the proposed probabilistic ClusRank is $O(|E|^3)+O(|V|^2)$.

C. FINDING INFLUENTIAL NODES

After following above steps, a list of nodes in different communities is obtained. The overlapping communities are sorted in their decreasing length and the largest community is picked. In the given example $H = \{1, 2, 3, 4, 8, 9, 10\}$ is largest. The weighted degree of the nodes are found out as given in table 10. Node 1 has highest degree, it is removed from the community and added to a vector S. S now contains $S = 1$. Again the largest community is picked and found out to be $H = \{4, 5, 6, 7, 8, 9, 10\}$, node 10 has highest weighted degree of 7, it is added to vector S. S now contains 1, 7. Similarly, the operation is carried out till 5 nodes are found out. The resultant vector $S = \{1, 7, 3, 10, 2\}$. These are the most influential nodes in the network.

D. FINDING SALIENT SENTENCES

The Algorithm 1 is explained here that deals with selection of salient sentences within budget b and cost c(budget b is summary length and cost is count ow words in a sentence). After the salient nodes are selected, if we assume them to be important sentences according to their weighted degree, there are chances that sentences with more terms will be selected in higher priority as they contain more terms. To overcome such situation, the weighted degree of node is divided by the length of sentence. This is termed as quality increase. These values are further stored in ω list. After creating an array called Rl[], the nodes are arranged as follows: according to the descending order of their improvement in quality increase. There is one option of selecting nodes with highest value in the Rl[] list.However, it could result in the selection of nodes that are overly restricted to a certain area of the text. They do not thereby generate a concept that can be broadly applied. Thus, we introduce a different approach to solve this problem. Initially, the top node is chosen from Rl[] list and is added to Salientsentence[] list. The second node/sentence is only added in Salientsentence[] list if it is not a neighbor of node currently in Salientsentence list. Second condition the cost incurred in adding second node/sentence does not surpass budget b. Addition of a scaling factor is introduced

TABLE 3. Index mapping of the edges in the example graph(given in Fig.4).

Edge	Map Index	Edge	Map Index	Edge	Map Index	Edge	Map Index
(1, 2)	1	(1, 3)	2	(1, 4)	3	(2, 3)	4
(2, 4)	5	(3, 4)	6	(4, 5)	7	(4, 6)	8
(4, 7)	9	(5, 6)	10	(5, 7)	11	(6, 7)	12
(7, 8)	13	(7, 9)	14	(7, 10)	15	(8, 9)	16
(8, 10)	17	(9, 10)	18	-	-	-	-

TABLE 4. FOAN of all edges in graph. The numbers shown in right column are the edge map indices for brevity, refer to table 3.

Edge	Elements in FOAN (Edge Map Indices)
(1, 2)	1, 2, 3, 4, 5
(1, 3)	1, 2, 3, 4, 6
(1, 4)	1, 2, 3, 5, 6, 7, 8, 9
(2, 3)	1, 2, 4, 5, 6
(2, 4)	1, 3, 4, 5, 6, 7, 8, 9
(3, 4)	2, 3, 4, 5, 6, 7, 8, 9
(4, 5)	3, 5, 6, 7, 8, 9, 10, 11
(4, 6)	3, 5, 6, 7, 8, 9, 10, 12
(4, 7)	3, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15
(5, 6)	7, 8, 10, 11, 12
(5, 7)	7, 9, 10, 11, 12, 13, 14, 15
(6, 7)	8, 9, 10, 11, 12, 13, 14, 15
(7, 8)	9, 11, 12, 13, 14, 15, 16, 17
(7, 9)	9, 11, 12, 13, 14, 15, 16, 18
(7, 10)	9, 11, 12, 13, 14, 15, 17, 18
(8, 9)	13, 14, 16, 17, 18
(8, 10)	13, 15, 16, 17, 18
(9, 10)	14, 15, 16, 17, 18

TABLE 5. SOAN of all edges in graph. The numbers shown in right column are the edge map indices for brevity, refer to table 3.

Edge	Elements in SOAN (Edge Map Indices)
(1, 2)	1, 2, 3, 4, 5
(1, 3)	1, 2, 3, 4, 6
(1, 4)	1, 2, 3, 5, 6, 7, 8, 9
(2, 3)	1, 2, 4, 5, 6
(2, 4)	1, 3, 4, 5, 6, 7, 8, 9
(3, 4)	2, 3, 4, 5, 6, 7, 8, 9
(4, 5)	3, 5, 6, 7, 8, 9, 10, 11
(4, 6)	3, 5, 6, 7, 8, 9, 10, 12
(4, 7)	3, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15
(5, 6)	7, 8, 10, 11, 12
(5, 7)	7, 9, 10, 11, 12, 13, 14, 15
(6, 7)	8, 9, 10, 11, 12, 13, 14, 15
(7, 8)	9, 11, 12, 13, 14, 15, 16, 17
(7, 9)	9, 11, 12, 13, 14, 15, 16, 18
(7, 10)	9, 11, 12, 13, 14, 15, 17, 18
(8, 9)	13, 14, 16, 17, 18
(8, 10)	13, 15, 16, 17, 18
(9, 10)	14, 15, 16, 17, 18

to adjust values of cost. Since weighted degree of a node and cost are not comparable.

E. REAL EXAMPLES

Given a document with text mentioned below, the probabilistic ClusRank algorithm is used to perform the extractive summarization. It detects $|\mathcal{H}| = 23$ communities and the sentences mentioned in quotes below are extracted. ROUGE-n gram score is computed to analyse the efficacy of the extracted summary(a brief discussion about ROUGE is given in Section VI-C). The results obtained in terms of recall, precision, and f-score presented in table 11 demonstrate the efficiency of the proposed approach.

