International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

An Improved Automatic Keyword Extraction using Graph based Modeling

S. Sam Karthik[1], S. Sivakumar[2], P. Yuvaraj[3]
[1]Assistant Professor, Department of Electrical and Electronics Engineering, Dhanalakshmi Srinivasan College of Engineering, Coimbatore, Tamilnadu, India.
[2] Associate Professor, Department of Computer Science and Engineering, Dhanalakshmi Srinivasan College of Engineering, Coimbatore, Tamilnadu, India.
[3]Assistant Professor, Department of Electrical and Electronics Engineering, Dhanalakshmi Srinivasan College of Engineering, Coimbatore, Tamilnadu, India.

Abstract: **Automatic extraction of key terms from a document is indispensable in the digital era to sum up the documents. For instance, instead of reviewing the full document from beginning to end, some of the author's keywords somewhat explain the discussions of the documents. However, the author's keywords are not adequate to recognize the whole idea of the document. Hence the necessity of automatic term extraction methods is essential. The automatic extraction approaches is mostly classified on some techniques such as Natural Language Processing, Statistical approaches, Graph Based approaches, Natural Inspired algorithmic approaches, etc. Even though there are abundant approaches accessible, the exact automatic keyword extraction is a most important dare in areas, which reveals around documents. In this paper, a study of Keyword extraction among Graph based approach is done. In the Graph based approach, the documents are robotically formed as graphs by applying centrality measures during the keyword extraction process. The results of centrality measures were compared and analyzed.**

Keywords: Graph based methods, Keyword Extraction, Centrality measures.

I. INTRODUCTION

Keywords are the least units that can sum up the idea of a document and are frequently used to identify the most appropriate information in a text. To retrieve documents while web hunting or to sum up the documents for cataloging and for such purposes Keywords are used. Keywords in a document present essential information about the content of the document. They can assist the users seek, through information more efficiently or decide whether to read a document or not to read the document. They can also be utilized for a range of language processing duties such as text classification and information recovery.Handing over the keywords physically is unfeasible because, over the past one decade, a enormous growth of computer technology has offered more reasonable and high configuration systems. The bursty data is mounting every day, so we have to preserve and examine the data for efficient use or processing. Data can be available in the form of picture, spatial form, manuscript; mostly manuscript data is represented in loads of ways like text, graphs, predicates, etc. keyword extraction play huge turn in text mining process since, in the broadsheet, social media are used for posting and messaging and all the information of company contained in the form of text.

Automatic keyword extraction is the method of opting terms and phrases from the text document that can finely portray the idea of the document exclusive of any human interference. The objective of automatic keyword extraction is the control and pace of present calculation abilities to crack the troubles in access and healing, and the evils related to information organization without involving human connections.

To classify the enduring expansion of vibrant formless documents is the major challenge and managing such unorganized documents causes pricier. The clustering of such active documents helps us to decrease the price. Document clustering by analyzing the keywords of the documents is one of the best methods to organize the unstructured dynamic documents.

II. LITERATURE REVIEW

Florian Boudin et al. [1] In this section, the author assess the past workings on keyword extraction methods and declare how unique these methods are discussed. Floarin Boudin presented the centrality measures, comparison for extraction of keywords in Graph based approaches, the closeness centrality obtains the optimum results on short – documents.

Rada Mihalcea et al. [2] The author proposed an innovative unsupervised method for automatic sentence extraction using graph based ranking algorithms and evaluate the method in the context of a text summarization task, and show that the results obtained compare favorably with previously published results on established benchmarks.

R.Nagarajan et al. [3] The authors have projected the new Graph based keyword extraction algorithm, the terms as vertices, the relationship between the terms as arcs, and the projected algorithm which gives more accurate result.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

Sonawane et al. [4] presented a centrality measure and compare the five centrality measures Degree centrality, Betweeness Centrality, Closeness Centrality, Eigenvector centrality, and Text rank). The authors deal with the document BOW, and finally, the authors say the graph based text representation is the best way and the result is better than the traditional model.

Fragkiskos et al. [5] introduced a novel graph-based approach for text categorization. Contrary to the traditional Bag-of-Words model for document representation, the author consider a model in which each document is represented by a graph that encodes relationships between the different terms. The importance of a term to a document is indicated using graph-theoretic node centrality criteria.

Marina Litvak et al. [6] introduced and compare between two novel approaches, supervised and unsupervised, for identifying the keywords to be used in extractive summarization of text documents which enhances the traditional vector-space model by taking into account some structural document features.

