Volume 6, Issue 10, October 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Keyword Extraction using Graph Based Approach

R. Nagarajan, Dr. S. Anu H Nair, Dr. P. Aruna, N. Puviarasan

Department of Computer Science & Engineering, Annamalai University,
Tamilnadu, India

Abstract— Organizing the rapid dynamic growth of unstructured documents is the major challenge. The documents are well utilized, when it is organized. Document clustering using keywords is one of most popular trendy method. In this research paper, a novel graph based keyword extraction algorithm has been proposed. In this algorithm, documents are represented as graphs, words of the documents are represented as nodes, and the relation between the words of the documents is represented as edges. This proposed algorithm gives more than 90% of accuracy.

Keywords— keyword extraction, graph based algorithm, document clustering, mathematical model, centrality Measures.

I. INTRODUCTION

Keywords can be defined as a small set of words or word phrases of the document that can best describe the meaning of the document. Keywords are high level summary of a document, presenting the concept of the document. Different terminology is used in studying the words that represents related information contained in the document, key phrase, key segments, key terms or just keywords. Keywords are more essential and widely used in areas such as Information Retrieval (IR), Natural Language Processing (NLP) etc. Relevant keywords can be used to make an automatic index for a document collection or alternatively can be used for document representation in categorization, classification or clustering tasks. Key word extraction plays a vital role in acquiring critical information from the document to improve the quality of the document processing applications like document summarization, document clustering, document indexing, data mining, information filtering, information retrieval etc. Keyword extraction refers to how to extract certain words or word phrases from a document automatically to present the document's subject accurately and can be considered as the core technology of all automatic processing of the documents.

Graph based keyword model is a mathematical model, which enables exploration of the relationships and structural information very effectively. Document is modeled as graph where terms are represented by vertices and relations among terms are represented by edges. Edge relation between two terms can be established on many principles exploiting different text range or relations for the graph construction.

II. LITERATURE REVIEW

In this section we review previous work on keyword extraction algorithms and how these algorithms differ themselves. In [1] presented a comparison of five centrality measures (Degree centrality, Closeness centrality, Betweenness centrality, Eigenvector centrality and TextRank) for graph-based Keyphrase extraction. In [2] proposed Conditional Random Field based keyword extraction approach. CRF model is a state-of-the-art sequence labelling method, which uses the features of documents more sufficiently and effectively and at the same time keyword extraction can be considered as the string labelling. In [3] presented SemanticRank, a graph-based ranking algorithm for keyword and sentence extraction from text. This proposed algorithm constructs a semantic graph using implicit links, based on semantic relatedness between nodes and consequently ranks nodes using different ranking algorithms. In [4] proposed and validated two graph-based methods, supervised and unsupervised graph based syntactic representation approaches, for cross-lingual keyword extraction to be used in extractive summarization of text documents. In [5] presented an innovative unsupervised approach for automatic extraction of sentences using graph-based ranking algorithms, this proposed algorithm identifies the important sentences in a text, a sentence recommends another sentence that address similar concepts, which helps to understand the text. Sentences that were highly recommended by other sentences were likely to be more informative for the given text and given high score. In [6] surveys methods and approaches for the task of keyword extraction, systematic review of methods was gathered to attain the review of existing approaches. In [7] applied the concept of k-core on the graph of words representation of text for single document keyword extraction maintaining only the nodes from the main core as addressing terms. In [8] an algorithm keygraph was proposed for keyword extraction, this algorithm based on the segmentation of a graph, represented the co-occurence between terms in a document, which forms clusters, each cluster represents the concept of the document and top ranked terms by a statistic based on each term's relationship to these clusters were selected as keywords. In [9] introduced TextRank a graph based ranking model for text processing used successfully in natural language applications. Furthermore, two innovative unsupervised approaches for keyword and sentence extraction has been proposed. The important aspect of the proposed TextRank was it does not require any deep linguistic knowledge, domain or language specific annotated corpora.