TABLE 6. Jaccard Similarity.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	∞																	
2	0.79	∞																
3	0.51	0.58	∞															
4	0.71	0.79	0.51	∞														
5	0.53	0.60	0.70	0.53	∞													
6	0.47	0.53	0.63	0.47	0.65	∞												
7	0.13	0.19	0.46	0.13	0.45	0.48	∞											
8	0.14	0.20	0.49	0.14	0.48	0.51	0.64	∞										
9	0.09	0.13	0.30	0.09	0.30	0.31	0.51	0.46	∞									
10	0.00	0.00	0.08	0.00	0.08	0.08	0.39	0.31	0.30	∞								
11	0.00	0.00	0.10	0.00	0.10	0.10	0.30	0.25	0.72	0.38	∞							
12	0.00	0.00	0.10	0.00	0.10	0.10	0.30	0.25	0.72	0.38	0.91	∞						
13	0.00	0.00	0.07	0.00	0.06	0.07	0.21	0.16	0.64	0.25	0.80	0.80	∞					
14	0.00	0.00	0.06	0.00	0.06	0.07	0.20	0.15	0.62	0.25	0.76	0.76	0.80	∞				
15	0.00	0.00	0.06	0.00	0.06	0.06	0.19	0.15	0.60	0.24	0.74	0.74	0.84	0.88	∞			
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.00	0.22	0.22	0.35	0.39	0.43	∞		
17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.00	0.33	0.33	0.47	0.51	0.55	0.64	∞	
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.00	0.22	0.22	0.35	0.39	0.43	0.45	0.64	∞

TABLE 7. CSOAN of all edges in graph. The numbers shown in right column are the edge map indices for brevity, refer to table 3.

Edge	Elements in CSOAN (Edge Map Indices)
(1, 2)	1, 2, 3, 4, 5
(1, 3)	1, 2, 3, 4, 5, 6
(1, 4)	1, 2, 3, 4, 5, 6
(2, 3)	1, 2, 3, 4, 5
(2, 4)	1, 2, 3, 4, 5, 6
(3, 4)	2, 3, 5, 6, 8
(4, 5)	7, 8, 9
(4, 6)	6, 7, 8
(4, 7)	7, 9, 11, 12, 13, 14, 15
(5, 6)	10
(5, 7)	9, 11, 12, 13, 14, 15
(6, 7)	9, 11, 12, 13, 14, 15
(7, 8)	9, 11, 12, 13, 14, 15
(7, 9)	9, 11, 12, 13, 14, 15, 17
(7, 10)	9, 11, 12, 13, 14, 15, 17
(8, 9)	16, 17
(8, 10)	14, 15, 16, 17, 18
(9, 10)	17, 18

Document:

Source: https://edition.cnn.com/2014/06/12/health/virus-chikungunya/index.htmlCNN News Article by Val Willingham

A debilitating, mosquito-borne virus called chikungunya has made its way to North Carolina, health officials say. It's the state's first reported case of the virus. The patient was likely infected in the Caribbean, according to the Forsyth County Department of Public Health. Chikungunya is primarily found in Africa, East Asia and the Caribbean islands, but the Centers for Disease

TABLE 8. CSOAN of all edges in graph after the loop stops. The numbers shown in right column are the edge map indices for brevity, refer to table 3.

Edge	Elements in CSOAN (Edge Map Indices)	Link Node Set	Conductance ϕ
(1, 2)	1, 2, 3, 4, 5, 6	3, 4, 1, 2	0.2
(1, 3)	1, 2, 3, 4, 5, 6	3, 4, 1, 2	0.2
(1, 4)	1, 2, 3, 4, 5, 6	3, 4, 1, 2	0.2
(2, 3)	1, 2, 3, 4, 5, 6	3, 4, 1, 2	0.2
(2, 4)	1, 2, 3, 4, 5, 6	3, 4, 1, 2	0.2
(3, 4)	1, 2, 3, 4, 5, 6	3, 4, 1, 2	0.2
(4, 5)	7	4, 5	0.778
(4, 6)	8	4, 6	0.778
(4, 7)	9, 11, 12, 13, 14, 15	8, 6, 5, 10, 7, 4, 9	0.334
(5, 6)	10	6, 5	0.667
(5, 7)	9, 11, 12, 13, 14, 15	8, 6, 5, 10, 7, 4, 9	0.334
(6, 7)	9, 11, 12, 13, 14, 15	8, 6, 5, 10, 7, 4, 9	0.334
(7, 8)	9, 11, 12, 13, 14, 15	8, 6, 5, 10, 7, 4, 9	0.334
(7, 9)	9, 11, 12, 13, 14, 15	8, 6, 5, 10, 7, 4, 9	0.334
(7, 10)	9, 11, 12, 13, 14, 15	8, 6, 5, 10, 7, 4, 9	0.334
(8, 9)	16	8, 9	0.667
(8, 10)	17, 18	8, 10, 9	0.334
(9, 10)	17, 18	8, 10, 9	0.334

TABLE 9. Communities Detected.

A	B	C	D	E
1, 2, 3, 4	1, 2, 3, 4, 5, 6	4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 8, 9, 10	8, 9, 10

TABLE 10. Weighted degree of the nodes in community $H = \{1, 2, 3, 4, 8, 9, 10\}$.

Node	Weighted Degree	Node	Weighted Degree	Node	Weighted Degree
1	20	2	18	3	18
4	12	8	6	9	8
		10	10		

Control and Prevention has been watching the virus,+ for fear that it could take hold in the United States – much like West Nile did more than a decade ago. The virus, which can cause joint pain and arthritis-like symptoms, has been on the U.S. public health radar for some time. About 25 to 28 infected travelers bring it to the United States each year, said Roger Nasci, chief of the CDC's Arboviral Disease Branch in the Division of Vector-Borne Diseases. ''We haven't had any locally transmitted cases in the U.S. thus far,'' Nasci said. But a major outbreak in the Caribbean this year – with more than 100,000 cases reported – has health officials concerned. Experts say American tourists are bringing chikungunya back home, and it's just a matter of time before it starts to spread within the United States. Study: Beer drinkers attract mosquitoes
Study: Beer drinkers attract mosquitoes 01:26 After all, the Caribbean is a popular one with American tourists, and summer is fast approaching. ''So far this year we've recorded eight travel-associated cases, and seven of them have come from countries in the Caribbean where we know the virus is being transmitted,'' Nasci said. Other states have also reported cases of chikungunya. The Tennessee Department of Health said the state has had multiple cases of the virus in people who have traveled to the Caribbean. The virus is not deadly, but it can be painful, with symptoms lasting for weeks. Those with weak immune systems, such as the elderly, are more likely to suffer from the virus' side effects than those who are healthier. The good news, said Dr. William Shaffner, an infectious disease expert with Vanderbilt University in Nashville, is that the United States is more sophisticated when it comes to controlling mosquitoes than many other nations. ''We live in a largely air-conditioned environment, and we have a lot of screening (window screens, porch screens),'' Shaffner said. ''So we can separate the humans from the mosquito population, but we cannot be completely be isolated.'' Chikungunya was originally identified in East Africa in the 1950s. The ecological makeup of the United States supports the spread of an illness such as this, especially in the tropical areas of Florida and other Southern states, according to the CDC. The other concern is the type of mosquito that carries the illness. Unlike most mosquitoes that breed and prosper outside from dusk to dawn, the chikungunya virus is most often spread to people by Aedes aegypti and Aedes albopictus mosquitoes. These are the same mosquitoes that transmit the virus that causes dengue fever. They bite mostly during the daytime. The disease is transmitted from mosquito to human, human to mosquito and so forth. A female mosquito of this type lives three to four weeks and can bite someone every three to four days.