Alwan M U et al. [7] suggested that MCL graph clustering algorithm (Markov Cluster Algorithm) can simplify information processing by identifying the characteristics of each vertex in the graph so that it will establish cluster vertices with a specific label. Further the process of expanding and inflating matrix will be the main process in the clustering of digital news that has been transformed into a graph database models that expand aims to show a new edge and remove the old edge in common not needed in the graph.

F.Sebastiani et al. [8] The automated categorization (or classification) of texts into predefined categories has witnessed a booming interest in the last 10 years, due to the increased availability of documents in digital form and the ensuing need to organize them. In the research community the dominant approach to this problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of pre-classified documents, the characteristics of the categories.

Adrien Bougouin et al. [9] the author states community the dominant approach to this problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of pre classified documents, the characteristics of the categories. The advantages of this approach over the knowledge engineering approach (consisting in the manual definition of a classifier by domain experts) are a very good effectiveness, considerable savings in terms of expert labor power, and straightforward portability to different domains.

Florian Boudin et al. [10] the author has dealt with key term extraction which to identify the main topics of the documents by topic rank, candidates key phrases which are clustered into topics, and the topic rank significantly out performs the state-of-art methods in graph analysis.

Shibamouli Lahiri et al.[11] the author discussed problems for keyword and key phrase extraction in NLP by applying the page rank algorithm. The authors have presented a survey for text summarization techniques, which deals with the necessity of text summarization; we review the various processes of text summarization and express the viability and deficiencies of the different techniques.

Mehdi Allahyari et al. [12] in recent years, there has been a explosion in the amount of text data from a variety of sources. This volume of text is an invaluable source of information and knowledge which needs to be effectively summarized to be useful. In this review, the main approaches to automatic text summarization are described. We review the different processes for summarization and describe the effectiveness and shortcomings of the different methods.

Paolo Tonelia et al. [13] reverse engineering techniques have the potential to support Web site understanding, by providing views that show the organization of a site and its navigational structure. However, representing each Web page as a node in the diagrams that are recovered from the source code of a Web site leads often to huge and unreadable graphs. Moreover, since the level of connectivity is typically high, the edges in such graphs make the overall result still less usable.

C.Abi Chahine et al. [14] author proposes an innovative method for an indexing support system. This system takes as input an ontology and a plain text document and provides as output contextualized keywords of the document. The method has been evaluated by exploiting Wikipedia's category links as a termino-ontological resources.

Santhosh Kumar et al.[15] introduced an investigation for extracting the key term without human intervention in text summarization, talk about the various tactic used for key term extraction and text summarization.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

A. Keyword Extraction Methods

Keyword extraction methods, mainly, are classified into Supervised and Unsupervised approaches. In the Supervised approach, training dataset is essential. In the Unsupervised approach, the approach doesn't need training data. In this paper, Graph based methods (Unsupervised methods) of keyword extraction have been analyzed.

B. Simple Statistics

These approaches are easy and don't need training information. The keyword statistics can be utilized to formulate the key terms:- n gram measurements, word recurrence, TFIDF, word co-event are some of the examples.

C. Linguistics

This approach utilizes the dialectal properties of the words, sentences and documents. A portion of the etymology includes the syntactic, lexical, etc. For NLP issues this technique can be applied.

D. Machine Learning

This approach is based upon the processed information to extract the keyword. It needs manual explanations for the learning dataset which is extremely tedious and mismatched. The SVM (support vector machine), and the Naïve Bayes are some of the examples.

E. Graph Based

These approaches are easy mathematical models, which allow the study of the associations and make the structural information significant. The graph is a discrete data structure consisting of nodes and edges ={V, E}

Vertices also referred to as node, and Edges are the lines or arcs, which connect the nodes in the graph. The document is modeled as graph the terms (words) are denoted by vertices (nodes) and their association is denoted as edges (link). In the Graph based method, the graph describes the significance of the document visually. At the start, the document is transformed into graph, and the keywords of the document are established. Graph nodes symbolize only important words of the documents

F. Other approaches

These approaches are integrated approaches of standard approaches such as Statistical approaches, Linguistics approaches, Machine learning approaches, and Graph based approaches.

