International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 12, No.1 (Jul-2022)

https://dx.doi.org/10.12785/ijcds/120130

SGAKE: Semantic Graph-based Automatic Keyword Extraction from Hindi Text Documents

Manju Lata Joshi[1], **Namita Mittal**[2] **and Nisheeth Joshi**[3]

[1]*Department of Computer Science, Banasthali University, Niwai, India*
[2]*Department of Computer Science Engineering, Malviya National Institute of Technology, Jaipur, India*

Received 6 Mar.2021, Revised 1 Apr. 2022, Accepted 25 May. 2022, Published 20 July 2022

Abstract: Automatic keyword extraction is an automated process to identify terms that best describe the subject of the document. These terms can be in the form of key terms or key phrases representing the most relevant information conveyed by the documents. Keyword extraction techniques can be Statistical based, Linguistic based, Machine Learning based, Graph-based, or Hybrid of any these. Each approach has its limitations and strengths. This paper focuses on Graph-based approaches. These approaches rely on the exploration of network properties like Degree, Structural Diversity Index, Strength, Clustering Coefficient, Neighborhood Size, Page Rank, Closeness, Betweenness, Eigenvector Centrality, Hub, and Authority Score. In the proposed approach, the graph is constructed using semantic linkages between the terms in the document. The semantic linkages between the document terms are extracted using Hindi Wordnet as a background knowledge source. Further, fourteen different graphical measures are applied to extract the keywords. The experiments are conducted on the Tourism and Health data set of the Hindi language. The results of the proposed approach are evaluated and compared with the state-of-the-art approach TextRank as well as with the Human Annotated keywords. The result shows that the closeness centrality measure produces better precision and recall as compared to other graphical measures in case of matching with human-annotated keywords while authority proved as a good graphical measure to produce keywords, matching with TextRank. The experiments prove that the proposed semantic graph-based approach performs better as compared to the state of art approach TextRank. This paper also explored the correlation between different graph-theoretic measures using different methods of correlations.

Keywords: Automatic Keyword Extraction, Semantic Graph-based Keyword Extraction, Semantic Network, Hindi Text Documents, Hindi WordNet

1. Introduction

In the era of the internet, a large number of text documents are accessible through a different set of sources as blogs, e-newspapers, e-libraries, etc. Automatic Keyword Extraction is a process to dig out representative terms from such digitized contents. Such automated systems are helping to regulate the documents without the assistance of a human.

Several keyword extraction approaches have been reported based on statistics, linguistics, machine learning, graph or hybrid approach, etc... Each approach has its strengths and limitations. This study focuses on graph-based approaches. The graph-theoretic approaches have been proved more suitable to find a significant relationship between terms as well as their weights. Still, it suffers from few limitations. One of the limitations is addressed in this study is concerned about the semantic linkages between terms of the document. For example different semantic relations such as hypernym-hyponym, meronym-holonym are not considered in keyword extraction models which leads to the extraction of the inappropriate keyword(s). To undertake this curb, several research studies investigated the use of semantic depiction of the document for different applications of text mining but this aspect of the document is still less explored.

The semantic Graph-based approach has been found appropriate for text mining processes, as these emphasize articulating distinct readings of sentences as different formulas that detain their instinctive meanings and structures. In this study, a semantic graph-based approach for keyword extraction is proposed. The objective of using a graph-based approach is to develop a new automatic keyword extraction method using graph modeling flexibility and analytical efficiency.

The prospective of semantic knowledge for automatic keyword extraction is a less explored aspect of text mining, specifically for Hindi text documents. The proposed study discovers the role of Hindi WordNet [46] ontology and represents text documents in form of a semantic graph to

E-mail address: manjulatajoshi@gmail.com, nmittal.cse@mnit.ac.in, jnisheeth@banasthali.in

extract keywords from Hindi text documents. The study explores different graphical measures for keyword extraction.

This study mainly encompasses three research objectives as To explore the role of semantics to extract keywords from text documents; To analyze the importance of different centrality measures to extract keywords and to verify that if there exists any correlation among the centrality measure through statistical analysis using Pearson's, Spearman's and Kendall's correlation methods.