Extracted: Chikungunya is primarily found in Africa, East Asia and the Caribbean islands. The virus, which can cause joint pain and arthritis-like

Algorithm 1 Salient Sentences Selection

procedure SALIENTSENTENCES(G, b)
  Assume the graph G contains m nodes.
  Salientsentence $\leftarrow \phi$, c $\leftarrow$ 0, count $\leftarrow$ 0, A $\leftarrow$ 0, d $\leftarrow$ 0, $\omega \leftarrow$ 0, Rl $\leftarrow$ 0
  for i = 1 to m **do**
    $\omega[c]$ = rl[c] / $(len(rl[c])^r)$
  Rl $\leftarrow$ Nodes are arranged in descending order according to ω value.
  while d < b and c $\neq$ m **do**
    if G[c] is not an adjacent neighbor in Rl **then**
      Salientsentence $\leftarrow$ Salientsentence + G[c]
      count $\leftarrow$ count + 1
      d $\leftarrow$ d + len(c)
      A $\leftarrow$ A $\cup$ Salientsentence
    c $\leftarrow$ c + 1
return A

TABLE 11. ROUGE-1, 2, 3 score on the given real example in Subsection V-E.

	Recall	Precision	F-score
ROUGE-1	0.69	0. 46	0.51
ROUGE-2	0.517241	0.40540541	0.45454545
ROUGE-3	0.246753	0.19322034	0.21673004

TABLE 12. Statistics of data sets.

	DUC 2001	DUC 2002	DUC2004	DUC2006
Count of clusters	30	59	50	50
Length of summary	100 words	100 words	665 bytes	250 words
Source of data	TREC-9	TREC-9	TREC-9	TREC-9
Count of documents	309	567	500	620

symptoms, has been on the U.S. public health radar for some time.

Golden Summary: North Carolina reports first case of mosquito-borne virus called chikungunya. Chikungunya is primarily found in Africa, East Asia and the Caribbean islands. Virus is not deadly, but it can be painful, with symptoms lasting for weeks.

VI. ANALYSIS OF EXPERIMENTS

To validate the efficiency of the proposed probabilistic approach, experiments are conducted on the data sets given by Document Understanding Conference (DUC). DUC is a method assessment competition that allows researchers to assess the efficiency of various summarization methods on similar data sets.

A. DATA SETS

A brief explanation about the data sets are given below:

- DUC 2001 - This dataset requires task summarizers to extract generic summaries from a total of 309 documents that consist of newswire or paper stories. These documents were further grouped into 30 clusters based on their relevance of topic. Each cluster was composed of 10, 50,100, and 200 words of fixed target length. Source of the documents was Text Retrieval Conference. The summarizers were asked to generate summary of 100 words summary length.
- DUC 2002 – It consists of documents related to newspaper articles. There are 59 clusters of documents in this collection of newspaper articles. There are one or two handwritten, roughly 100-word abstracts for each document in the collections. Source of the documents are from Text Retrieval Conference.
- DUC2004- This dataset is intended solely for testing purposes. There are 500 news related stories in total, and four human-written summaries accompanying each other. Particularly, it comprises fifty groups of Text Retrieval Conference (TREC) records, sourced from the subsequent compilations: Xinhua News Agency, .AP newswire, New York Times newswire, 1998-2000. Total number of clusters were 50 with 10 documents in each cluster.
- DUC2006 - DUC 2006 is a difficult question-focused summary assignment. In order to respond to a question or series of questions presented in a DUC subject, summarizers had to gather together information from several documents. Fifty DUC subjects in total were created by NIST Assessors to serve as test data. In order to create a topic statement—a request for information that could be addressed with the help of the chosen papers—the assessor first chose 25 relevant documents from the newswire for each topic.

A brief description about the statistics of the data set is presented in Table 12. DUC 2001, DUC 2002, DUC 2004, and DUC 2006 data sets are used for experimental analysis and a comparative study is made for the system generated summaries and golden summaries.

B. PRE-PROCESSING

Some linguistic techniques such as stemming, upper case removal, removal of stop words, and segmentation of sentences has been used during pre-processing phase of the document in this experiment. The textual content of the document is divided into sentences during the process of segmentation. Words appearing frequently such as a, an, the etc. within the text are removed during stop word removal process, since they are considered irrelevant. The process of stemming involves reduction of a word to its root stem. Porter Stemmer is used for the stemming of word. The process of pre-processing is performed before the execution of the algorithm.

C. EVALUATION METRIC

DUC has adopted ROUGE metric [47] for evaluation of automatically generated summary, hence, the proposed research uses this metric for performance analysis of the proposed research. The quality of summary is measured using ROUGE metric, that basically counts the number of overlapping units such as word pairs, word sequences, and n-grams between the reference summary and candidate summary. ROUGE-1 and ROUGE-2 recall score is used in this research for the evaluation of automatic summaries.

D. EVALUATION OF PERFORMANCE

This section deals with a comparative analysis of the performance of the proposed probabilistic approach with some recent works. The proposed probabilistic ClusRank method is compared with some baseline methods: (a) ESDS [24] – a search algorithm based on binary optimization, (b) manifold ranking [48] - greedy search involving probabilistic approach, (c) NetSum [49] - approach based on neural networks,

TABLE 13. R-1 and R-2 recall score for DUC01 and DUC02 dataset.