III. GRAPH BASED METHOD

Generally, a graph is an ordered pair *G = (V, E)* where is the set of vertices and E⊆VxV is the set of edges. A graph is directed if the edges have a direction associated with them. A graph is weighted if there is a weight function w that assigns value (real number) to each number. We use $N = |V|$ and $K = |E|$ as shorthand for the number of vertices and edges in a graph. A path in a graph is a chain of edges which connects a chain of vertices which are distinct from one another. A shortest path between two vertices u and v is a path with the shortest length and it is called distance between u and v.

IV. VARIOUS TYPES OF GRAPHICAL KEYWORD EXTRACTION METHODS

The classifications of graph type widely focused on Vertices and Edges. In vertex representation models, vertices represent advanced concepts which can be homogeneous or heterogeneous (Slobodan Beliga et al., 2015). The homogeneous graph model is generally used to represent the grammatical associations between words or semantic similarities. Additionally, vertices be weighted or unweighted which conditions the representation model, which is respectively weighted graph or unweighted graph. Weighted vertices in this case commonly indicate the importance of the vertex in the graph, and different measures are used to calculate the importance of vertex.

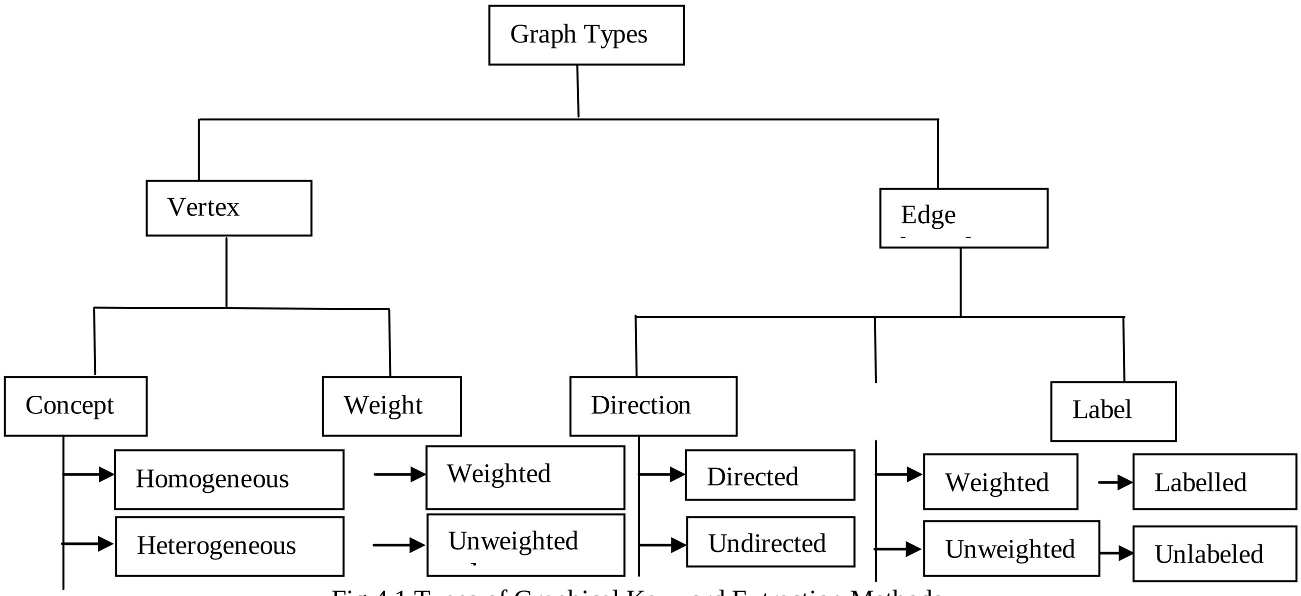

Fig.4.1 Types of Graphical Keyword Extraction Methods

In Edge representation model, graphs can be either directed (called digraph, e.g for word order in text) or undirected (for connecting related words). Edges can also be weighted or unweighted, depending on relationships between vertices. In a language complex network, weight could be the distance of two words in paragraphs or text or the frequency of word pair's co-occurrences. Besides weights, edge models can be labelled or unlabeled, it is almost conventional to explain the relationships or rules between related vertices by the edge label in many graph models. In related work of graphs in the language's edge label can be denote POS (part of speech), grammatical rule of word etc.