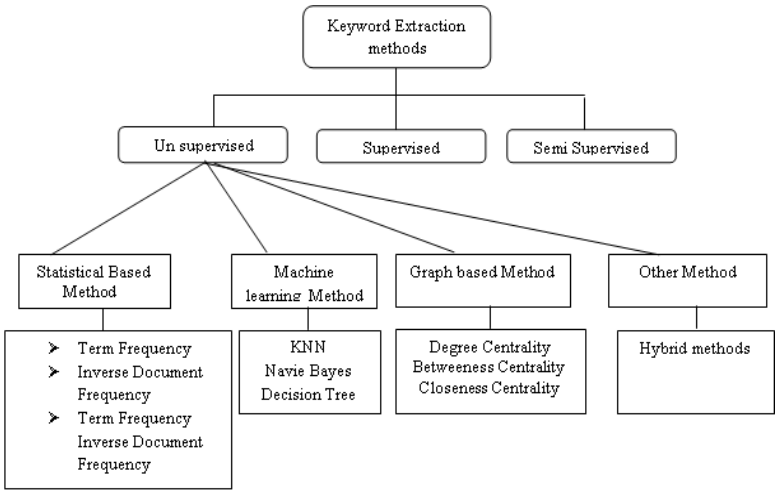

Fig. 1. Different Keyword Extraction methods

III. METHODOLOGY

This section focuses concentration on the various accessible keyword extraction approaches found in the Graph based methods.

A. Graph Based Method

Usually, Graphs are mathematical model, generally graphs as represented as G= {V,E}

V- Vertices

E- Edges

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

In Graph based keyword extraction methods, the important terms are denoted as vertices, the association among the vertices are connected, the connections are denoted as Edges. Centrality measures are mostly used to locate the key terms from the Graph in the Graph based keyword extraction methods

Degree Centrality

TheDegree centrality, it takes in to description the number of the adjoined nodes. If the network is directed, two types of the measures, in-degree : calculates the number of innermost links or the number of predecessor nodes; out-degree: computes the number of leaving links or the number of successor nodes.

As per degree centrality concerns, a node is important if it has many adjoined nodes.

The degree centrality of a node v_i is calculated as:

$$CD(Vi) = \frac{|N(Vi)|}{|V|-1} \quad (1)$$

where,

- $CD(V_i)$ is the degree the centrality of node V_i
- V is the set of nodes
- $N(V_i)$ is the set of nodes connected to the node V_i

Closeness Centrality

This centrality depicts as the equivalent of the cumulative of separations of all nodes to certain nodes, i.e., contrary of farness. The closeness centrality of the node V_i is given in the following equation.

In a linked graph, the closeness centrality of a node is a calculated centrality in a graph, which is calculated as the sum of the space of the shortest paths between the node and all other nodes in the graph. Thus, the more vital a node is, the nearer it is to all other nodes.

$$C_C(V_i) = \frac{(|V|-1)}{Vi€V\ dist(Vi,Vj)} \quad (2)$$

where

- The $C_C(V_i)$ is closeness centrality of the node V_i
- The V is the set of nodes (words) in the graph G
- The dist(V_i,V_j) is the shortest distance between nodes V_i and V_j

Betweeness Centrality:

Betweeness Centrality is a method of centrality in a graph depending on the smallest way. For each couple of vertices in a connected graph, there exists, at any, rate one most short way among the vertices to such an extent that either the quantity of edges that the way goes through (for unweighted charts) or the whole of the loads of the edges (for weighted charts) are limited. In the betweeness centrality, for every vertex, the quantity of these most limited ways goes through the vertex.

$$C_B(V_i) = \frac{\sum_{V_i \neq V_j \neq V_k \in V} \frac{\sigma(V_j, V_k|V_i)}{\sigma(V_j, V_k)}}{(|V|-1)(|V|-2)/2} \quad (3)$$

Where

The $C_B(Vi)$ is Betweeness Centrality of node V_i

The $\sigma(Vj, Vk)$ is the number of shortest paths from node$V_j$to nodeV_k

The $\sigma(V_j, V_k|\ V_i)$ is the number of those path that pass through the node V_i

IV. EXPERIMENTAL RESULT

To analyze graph based keyword extraction methods, 14 documents have been taken, which are shown in the following Table -1.