Methods	DUC01				DUC02			
	R-1	Rank	R-2	Rank	R-1	Rank	R-2	Rank
FEOM	0.4773	1	0.1855	5	0.4658	7	0.1249	8
ClusRank	0.4725	2	0.2011	2	0.4906	1	0.2306	1
NetSum	0.4643	3	0.1770	7	0.4496	8	0.1117	9
CRF	0.4551	4	0.1773	9	0.4401	10	0.1092	10
ESDS	0.4540	5	0.1957	4	0.4790	5	0.2214	4
UnifiedRank	0.4538	6	0.1765	8	0.4849	2	0.2146	5
MA	0.4486	7	0.2014	1	0.4828	3	0.2284	3
QCS	0.4485	8	0.1852	6	0.4487	9	0.1877	7
LexRank	0.4468	9	0.1989	3	0.4796	4	0.2295	2
SVM	0.4463	10	0.1702	10	0.4324	11	0.1087	11
CollabSum	0.4404	11	0.1623	12	0.4719	6	0.2010	6
ManifoldRanking	0.4336	12	0.1664	11	0.4233	12	0.1068	12
DPSO	0.3993	13	0.0832	13	0.4172	13	0.1026	13
0-1 non-linear	0.0.3876	14	0.0778	14	0.4097	14	0.0937	14

TABLE 14. R-1 and R-2 recall score for DUC04 and DUC06 data set.

Methods	DUC04				DUC06			
	R-1	Rank	R-2	Rank	R-1	Rank	R-2	Rank
FEOM	0.4872	2	0.1870	4	0.4560	5	0.2981	3
ClusRank	0.4816	3	0.1770	5	0.4770	1	0.3000	2
NetSum	0.4608	6	0.1559	8	0.4321	8	0.2916	4
CRF	0.4519	7	0.1489	9	0.4403	7	0.2552	10
ESDS	0.4493	8	0.2112	1	0.4423	6	0.3116	1
UnifiedRank	0.4891	1	0.1669	6	0.4623	2	0.2871	6
MA	0.4776	4	0.1660	7	0.4604	3	0.2881	5
QCS	0.4226	9	0.2021	2	0.4111	10	0.2772	8
LexRank	0.4714	5	0.1989	3	0.4291	9	0.2661	9
SVM	0.4008	10	0.1361	10	0.4591	4	0.2790	7
CollabSum	0.3872	12	0.1244	12	0.4008	11	0.2443	11
ManifoldRanking	0.3791	15	0.1221	13	0.3920	13	0.2311	12
DPSO	0.3981	11	0.1191	14	0.3983	12	0.2001	14
0-1 non-linear	0.3868	13	0.1254	11	0.3801	14	0.2291	13
GRU+GCN	0.3823	14	0.0948	15				

TABLE 15. Relative improvement of ClusRank with other methods(DUC01 and DUC02).

Methods	DUC01		DUC02	
	R-1	R-2	R-1	R-2
FEOM	(-)0.96	(+) 8.46	(+) 5.37	(+) 84.87
NetSum	(+)1.81	(+) 13.67	(+) 9.16	(+) 106.71
CRF	(+)3.87	(+) 16.10	(+) 11.52	(+) 111.45
ESDS	(+)4.12	(+) 2.81	(+) 2.46	(+) 4.29
UnifiedRank	(+)4.16	(+) 13.99	(+) 1.22	(+) 7.60
MA	(+)5.37	(-) 0.10	(+) 1.66	(+) 1.09
QCS	(+)5.40	(+) 8.64	(+) 9.38	(+) 23.02
LexRank	(+)5.80	(+) 1.16	(+) 2.34	(+) 0.61
SVM	(+)5.92	(+) 18.21	(+) 13.51	(+) 112.42
CollabSum	(+)7.33	(+) 23.97	(+) 4.01	(+) 14.88
ManifoldRanking	(+)9.02	(+) 20.91	(+) 15.95	(+) 116.20
DPSO	(+)18.38	(+) 141.83	(+) 17.64	(+) 125.05
0-1 non-linear	(+)21.96	(+) 158.61	(+) 19.79	(+) 146.42

TABLE 16. Relative improvement of ClusRank with other methods (DUC04 and DUC06 data set).

Methods	DUC04		DUC06	
	R-1	R-2	R-1	R-2
FEOM	(-)0.22	(-)5.34	(+) 1.14	(+) 0.63
NetSum	(+)4.51	(+) 13.53	(+) 10.39	(+) 2.88
CRF	(+)6.57	(+) 18.87	(+) 8.33	(+) 17.55
ESDS	(+)7.19	(-) 16.19	(+) 7.84	(-) 3.72
UnifiedRank	(-)1.53	(+) 6.05	(+) 3.17	(+) 4.49
MA	(+)0.83	(+) 6.02	(+)3.60	(+) 4.13
QCS	(+)13.96	(-) 12.42	(+) 16.03	(+) 8.22
LexRank	(+)2.16	(-) 11.01	(+) 11.16	(+) 12.73
SVM	(+)20.16	(+) 30.05	(+) 3.89	(+) 7.52
CollabSum	(+)24.38	(+) 42.28	(+) 19.01	(+) 22.79
ManifoldRanking	(+)27.04	(+) 44.96	(+) 21.68	(+) 29.81
DPSO	(+)20.97	(+) 48.61	(+) 19.75	(+) 49.93
0-1 non-linear	(+)24.51	(+) 41.15	(+) 25.49	(+) 30.94

(d) MA [9]- local search and genetic operators based meta-heuristic approach, (e) CRF [41] - approach based on conditional random field, (f) FEOM [50] – evolutionary algorithm involving fuzzy approach, (g) SVM [51] - mathematical approach, (h) QC [52] - hidden markov model based approach, (i) 0–1 non linear [53] - evolutionary algorithm approach based on binary PSO, (j) UnifiedRank [28] - ranking approach based on graph, (k) CollabSum [54] - clustering approach along with ranking of graphs, (l) LexRank [20] - method involving ranking of graphs, (m) DPSO [25] - optimization approach based on evolutionary algorithm. The above methods have accomplished good results on DUC01 and DUC02 data sets, due to this reason they are selected for comparison with the proposed probabilistic ClusRank method.