V. KEYWORD EXTRACTION USING GRAPH BASED MODEL

Extraction of keywords from a document can be divided in three steps

1. Graph Construction
2. Word determination
3. Keyword generation

5.1 Graph Construction

In this first step, documents are represented as graph, an undirected word graph is constructed for each document in a document corpus, in which each of words of the document are represented as nodes, the co-occurrence relations between words of the documents are represented as edges. Nodes of graph are filtered by syntactic filters. Edges are weighted according to the co-occurrence count of the words they connect.

5.2 Word Determination using Centrality Measures

Once the word graph is constructed, the important word determination step is followed, for which certain centrality measures applied to assign the rank to each node in a graph. In graph theory, centrality measures refer to indicators which indentify the most important vertices within a graph and that approach is used for the task of ranking the vertices. In the domain of keyword extraction, various centrality measures are used for the task of ranking the words in a text.

Centrality measure is local graph measures, focused on a single vertex and its neighbourhood in a graph. The neighbourhood of a vertex in a graph G is defined as a set of neighbours of a vertex *v* and is denoted by *N(v).* The neighbourhood size is the number of immediate neighbours to a vertex. The number of edges between all neighbours of a vertex is denoted by E(v). In the directed graph, the set of *Nin(v)* is the set of vertices that point to a vertex *v(predecessors)* and set of *Nout(v)* is the set of vertices that vertex *v* points to *v(successors).*

The clustering coefficient of a vertex measures the density of edges among the immediate neighbours of a vertex. It determines the probability of the presence of an edge between any two neighbours of a vertex. It is calculated as a ratio between the number edges *Ei* that actually exist among these and the total possible number of edges among neighbours.

i) Degree Centrality

In this centrality measure, degree centrality of a node is defined by the number of edges incidents the node (word). The higher of number edges incidents the node shows the stronger node (word) to the graph (document), alternately nodes has lesser edge incidents shows the weaker node to the graph. The degree centrality of a node v_i is given by

$$CD(V_i) = |N(V_i)| / |V|-1 \qquad (5.1)$$

where

- $CD(V_i)$ is the degree 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

ii) Closeness Centrality

Closeness Centrality is defined as the reciprocal of the sum of distances of all nodes to some node, i.e., inverse of farness. The Closeness Centrality of a node V_i is given by the following equation

$$CC(Vi) =(|V|-1) / \sum_{Vi€V} dist(V_i,V_j) \qquad (5.2)$$

where

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

5.3 Keyword Generation

This is twostep process, first, keywords are extracted from the document. Sequence of adjacent words are considered as keyword candidates. In the second step, the score of a candidate keyword k is computed by summing the scores of the words it contains normalized by its length+1. The score can be given by the equation

$$Score(k) = (\sum_{word€k} Score(word)) / (length(k)+1) \qquad (5.3)$$

After scoring the keywords, redundant keywords are eliminated and the resulting keywords are ranked by the descending scores of the keywords.

Algorithm 1 presents the steps to extract keyword from graph based approach.

Step 1	:	*For the given document D, construct word graph G*
Step 2	:	*Do for each w € W (W is the words in a document D)*
Step 3	:	*For each edge(w1,w2) in G, calculate the centrality measure (w1,w2)*
Step 4	:	*Update word importance until convergence, get related word rank rw*
Step 5	:	*Merge rw(w€W) and generate final word rank r.*

VI. EXPERIMENTAL RESULTS

To test our proposed graph based keyword extraction algorithm, the following sample of text have been taken, which consist of 398 words in 13 sentences.

In imaging science, image processing is processing of images using mathematical operations by using any form of signal processing. 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. Most image-processing techniques involve treating the image as a two-dimensional signal and applying standard signal-processing techniques to it. Images are also processed as three-dimensional signals where the third-dimension being time or the z-axis. Image processing usually refers to digital image processing, but optical and analog image processing also are possible. 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. 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. 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). 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. 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. 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. 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. The interaction and understanding of computers and interpretation of data has been made easier because of computer graphics. 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.*

Fig. 5.1 Sample of text

By applying our proposed graph based keyword extraction algorithm, a graph is constructed as in the Fig. 5.2.