Table–I: Sample Documents

Document Id	Document
D1	In imaging science, image processing is processing of images using mathematical operations by using any form of signal processing.
D2	In Image Processing, the input is an image, a series of images, or a video, such as a photograph or video frame; the output of image processing may be either an image or a set of characteristics or parameters related to the image.
D3	Most image-processing techniques involve treating the image as a two-dimensional signal and applying the standard signal-processing techniques to it.
D4	Images are also processed as three-dimensional signals where the third-dimension being time or the z-axis.
D5	Image processing usually refers to digital image processing, but the optical and analog image processing also are possible.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

D6	This article is about general techniques that apply to all of them. The acquisition of images (producing the input image in the first place) is referred to as imaging. Closely related to image processing are computer graphics and computer vision.
D7	In computer graphics, images are manually made from physical models of objects, environments, and lighting, instead of being acquired (via imaging devices such as cameras) from natural scenes, as in most animated movies.
D8	Computer vision, on the other hand, is often considered high-level image processing out of which a machine/computer/software intends to decipher the physical contents of an image or a sequence of images (e.g., videos or 3D full-body magnetic resonance scans).
D9	Computer graphics are pictures and movies created using computers, such as CGI - usually referring to image data created by a computer specifically with help from specialized graphical hardware and software.
D10	It is a vast and recent area in computer science. The phrase was coined by computer graphics researchers Verne Hudson and William Fetter of Boeing in 1960.
D11	Another name for the field is computer-generated imagery, or simply CGI. Important topics in computer graphics include user interface design, sprite graphics, vector graphics, 3D modeling, shaders, GPU design, and computer vision, among others.
D12	The overall methodology depends heavily on the underlying sciences of geometry, optics, and physics. Computer graphics is responsible for displaying art and image data effectively and beautifully to the user, and processing image data received from the physical world.
D13	The interaction and understanding of computers and interpretation of data has been made easier because of computer graphics.
D14	Computer graphic development has had a significant impact on many types of media and has revolutionized animation, movies, advertising, video games, and graphic design generally.

In table -1, the first column indicates the document ID, the second column indicates documents. At this point, the documents include all the terms. In the analysis, the standard preprocessing techniques is used in sinking the size of the document. The standard preprocessing techniques include Tokenization, Stop word Removal and Stemming. Later than applying the preprocessing techniques, the term 'in', 'is', 'of',' using', 'by', 'any', 'form', and 'of' are removed with the help of stop word removal technique. Similarly, the terms 'imaging', and 'images' denote the root word 'image'. With the assist of stemming process, this technique identifies the root word, and all the documents are automatically preprocessed. Lastly, the noise detached terms are maintained in the following table-2 for the further process.