TABLE 17. New resultant ranklist.

Methods	R_k														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Rank
ClusRank	3	2	2	0	0	0	0	0	0	0	0	0	0	0	3.9
FEOM	0	0	0	0	1	1	1	0	0	0	0	0	0	0	2.6
ESDS	0	0	1	0	1	1	2	0	0	0	0	0	0	0	2.9
MA	2	0	1	0	0	0	0	0	0	1	0	0	0	0	2.8
LexRank	0	0	1	1	0	1	0	0	0	0	0	0	0	0	2.8
CRF	0	0	0	1	0	0	0	0	1	0	0	2	0	0	1.7
SVM	0	0	0	0	0	0	2	0	2	0	0	0	0	0	1.1
CollabSum	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1.7
0-1 non-linear	3	2	2	0	0	0	0	0	0	0	0	0	0	0	1.19
DPSO	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0.4
UnifiedRank	1	0	0	1	0	0	1	0	0	1	0	0	0	0	2.6
ManifoldRanking	0	0	0	0	0	0	0	0	1	0	0	2	0	0	0.7
Netsum	0	0	1	0	0	0	0	1	1	0	0	0	0	0	2.1
QCS	0	0	2	0	0	1	0	1	0	0	0	0	0	0	1.9

Table 13 presents R-1 and R-2 recall score for DUC01 and DUC02 data sets. As is concerned for DUC01 data set, FEOM and CluRank acheive first and second place for R-1 recall score, whereas MA and ClusRank accomplish first and second place for R-2 recall score. For DUC02 data set, ClusRank is performing best for both R-1 and R-2 recall score, with DPSO and 0-1 non-linear performing worst in every case. The R-1 and R-2 score for DUC04 and DUC06 is presented in table 14. ClusRank secures third and first position for DUC04 and DUC06 data sets.

A significant improvement was observed in the performance of the proposed approach for R-1 and R-2 recall score in comparison with other methods. The relative improvement of ClusRank w.r.t other approaches is presented in Table 15 and Table 16. As is evident, ClusRank outperforms state-of-the-art methods and achieves highest R-1 and R-2 recall score for DUC01, DUC02, and DUC06 data sets. Relative improvement is used as a measure for comparison. The formula for calculating relative improvement is $\frac{(c-b)}{b} \times 100$, when a comparison of c is made with b. For showing the relative improvement of ClusRank with other methods "+" sign is used, whereas "-" means opposite. As shown in Table 13 and table 14, ClusRank has outperformed other methods for R-1 and R-2 recall score on DCU01, DUC02 and DUC04 data sets. For DUC01 data set, only FEOM and MA have performed better than ClusRank. FEOM has shown an improvement of 0.96% for R-1 metric whereas, MA is performing 0.10% better than ClusRank for R-2 metric. ClusRank has outperformed state-of-the-art methods for DUC02 and DUC06 data sets.However, in case of DUC04, the performance is slightly good. The reason for ClusRank's good performance is, it is clustering sentences based on large sub-topics that are carrying more weight, and obtained salient sentences having maximum diversity and coverage.

As visible from Table 13, and Table 14, the rank of ClusRank on all the four data sets for R-1 and R-2 are different, thus, we cannot validate the efficiency of the ClusRank. Hence, to get a clear status of the ranking methods, this work uses a combined ranking approach proposed by [54], that considers the rank of every individual method for every measure. The formula for the ranking approach is:

$$Rank = \sum_{k=1}^{C_m} \frac{C_m - k + 1}{C_m} \times R_k \tag{8}$$

The new resultant rank list obtained from eq.8 is presented in Table 17. Based on Table 17 observation, we can draw the following conclusion:

- The method ClusRank achieves first rank in the new resultant rank list shown in table 17 computed according to eq.8, in comparison with other state-of-the art methods such as MA and FEOM, which had comparatively better ranking on DCU01 and DUC04 data sets for R-1 and R-2 measure (seen in table 13 and table 14).
- The proposed approach ClusRank has outperformed state-of-the-art-methods for DUC02 data set and performed competitively well for DUC01 data set except for DE and MA method.
- Although LexRank and UnifiedRank are graph based approaches, but its performance is less in comparison to ClusRank, which is a combination of clustering and graph ranking method.
- Optimization methods such as ESDS and MA based on a combination of three features namely, position of the sentence, length of the sentence, and coverage of sentence secure second and third position in the rank list presented in table 17.
- ClusRank, FEOM, UnifiedRank, ESDS, LexRank, and MA are unsupervised methods that have performed better than SVM, a supervised approach.
- The performance of ClusRank, which uses clustering approach for summary generation is higher than optimization methods such as Unified rank, and FEOM.
- 0-1 non-linear and DPSO have failed to perform since they don't use the clustering concept.

E. ANALYSIS OF STRUCTURAL FINDINGS

As observed from table 13 - 17, that the proposed approach has performed better than other competitive methods. The increased performance is due to the fact: (a) The proposed

TABLE 18. Comparative analysis of t-test values.

Methods	t-test (p value)
FEOM	<0.0001
ESDS	<0.0002
LexRank	<0.0003
CRF	<0.0002
SVM	<0.00001
CollabSum	<0.00013
0-1 non-linear	<0.0001
DPSO	<0.0002
UnifiedRank	<0.0002
ManifoldRanking	<0.0001
Netsum	<0.00011
QCS	<0.00021

approach is a combination of clustering among sentences and then extraction of salient sentences from them. In real-world, communities exist within networks, that are overlapping, disjoint, and nested. The proposed community detection algorithm is based on the concept of link mutual reciprocity that is able to detect overlapping, disjoint, and nested communities. (b) Further, a sentence selection algorithm is introduced that is extracting sentences within budget b and cost c.

Upon applying t-Test with null hypothesis: ''The proposed system is equal or inferior to other competitive methods in ROUGE-2'', the one-tailed p-value was less than 0.0002 implying the difference to be extremely statistically significant. Thus we reject the null hypothesis and conclude that the system is better than other competitive methods. The results of t-test are shown in table 18.