The proposed work is different in the sense that it fills the existing research gap for a detailed study of WordNet as a tool for automatic keyword extraction. The other different aspect of our work is that it constructs a semantic graph-based on semantic links between document terms while most of the studies conducted earlier rely on word co-occurrence graphs. Apart from this the target corpus will be in the Hindi Language wherein very little work has been done as compared to the English language.

This research paper is organized as follows: In section 2, related work is presented for different approaches being used for keyword extraction along with their strengths and limitations. Particularly, more emphasis is given to graph-based approaches. Section 3 of the paper presents the proposed approach for keyword extraction. It includes the proposed methodology and step-by-step explanation of the algorithm to extract keywords using the proposed approach. Section 4 discusses experiments performed on two data sets. Further, it discusses results and performance evaluation of the experiments, this section is concluded showing the computation of correlation between different graphical measures and result implications. Section 5 presents the conclusion of this study along with future work.

2. Related Work

Various approaches have been proposed by many researchers to automatically generate keywords from the document(s).

These approaches include statistical-based, linguistic-based, machine learning-based, graph-based, and hybrid. Statistical-based approaches concentrate on non-linguistic features as word frequency, TF-IDF (Term Frequency-Inverse Document Frequency), word co-occurrences and PAT-tree, etc. [7]. A linguistic approach utilizes syntactic analysis [18], discourse analysis [30], and lexical analysis [8]. The resources used for lexical analysis may include WordNet [26] electronic dictionaries [12], POS taggers [9], stemmers [2] etc. Machine learning-based model [41] train data models using different models like Hidden Markov Model (HMM) [37], Support Vector Machine (SVM) [44], Naïve Bayes Classifier [15], Bagging [18] etc. The popularly known model based on machine learning is Keyphrase Extraction Algorithm (KEA) [40]. The hybrid approaches ([34], [5], [10]) coalesces any of the above two or more approaches or heuristics like length, position, features of words, Html tags nearby the words, etc. in the document. Hybrid algorithms are best designed to take the finest features of the above-discussed approaches [43].

A set of studies found using graph-based approaches to extract keywords. Such studies focus on either frequency of term to document or co-occurrences of the terms in the document ([31], [23], [24], [39], [28]). Using different Graph-based approaches [25] there are considerable studies available for automatic keyword extraction ([14], [21], [24], [27], [35], [42]).

The TextRank models proposed by [24] originated from PageRank and initiated text processing depending on graphs for sentence and keyword extraction. The scores of every node indicating its importance are derived from the importance of its neighboring nodes. The performance achieved by TextRank is favorably compared with the n-gram approach based on a supervised model. Further, the study presented in [6] compares several centrality measures applied to French and English data sets with TextRank and attain analogous results. A graph-based approach called DegExt proposed by [21] explores the sequence-based co-occurrence graph. It is a language-independent approach and it surpasses other approaches in terms of computational complexity and implementation simplicity. The approach introduced in [45] explored the significance of closeness centrality to compute keyword candidates and exhibit preferable results over the bag of words approach. The study presented in [1] also put the tweets in form of a lexical network and uses centrality measures to extract keywords. The performance of the approach shows good results in terms of computation in comparison to KEA. In the study [4] a co-occurrence network constructed and the frequency of co-occurrence of terms was considered at the first level of selection. Further, the node degree to node strength (weighted degree) ratio is measured for each node. It is observed that this ratio overpowered other standard centrality measures. The study in [27] put forwards a different technique for automatic keyword extraction based on a graph as well as the influence of word embeddings. The paper compared three diverse methods for weighting the graph of words to measure the effect of word embeddings. In the first approach weights to the nodes are assigned using word co-occurrence. The second approach is based on co-occurrence with word embeddings to measure weight and the third approach is based on Weighting with word embeddings only. Once the graph is constructed, the node ranking is done using various graph algorithms such as TextRank, HITS, Centrality Measures, Degree, etc. The experimental result shows that combining the use of word embedding with the above-mentioned measures neither increases nor decreases the performance. A recent study, presented in [38] is a comparison of nine different centrality measures (Betweenness, Clustering Coefficient, Closeness, Degree, Eccentricity, Eigenvector, K-Core, PageRank, and Structural Holes) for graph-based keyword extraction. The study constructed the graph, based on the co-occurrence of window size 3. They demonstrated that the measures produced the similar type of results. Further, they also confirmed the correlation among all graphical measures using Pearson's and Spearman's correlation coefficient. The study also proposed a new research path on multi-centrality approaches for combining graphical measures to improve the results. They conducted the experiments on three data sets as Semeval 2010 and Marujo 2012 [18]. The results exhibit that the proposed approach outperforms as compared to different centrality measures used separately.