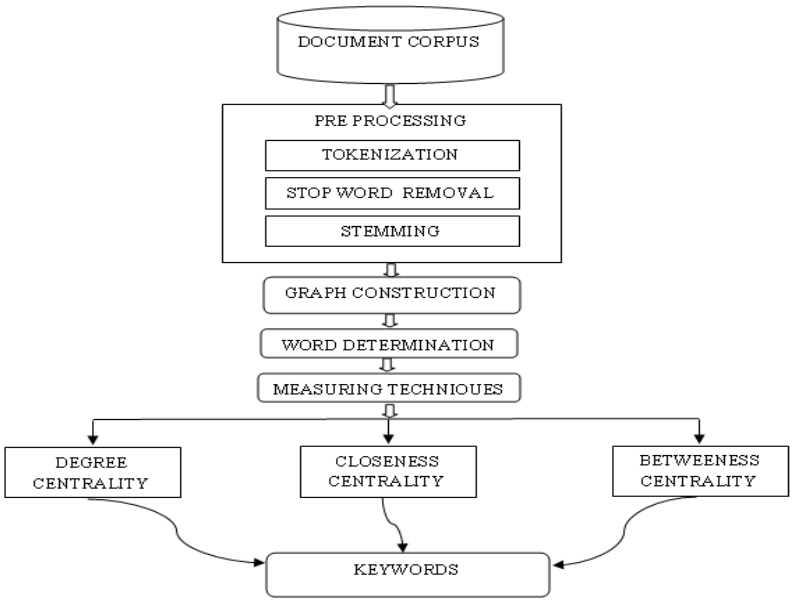

Fig. 2. Proposed Graph based Keyword Extraction method

Table-II: Extracted Terms in the Documents

Document Id	Extracted Terms
D1	'IMAGE' , 'SCIENCE', 'PROCESSING', 'SIGNAL', 'MATHEMATICAL', 'OPERATION'
D2	'IMAGE', 'SERIES', 'PROCESSING', 'VIDEO', 'PHOTOGRAPH' ,'FRAME'
D3	'IMAGE', 'PROCESSING', 'TELEVISION', 'TWO-DIMENSION', 'SIGNAL'
D4	'IMAGE', 'PROCESSING', 'THREE-DIMENSION', 'Z-AXIS' ,'SIGNAL'
D5	'IMAGE', 'PROCESSING', 'DIGITAL', 'ANALOG', 'OPTICAL'
D6	'ACQUISITION', 'IMAGE', 'PROCESSING', 'COMPUTER' ,'VISION', 'GRAPHICS'
D7	'COMPUTER','GRAPHICS','IMAGE','ACQUIRED','LIGHT','ENVIRONMENT' , 'PHYSICAL' ,'MODEL', 'OBJECT', 'ANIMATION', 'MOVIE'
D8	'COMPUTER', 'VISION', 'IMAGE', 'PROCESSING'
D9	'COMPUTER' ,'GRAPHICS', 'PICTURE', 'MOVIE', 'IMAGE', 'DATA', 'SOFTWARE'
D10	'COMPUTER' 'SCIENCE' 'GRAPHICS'
D11	'COMPUTER' 'VIDEO' 'IMAGE' 'GRAPHICS' 'SPRITE' 'VECTOR'
D12	'SCIENCE' 'OPTICS' 'PHYSICS' 'GEOMETRY' 'COMPUTER' 'GRAPHICS' 'IMAGE' 'DATA' 'PROCESSING'
D13	'COMPUTER' 'DATA' 'GRAPHICS'
D14	'COMPUTER' 'GRAPHICS' 'ANIMATION' 'MOVIE' 'ADVERTISEMENT' 'VIDEO GRAPH'

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

A. Graph Construction

At this phase, the undirected word graph is built for all documents in a corpus, in which the documents are denoted as graph. In each document, the terms are indicated as nodes; and the repetitive associations, among the words of the documents, are referred as the edges. By using syntactic filters, the vertices of graph are filtered; the repetitive count of the words determines the arcs.

By using the above said strategy, the extracted terms, from the table-2, were in use to built the graphs, which best reveal the document's formation. During the graph creation, the association between the nodes is drawn as the edges based on their position in the document. For instance, the subsequent graph-1 denotes document id 1. From the subsequent graph, it is observed that the nodes image, processing, science and signals form a separate connected component of the graph and the nodes, mathematics and operation form a separate connected component of the graph. This shows the document id 1, which deals with two slightly different concepts.

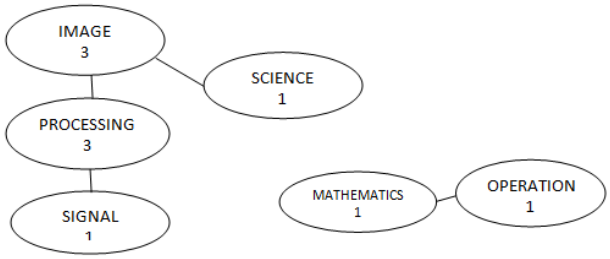

Fig. 3. Graph representation of D1

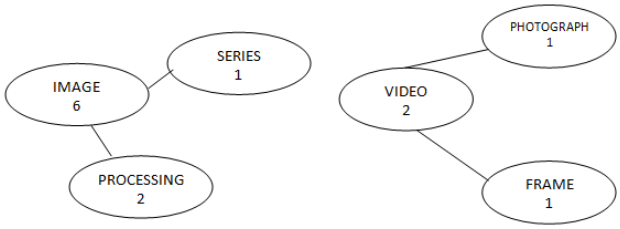

Fig. 4. Graph representation of D2

The following graph 2 - represents the graph form of document id 2. From the above graph, it is observed that the nodes image, processing, and series form a part connected component of a graph and the nodes videos, photograph and frame form a separate connected component of the graph. This shows the document id 2, which portray the two different concepts (indirectly related). Likewise, the remaining documents have been processed. Similarly all the 14 documents were converted into graphs.

B. Word Determination

When the word graph is built, the important word determination pace is pursued, for which definite centrality estimates are useful to dole out the location of every node in a graph. In graph hypothesis, centrality measures allude the markers which identify the most important vertices inside a graph and that approach is utilized for the errand of positioning the nodes.