VII. CONCLUSION

This research work proposes a probabilistic method named as ClusRank: a clustering combined graph ranking approach for generating extractive summaries. It aims to cover two aspects: (a) diversity – obtained summary should not cover redundant information; (b) coverage – resultant summary should contain different main topics, sub-topics of the original source document. We presented an approach for disjoint, nested and overlapping communities' identification in complex networks, which is based on link approximation. The suggested technique combines link communities' advantages with rough set concept. The suggested approach successfully identifies coexisting fragmented, nested, and overlapping community structures in intricate real-world networks, as demonstrated by the experiments. Although the suggested technique offers a significant methodological advancement to the difficult community discovery problem, it is limited to unweighted and undirected networks. Above that, a senetence slection algorithm is proposed that extracts summary within budget and cost. Summarizing, initially, a clustering algorithm is proposed that groups sentences into clusters based on topics and sub-topics. Then, from every cluster, most salient and representative sentences are selected using proposed probabilistic algorithm.

The performance of the probabilistic ClusRank algorithm is validated on DUC01,DUC02, DUC04 and DUC06 data sets in terms of Recall-1 and Recall-2 measure. The proposed approach obtains best results for DUC02 and DUC06 data sets beating the best performing MA approach by 1.66% and 2.91%. It obtains good results for DUC01 and DUC04 data sets, however, slightly lagging behind FEOM approach.When combined ranking approach is used, the proposed ClusRank secures first position leading ahead against two popular methods namely,FEOM, and MA. Other graph based approaches such as Lex Rank and Unified Ranking are not so effective since they lack the concept of clustering. The reason for the promising results of the proposed probabilistic approach is, it is able to extract main topics and sub-topics from the main text with maximum coverage and diversity.

The future work remains to include more techniques based on optimization, use different combinations of similarity measures for extracting summaries, incorporate different similarity measures in the clustering procedure, and propose algorithms for dynamic overlapping community detection.

REFERENCES

[1] P. Viswanath, N. M. Murty, and B. Shalabh, ''Pattern synthesis for nonparametric pattern recognition,'' in *Encyclopedia of Data Warehousing and Mining*, 2nd ed. Hershey, PA, USA: IGI Global, 2009, pp. 1511–1516.

[2] I. Mani and M. Maybury, *Automatic Summarization*. The Netherlands: John Benjamin's, 2001, pp. 1–22.

[3] B. Y. Ricardo and R. N. Berthier, *Modern Information Retrieval*. Bengaluru, India: Pearson Education India, 1999, pp. 23–38.

[4] O. Zamir and O. Etzioni, ''Web document clustering: A feasibility demonstration,'' in *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 1998, pp. 46–54.

[5] A. Turpin, Y. Tsegay, D. Hawking, and H. E. Williams, ''Fast generation of result snippets in web search,'' in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2007, pp. 127–134.

[6] I. S. Dhillon, S. Mallela, and D. S. Modha, ''Information-theoretic co-clustering,'' in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2003, pp. 89–98.

[7] H. Jing and K. McKeown, ''Cut and paste based text summarization,'' in *Proc. 1st Meeting North Amer. Chapter Assoc. Comput. Linguistics*, 2000, pp. 178–185.

[8] K. Knight and D. Marcu, ''Summarization beyond sentence extraction: A probabilistic approach to sentence compression,'' *Artif. Intell.*, vol. 139, no. 1, pp. 91–107, Jul. 2002.

[9] M. Mendoza, S. Bonilla, C. Noguera, C. Cobos, and E. León, ''Extractive single-document summarization based on genetic operators and guided local search,'' *Expert Syst. Appl.*, vol. 41, no. 9, pp. 4158–4169, Jul. 2014.

[10] M. G. Ozsoy, F. N. Alpaslan, and I. Cicekli, ''Text summarization using latent semantic analysis,'' *J. Inf. Sci.*, vol. 37, no. 4, pp. 405–417, Aug. 2011.

[11] M. Jang and P. Kang, ''Learning-free unsupervised extractive summarization model,'' *IEEE Access*, vol. 9, pp. 14358–14368, 2021.

[12] X. Cai and W. Li, ''A spectral analysis approach to document summarization: Clustering and ranking sentences simultaneously,'' *Inf. Sci.*, vol. 181, no. 18, pp. 3816–3827, Sep. 2011.

[13] X. Wan and J. Yang, ''CollabSum: Exploiting multiple document clustering for collaborative single document summarizations,'' in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2007, pp. 143–150.

[14] X. Cai, W. Li, and R. Zhang, ''Combining co-clustering with noise detection for theme-based summarization,'' *ACM Trans. Speech Lang. Process.*, vol. 10, no. 4, pp. 1–27, Dec. 2013.

[15] X. Cai, W. Li, and R. Zhang, ''Enhancing diversity and coverage of document summaries through subspace clustering and clustering-based optimization,'' *Inf. Sci.*, vol. 279, pp. 764–775, Sep. 2014.

[16] R. Guo, J. Zou, Q. Bai, W. Wang, and X. Chang, ''Community detection fusing graph attention network,'' *Mathematics*, vol. 10, no. 21, p. 4155, Nov. 2022.