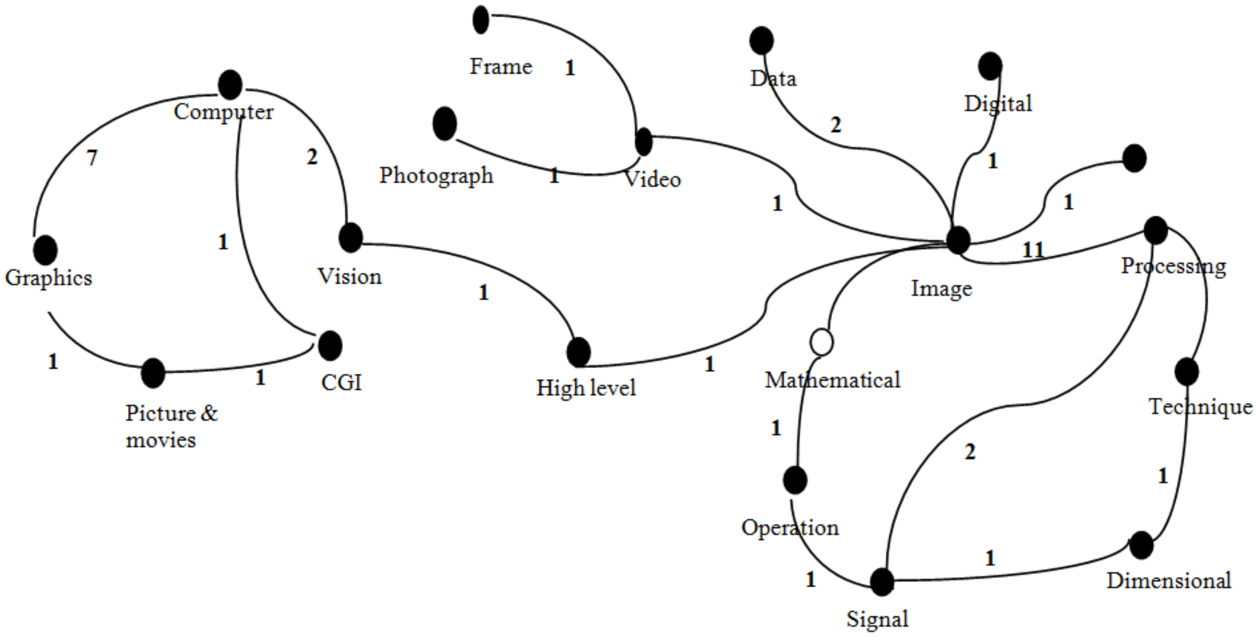

Fig. 5.2 Graph representation of sample text

Figure 5.2 shows the graph representation of our sample text (Figure 5.1), this graph is undirected weighted graph. In which nodes represents the words of the sample text and edges represents the co-occurrences relationship between the words. From the graph, it is interesting to note that the sample text majorly discussed about Image processing and computer graphics, and also there is a relation between the two discussed concepts. The numbers mentioned in between the nodes shows the number of co-occurrence between two nodes in the sample text. For example, in the sample text nodes 'image' and 'processing' co-occurs 11 times, nodes 'computer' and 'graphics' co-occurs 7 times, likewise nodes 'computer' and 'vision', 'image' and 'data', 'signal' and 'processing' co-occurs 2 times.

The remaining nodes co-occur only one time. The higher number of co-occurrence nodes are stronger keyword of the sample text. The following Table 5.1 shows the degree of centrality of the nodes in the graph of Figure 5.2.