Table-V: Closeness Centrality

Term	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	C_C(Vi)
Image	0.7500	1.0000	0.6600	0.4444	0.8000	0.5000	0.6666	0.6000	0.4666	0.0000	0.4545	0.8000	0.0000	0.0000	0.5102
Computer	0.0000	0.0000	0.0000	0.0000	0.0000	0.6250	0.4705	0.3333	0.6363	1.0000	0.7142	0.4444	1.0000	0.5555	0.4128
Graphics	0.0000	0.0000	0.0000	0.0000	0.0000	0.4166	0.3333	0.0000	0.5384	0.6666	0.7142	0.6666	0.6666	1.0000	0.3573
Processing	0.7500	0.6600	0.8000	0.6666	0.5000	0.6250	0.0000	0.4000	0.0000	0.0000	0.0000	0.5714	0.0000	0.0000	0.3552
Science	0.5000	0.6600	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.6666	0.0000	1.0000	0.0000	0.0000	0.2019
Signal	0.5000	0.0000	0.6600	0.6666	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1305
Data	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3333	0.0000	0.0000	0.5714	0.6666	0.0000	0.1122
Video	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5555	0.1111
Vision	0.0000	0.0000	0.0000	0.0000	0.0000	0.4166	0.0000	0.4615	0.0000	0.0000	0.4545	0.0000	0.0000	0.0000	0.0952

Table-III: Degree Centrality

Term	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	C_D(Vi)
Image	0.4000	0.4000	0.5000	0.2500	0.7500	0.4000	0.5000	0.3300	0.2800	0.0000	0.2000	0.3750	0.0000	0.0000	0.3132
Computer	0.0000	0.0000	0.0000	0.0000	0.0000	0.6000	0.2000	0.1600	0.4200	1.0000	0.6000	0.1250	1.0000	0.2000	0.3075
Graphics	0.0000	0.0000	0.0000	0.0000	0.0000	0.2000	0.1000	0.0000	0.4200	0.5000	0.6000	0.2500	0.5000	1.0000	0.2550
Processing	0.4000	0.2000	0.5000	0.5000	0.2500	0.4000	0.0000	0.1600	0.0000	0.0000	0.0000	0.2500	0.0000	0.0000	0.1900

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

Science	0.2000	0.2000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5000	0.0000	0.3750	0.0000	0.0000	0.0911
Signal	0.2000	0.0000	0.5000	0.2500	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0679
Data	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1400	0.0000	0.0000	0.2500	0.5000	0.0000	0.0636
Vision	0.0000	0.0000	0.0000	0.0000	0.0000	0.2000	0.0000	0.3300	0.0000	0.0000	0.2000	0.0000	0.0000	0.0000	0.0521
Video	0.0000	0.4000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2000	0.0429

Table-IV: Betweeness Centrality

Term	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**	**D11**	**D12**	**D13**	**D14**	**C_B(Vi)**
Image	2.0000	1.0000	1.0000	0.0000	5.0000	4.0000	24.0000	8.0000	6.0000	0.0000	0.0000	4.0000	0.0000	0.0000	3.9286
Computer	0.0000	0.0000	0.0000	0.0000	0.0000	6.0000	7.0000	0.0000	10.0000	1.0000	6.0000	0.0000	1.0000	0.0000	2.2143
Graphics	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	10.0000	0.0000	6.0000	3.0000	0.0000	10.0000	2.0714
Processing	2.0000	0.0000	3.5000	3.0000	0.0000	6.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0357
Physical	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	12.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.8571
3d	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	8.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5714
Model	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	7.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5000
Picture	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	6.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4286
3-dimension	0.0000	0.0000	0.0000	5.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3571

In the area of keyword extraction, different centrality measures are utilized for the errand of positioning the words in documents.

C. Centrality Measures

After the conversion of documents into graphs, to extract the important terms from the graphs, three centrality measures namely Degree Centrality, Betweeness Centrality and Closeness Centrality measures had been applied to the graphs. The resulted numeric values were placed in Table 3, Table 4 and Table 5 respectively.

- *Degree Centrality:* Initially, when this method is applied for all the documents, the result of Degree centrality method with 40 keywords are identified based on the thrash hold value (automatically assigned based on the size of the document), which will minimize the size of the keyword. Finally top-ranked keywords are taken are shown in the Table 3.

- *Betweeness Centrality:* While the Betweeness centrality method is applied for all the 14 documents, result of the Betweeness centrality method 16 keywords are identified based on the thrash hold value, it will produce top-ranked keywords are shown in the Table 4.