[17] J.-P. Mei and L. Chen, "SumCR: A new subtopic-based extractive approach for text summarization," *Knowl. Inf. Syst.*, vol. 31, no. 3, pp. 527–545, Jun. 2012.
[18] E. Lloret and M. Palomar, "COMPENDIUM: A text summarisation tool for generating summaries of multiple purposes, domains, and genres," *Natural Lang. Eng.*, vol. 19, no. 2, pp. 147–186, Apr. 2013.
[19] W. Luo, F. Zhuang, Q. He, and Z. Shi, "Exploiting relevance, coverage, and novelty for query-focused multi-document summarization," *Knowl.-Based Syst.*, vol. 46, pp. 33–42, Jul. 2013.
[20] C. C. Yang and F. L. Wang, "Hierarchical summarization of large documents," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 59, no. 6, pp. 887–902, Apr. 2008.
[21] R. Ferreira, L. de Souza Cabral, F. Freitas, R. D. Lins, G. de França Silva, S. J. Simske, and L. Favaro, "A multi-document summarization system based on statistics and linguistic treatment," *Expert Syst. Appl.*, vol. 41, no. 13, pp. 5780–5787, Oct. 2014.
[22] Y. Li and S. Li, "Query-focused multi-document summarization: Combining a topic model with graph-based semi-supervised learning," in *Proc. 25th Int. Conf. Comput. Linguistics (COLING)*, 2014, pp. 1197–1207.
[23] J. Balaji, T. V. Geetha, and R. Parthasarathi, "A graph based query focused multi-document summarization," *Int. J. Intell. Inf. Technol.*, vol. 10, no. 1, pp. 16–41, Jan. 2014.
[24] X. Wan, J. Yang, and J. Xiao, "Manifold-ranking based topic-focused multi-document summarization," in *Proc. 20th Int. Joint Conf. Artif. Intell.*, 2007, pp. 2903–2908.
[25] R. M. Alguliev, R. M. Aliguliyev, and N. R. Isazade, "DESAMC+DocSum: Differential evolution with self-adaptive mutation and crossover parameters for multi-document summarization," *Knowl.-Based Syst.*, vol. 36, pp. 21–38, Dec. 2012.
[26] D. Hingu, D. Shah, and S. S. Udmale, "Automatic text summarization of Wikipedia articles," in *Proc. Int. Conf. Commun., Inf. Comput. Technol. (ICCICT)*, Jan. 2015, pp. 1–4.
[27] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *J. Artif. Intell. Res.*, vol. 22, pp. 457–479, Dec. 2004.
[28] X. Wan, "Towards a unified approach to simultaneous single-document and multi-document summarizations," in *Proc. 23rd Int. Conf. Comput. Linguistics*, 2010, pp. 1137–1145.
[29] P. Yang, X. Liu, and G. Xu, "A dynamic weighted TOPSIS method for identifying influential nodes in complex networks," *Modern Phys. Lett. B*, vol. 32, no. 19, Jul. 2018, Art. no. 1850216.
[30] Z. Ding, X. Zhang, D. Sun, and B. Luo, "Overlapping community detection based on network decomposition," *Sci. Rep.*, vol. 6, no. 1, pp. 1–11, Apr. 2016.
[31] Z. Zhang, X. Li, and C. Gan, "Identifying influential nodes in social networks via community structure and influence distribution difference," *Digit. Commun. Netw.*, vol. 7, no. 1, pp. 131–139, Feb. 2021.
[32] A. Ahmad, T. Ahmad, and A. Bhatt, "HWSMCB: A community-based hybrid approach for identifying influential nodes in the social network," *Phys. A, Stat. Mech. Appl.*, vol. 545, May 2020, Art. no. 123590.
[33] S. Gupta and P. Kumar, "An overlapping community detection algorithm based on rough clustering of links," *Data Knowl. Eng.*, vol. 125, Jan. 2020, Art. no. 101777.
[34] G. Glavaš and J. Šnajder, "Event graphs for information retrieval and multi-document summarization," *Expert Syst. Appl.*, vol. 41, no. 15, pp. 6904–6916, Nov. 2014.
[35] N. Abdelaleem, H. A. Elkader, R. Salem, D. D. Salama, and A. Elminaam, "Extractive text summarization using neural network," in *Proc. 36th IBIMA Conf.*, 2020, pp. 13119–13131.
[36] W. Yan and J. Guo, "Joint hierarchical semantic clipping and sentence extraction for document summarization," *J. Inf. Process. Syst.*, vol. 16, no. 4, pp. 820–831, 2020.
[37] J. Batista, R. D. Lins, R. Lima, S. J. Simske, and M. Riss, "Towards cohesive extractive summarization through anaphoric expression resolution," in *Proc. ACM Symp. Document Eng.*, Sep. 2016, pp. 201–204.
[38] M. Bonzanini, M. Martinez-Alvarez, and T. Roelleke, "Extractive summarisation via sentence removal: Condensing relevant sentences into a short summary," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2013, pp. 893–896.
[39] S. Sonawane, P. Kulkarni, C. Deshpande, and B. Athawale, "Extractive summarization using semigraph (ESSg)," *Evolving Syst.*, vol. 10, no. 3, pp. 409–424, Sep. 2019.
[40] D. Parveen, H.-M. Ramsl, and M. Strube, "Topical coherence for graph-based extractive summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1949–1954.
[41] M. Nishino, N. Yasuda, T. Hirao, J. Suzuki, and M. Nagata, "Lagrangian relaxation for scalable text summarization while maximizing multiple objectives," *Inf. Media Technol.*, vol. 8, no. 4, pp. 1017–1025, 2013.
[42] S. Pattnaik and A. K. Nayak, "A simple and efficient text summarization model for Odia text documents," *Indian J. Comput. Sci. Eng.*, vol. 11, no. 6, pp. 825–834, Dec. 2020.
[43] S. A. Anam, A. M. Muntasir Rahman, N. N. Saleheen, and H. Arif, "Automatic text summarization using fuzzy C-means clustering," in *Proc. Joint 7th Int. Conf. Informat., Electron. Vis. (ICIEV), 2nd Int. Conf. Imag., Vis. Pattern Recognit. (icIVPR)*, Jun. 2018, pp. 180–184.
[44] Y. Chali, S. A. Hasan, and S. R. Joty, "A SVM-based ensemble approach to multi-document summarization," in *Proc. 22nd Can. Conf. Artif. Intell.*, Kelowna, BC, Canada, May 2009, pp. 199–202.
[45] E. Griechisch and A. Pluhár, "Community detection by using the extended modularity," *Acta Cybernetica*, vol. 20, no. 1, pp. 69–85, 2011.
[46] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, Jul. 2004, pp. 74–81.
[47] M. Mendoza, C. Cobos, and E. León, "Extractive single-document summarization based on global-best harmony search and a greedy local optimizer," in *Proc. 14th Mex. Int. Conf. Artif. Intell.*, Morelos, Mexico, Oct. 2015, pp. 52–66.
[48] K. Svore, L. Vanderwende, and C. Burges, "Enhancing single-document summarization by combining RankNet and third-party sources," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn. (EMNLP-CoNLL)*, 2007, pp. 448–457.
[49] D. Shen, Q. Yang, and Z. Chen, "Noise reduction through summarization for web-page classification," *Inf. Process. Manage.*, vol. 43, no. 6, pp. 1735–1747, Nov. 2007.
[50] W. Song, J. Z. Liang, and S. C. Park, "Fuzzy control GA with a novel hybrid semantic similarity strategy for text clustering," *Inf. Sci.*, vol. 273, pp. 156–170, Jul. 2014.
[51] J.-Y. Yeh, H.-R. Ke, W.-P. Yang, and I.-H. Meng, "Text summarization using a trainable summarizer and latent semantic analysis," *Inf. Process. Manage.*, vol. 41, no. 1, pp. 75–95, Jan. 2005.
[52] D. M. Dunlavy, D. P. O'Leary, J. M. Conroy, and J. D. Schlesinger, "QCS: A system for querying, clustering and summarizing documents," *Inf. Process. Manage.*, vol. 43, no. 6, pp. 1588–1605, Nov. 2007.
[53] R. M. Alguliyev, Y. N. Imamverdiyev, and F. J. Abdullayeva, "PSO-based load balancing method in cloud computing," *Autom. Control Comput. Sci.*, vol. 53, no. 1, pp. 45–55, Jan. 2019.
[54] C. Fang, D. Mu, Z. Deng, and Z. Wu, "Word-sentence co-ranking for automatic extractive text summarization," *Expert Syst. Appl.*, vol. 72, pp. 189–195, Apr. 2017.
[55] S. Yuan, H. Zeng, Z. Zuo, and C. Wang, "Overlapping community detection on complex networks with graph convolutional networks," *Comput. Commun.*, vol. 199, pp. 62–71, Feb. 2023.
[56] L. Wu, Y. Chen, K. Shen, X. Guo, H. Gao, S. Li, J. Pei, and B. Long, "Graph neural networks for natural language processing: A survey," *Found. Trends Mach. Learn.*, vol. 16, no. 2, pp. 119–328, 2023.