Table 5.1 Degree of Centrality measures on Sample Text

S.No.	Nodes (V_i))	Incidents	CD(V_i)
1	Image	7	0.3684
2	Processing	3	0.1578
3	Signal	3	0.1578
4	Computer	2	0.1052
5	Graphics	2	0.1052
6	Vision	2	0.1052
7	Highlevel	2	0.1052
8	Picture and Movies	2	0.1052
9	CGI	2	0.1052
10	Techniques	1	0.0526
11	2 Dimensional	1	0.0526
12	Mathematical	1	0.0526
13	Operation	1	0.0526
14	3 Dimensional	1	0.0526

15	Digital	1	0.0526
16	Science	1	0.0526
17	Data	1	0.0526
18	Video	1	0.0526
19	Frame	1	0.0526
20	Photograph	1	0.0526

The higher value in degree of centrality measure of a node comes in top ranks, from the table 3.4, it is very clear that the highest centrality measure value of node is ‘image’, so it is ranked as first, followed by ‘processing’, ‘graphics’, ‘vision’, ‘highlevel’, ‘pictures and movies’ and ‘CGI’. The centrality value lesser than 0.01 are not considered as keywords here.

Closeness centrality measure is used to measures the linkage strength between the nodes, this measure is the reciprocal of the sum of distances of all nodes to some node. Because of this inverse fairness, the lower value nodes are ranked top. Table 5.2 shows the closeness measures of sample text.

Table 5.2 Closeness Centrality measures on Sample Text

S.No	Nodes (V_i,V_j)	Co-occurrences	Closeness measure
1	Image processing	11	1.81
2	Computer graphics	7	2.86
3	Computer vision	2	3.72
4	Image data	2	4.74
5	Signal processing	2	4.97

Because of the inverse calculation in the closeness centrality measure, lower value nodes ranked first, as such, the closeness measure value of nodes ‘image’ and ‘processing’ is lower which is 1.81 and ranked first, the second lowest value is 2.86 of nodes ‘computer’ and ‘graphics’, followed by ‘computer’ and ‘vision’ , ‘image’ and ‘data’, ’Signal’ and ‘Processing’. The nodes having closeness measure value greater than 5 are negligible.

VII. CONCLUSION

In this research paper an attempt has been made to cluster the documents using graph based keyword extraction method. In this method, initially documents are represented as graphs, then keyword are determined from the graph using centrality and closeness measures, Finally, documents are clustering using keywords. This research demonstrates the feasibility of graph based keyword extraction method as a attainable approach for document clustering.

REFERENCES

[1] Florian Boudin, *A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction*, International Joint Conference on Natural Language Processing, 2013, p. 834 – 838.

[2] Chengzhi Zhang, Huilin Wang, Yau Liu, Dan Wu, Yi Liao, Bo Wang, *Automatic Keyword Extraction from Documents using Conditional Random Fields*, Journal of Computational Information Systems, 2008, vol.4 (3), p. 1169 – 1880.

[3] George Tsatsaronis, Iraklis Varlamis, Kjetil Norvag, *SemanticRank : Ranking Keywords and Sentences using Semantic Graphs,* 23rd International Conference on Computational Linguistics-Coling 2010, p. 1074 – 1082.

[4] Marina Litvak, Mark Last, *Graph-Based Keyword Extraction for Single-Document Summarization,* Coling 2008: Proceedings of the workshop Multi-source Multilingual Information Extraction and Summarization, p. 17 – 24.

[5] Rada Mihalcea, *Graph-based Ranking Algorithms for Sentence Extraction, applied to Text Summarization*, Proceedings of ACL 2014, pp. 8-12.

[6] Slobodan Beliga, Ana Mestrovic, Sanda Martincic-Ipsic, *An Overview of Graph-Based Keyword Extraction Methods and Approaches,* Journal of Information and Organizational Science (JIOS), (2002), Vol. 39 (1), p. 1-20.

[7] François Rousseau and Michalis Vazirgiannis, *Main Core Retention on Graph-of-Words for Single-Document Keyword Extraction,* Advances in Information Retrieval (Springer) 2015, Vol. 9022, p. 382 – 393.

[8] Yukio Ohsawa, Nels E. Benson and Masahiko Yachida, *KeyGraph: automatic indexing by co-occurrence graph based on building construction metaphor,* Research and Technology Advances in Digital Libraries(IEEE), 2002 p. 12 – 18.

[9] Rada Mihalcea, *TextRank: Bringing order into Texts,* EMNLP 2014-Conference on Empirical Methods in Natural Language Processing (ACL), 2014 p. 404 – 411.