- *Closeness Centrality:* Finally, when the Closeness centrality method is applied for the same documents, the result of the Closeness centrality method with 40 keywords are identified, based on the thrash hold value the top-ranked keywords are identified in the Table 5.

In the graph based methods Degree centrality, Betweeness centrality and Closeness centrality results are analyzed and the first four top-ranked keywords are identified: which are, 'image', 'computer', 'graphics', 'processing'. The result is the same for all the three
methods. The remaining keywords are found with same degree of centrality and Closeness centrality, and little difference in Betweeness centrality, since the properties of the Degree centrality and Closeness centrality are nearly same.

The node appearing in the graph gets some weight, but in Betweeness centrality it is different from above methods; in the Betweeness centrality, the method weight of the node is calculated only when the particular node has more number of shortest paths pass via the node. In this nature, the Degree centrality and the Closeness centrality is found more similar than the Betweeness centrality.

V. CONCLUSION

In this paper, an endeavor has been made to investigate the Automatic Keyword Extraction Methods of Graph Based Approaches. In Graph based approach, top position keywords were extracted by applying the centrality measures on graphs, which symbolize the documents. Because of the Centrality measures, the important terms of graphs, which denote the document is extracted wisely. To sum up, graph based approaches, for keyword extraction, may yield better results.

REFERENCES

[1] Florian Boudin, "A Comparison of Centrality Measures for Graph Based Keypharse Extraction," International Joint Conference on Natural Language Processing, 2013, p.834-838.
[2] Rada Mihalcea , "Graph Based Ranking Algorithms for Sentence Extraction, applied to Text Summarization," Proceedings of ACL 2014, p.8-12.
[3] R.Nagarajan, S.Anu H Nair, P.Aruna, N.Puviarasan, "Keyword Extraction using Graph Based Approach," International Journal of Advanced Research in Computer Science and Software Engineering, 2016, p.25-29.
[4] S.S. Sonawane, P.A. Kulkarni, "Graph Representation and analysis of Text Document: A Survey of Techniques, International Journal of Computer Applications," Volume 96 – No. 19, June 2014.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

[5] Fragkiskos D.Malliaros, Konstantinos Skianis, "Graph Based Term Weighting for Text Categorization," IEEE/ACM International Conference on Advanced in Social Networks Analysis and Mining, 2015, p.1473-1479.
[6] Marina Litvak, Mark Last,Graph Based Keyword Extraction for Single-Document Summarization, Coling,2008, Proceedings of the workshop Multilingual Information Extraction and Summarizartion, p.17-24.
[7] Alwan M Ubaidillah Al-Fath, Kemas Rahmat Saleh W., M.Eng, Siti Sa'adah,M.T., "Implementation of MCL Algorithm in Clustering Digital News with Graph Representation," IEEE, 2016.
[8] F.Sebastiani, "Machine Learning in automated text categorization," ACM Computer surv., Vol. 34, no.1, p.1-47, 2002.
[9] Adrien Bougouin, Florian Boudin, Beatrice Daille, "Topic Rank: Graph Based Topic Ranking for Keypharse Extraction," International Joint Conference on Natural Language Processing, 2013, p.543-551.
[10] Florian Boudin, "A Comparison of Centrality Measures for Graph Based Keypharse Extraction," International Joint Conference on Natural Language Processing, 2013,p.834-838.
[11] Shibamouli Lahiri, Sagnik Ray Choudhury, Cornelia Caragea, "Keyword and Keypharse Extraction Using Centrality Measures on Collocation Networks," arXiv:1401.6571vl [cs.CL], 2014.
[12] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi, Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierres, Krys Kochut, Text Summarization Techniques: A Brief Survey, arXiv, USA, 2017.
[13] Paolo Tonelia, Filippo Ricca, Emanuele Pianta and Christian Girardi, "Using Keyword Extraction for website Clustering," Proceedings of the Fifth IEEE International Workshop on Web Site Evolution (WSE'03), 2003.
[14] C.Abi Chahine, N. Chaignaud, JPh Kotowicz, JP Pecuchet, "Context and Keyword Extraction in Plain Text Using a Graph Representation," IEEE, 2008.
[15] Santhosh Kumar Bharathi, Korra sathya Babu, Sanjay Kumar Jena, "Automatic Keyword Extraction for Text Summarization: A Survey," NIT, Rurkela, Odisha,2017.