AMREEN AHMAD received the M.Tech. degree in CSE from GGSIPU and the Ph.D. degree in CSE from Jamia Millia Islamia. She is currently an Associate Professor with the Department of Computer Science and Engineering, Galgotia College of Engineering and Technology, Greater Noida, Uttar Pradesh, India. Before that, she was an Associate Professor with Chandigarh University, Chandigarh. She has more than six years of experience in the academics. She has authored more than 20 research papers in reputed conferences and journals, including Web of Science, SCI, and Scopus. She has worked as an editorial board member and a reviewer of various major conferences and journals, including IEEE, Springer, Elsevier, and other international journals with a Scopus index.

TANVIR AHMAD is currently a Full Professor with the Department of Computer Engineering, Jamia Millia Islamia (A Central University), New Delhi, India. He has published over 130 research articles in international journals, books, and conference proceedings, including seven in IEEE/ACM TRANSACTIONS. His research interests include the development of innovative data mining, machine learning, and network analysis techniques to address real-world societal and industrial problems, particularly for text mining, social network analysis, figurative language detection, rumor detection, sentiment and emotion analysis, health informatics, and data-driven cybersecurity.

SARFARAZ MASOOD received the Ph.D. degree in computer engineering, with a focus on the field of artificial neural networks. He is currently an Associate Professor with the Department of Computer Engineering, Faculty of Engineering and Technology, Jamia Millia Islamia, New Delhi, India. He has a rich experience in teaching and research for more than 16 years in the field of computer science and engineering. He also worked in the software industry for around a year before joining academics. His research interests include machine learning and deep learning applications in healthcare, vehicular safety, and agriculture. He has published more than 60 research papers in various reputed international journals and conferences in his research domain. He has been granted an international patent titled "A feature boosted Web-based product purchase recommendation method and system." He is also serving as a Technical Expert for multiple committees at various government institutions in India, including UPSC and CEC-UGC.

MOHD. KHIZIR SIDDIQUI received the B.Tech. degree from BITS, Goa. He was an Undergraduate Researcher in computational social system with IIIT Delhi. He is currently an active Researcher in machine learning and text summarization.

BASMA ABD EL-RAHIEM (Member, IEEE) received the B.Sc. and Master of Science degrees from the Faculty of Science, Menoufia University, Egypt, in 2015 and 2019, respectively. She is currently a Teaching Assistant with the Faculty of Science, Menoufia University. She had published several papers in SCI/SCI journals. Her research interests include deep learning, information security, image processing, and biometrics.

PAWEL PLAWIAK was born in Ostrowiec, Poland, in 1984. He received the B.Eng. and M.Sc. degrees in electronics and telecommunications and the Ph.D. degree (Hons.) in biocybernetics and biomedical engineering from the AGH University of Science and Technology, Kraków, Poland, in 2012 and 2016, respectively, and the D.Sc. degree in technical computer science and telecommunications from the Silesian University of Technology, Gliwice, Poland, in 2020. He is currently the Dean of the Faculty of Computer Science and Telecommunications and an Associate Professor with the Cracow University of Technology, Kraków, and the Deputy Director of Research with the National Institute of Telecommunications, Warsaw, and an Associate Professor with the Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Gliwice. He has published more than 50 articles in refereed international SCI-IF journals. His research interests include machine learning and computational intelligence (e.g., artificial neural networks, genetic algorithms, fuzzy systems, support vector machines, k-nearest neighbors, and hybrid systems), ensemble learning, deep learning, evolutionary computation, classification, pattern recognition, signal processing and analysis, data analysis and data mining, sensor techniques, medicine, biocybernetics, biomedical engineering, and telecommunications. He is an academic editor and a reviewer of many prestigious and reputed journals.

FAHAD ALBLEHAI (Member, IEEE) received the B.S. degree in education in the field of computer, in 2001, the M.S. degree in information technology and communication, in 2010, and the Ph.D. degree in E-learning/web-/internet based teaching and learning, in 2017. He has been an Associate Professor with the Community College, King Saud University (KSU), since 2019. His research interests include web applications, digital transformation, augmented reality, virtual reality, cloud computing, virtual learning environments, E-learning, M-learning, AI, and human–computer interaction.