As the above-discussed approaches ([24], [21], [14], [27], [38]) mostly are based on a co-occurrence graph that focuses on the positional and lexicographic equivalent of the terms in the document. These approaches have their strengths and limitations. As lexicographic-based similarity considers the only word-to-word matching and it does not take into consideration the meaning of the words to measure any semantic relationship between the terms. Therefore, such approaches are not effective. In such a condition, an approach is needed that considers lexical match as well as semantic match of documents words/terms. To get the semantics of document terms, some background knowledge source is required to find the semantic relationship among document terms. Ontologies [16] have been proved as an effective tool for the same. There are many ontologies available depending upon their suitability of applications, domain, and language.

A few studies ([19], [22], [36], [11]) investigate the role of the semantics of the document to extract keywords. The study discussed in [42] suggested a semantic network-based approach for keyword extraction and to comment on the nature of documents for text documents. For keyword extraction, they performed experiments on the documents related to specific queries. The approach considers a set of relevant documents of these queries. The algorithm identifies keywords of each relevant document by applying graph-theoretic measures such as Eccentricity, Degree; Centrality, etc. the nodes with a maximum degree of these measures are considered representative words of the document. If these extracted representative words are presented in a query or are semantically associated with the query, then they are considered as the relevant result of the experiment. They performed experiments on 50 queries and obtained Inspiring outcomes. The experiment depicts that the eccentricity and closeness centrality of a node provides a fine base for keyword extraction. Another approach based on unsupervised learning mentioned in [35] makes use of Wordnet-based semantic relations to construct the graph. The approach extracts nouns and applies Depth-First search (DFS) to find related Hypernym-Hyponyms [33] and pertain betweenness centrality measures and closeness centrality measures.

For this study, we are exploring the semantic graph-based approach more exhaustively for keyword extraction. Hindi WordNet [46] ontology is used as a background knowledge source to find semantic relations between document terms. It's a kind of thesaurus that is organized in the form of concepts. Our approach constructs a semantic graph of document terms. The nodes of the semantic graphs are representing noun terms of the document and edges symbolizes semantic relation (extracted from Hindi Wordnet) between the nodes (document terms). Various type of Semantic relation is defined in Hindi Wordnet [29]. For this paper, hypernym-hyponym and meronym-holonym relationships between the document terms are used to construct the graph. In our approach, the reason behind considering noun terms only is that it is observed that most of the meaning of a document is conveyed by the content terms only and as the objective of the study is to extract keywords, there are least chances that other POS like Verb, Adjectives or Adverb can be keywords for the document.

Any natural-language text can be symbolized as a semantic network where nodes can be used to represent a concept while edges can be used to represent some kind of semantic relationship among concepts in the same. Once a document(s) is represented in form of a semantic network or semantic graph, a set of measures from the graph and network analysis can be applied to perform quantitative analysis. Further, its outcome can be utilized in several potential applications of text mining such as to detect closely related concepts and discover the most dominant concepts. Such a concept generates some meaning of the text and provides a method to understand the text's structure in a better way [32].

3. Proposed Approach

The proposed approach is applied to construct a semantic graph for a document or data set irrespective of any language or domain. The only requirement is that the linguistic ontology for that language should be available. The following figure shows a Process flow for the proposed methodology.

In step 1 this approach accepts a text document D as input and then in step 2, document D is preprocessed by removing stop words, and the obtained preprocessed document is called D'. In step 3 after preprocessing all the remaining terms of the document are tagged with their respective part of speech which leads to the extraction of unique noun terms (N1, N2......Nm) creating a master list of nouns of the document. In step 4 semantic graph G (V, E) is constructed where V = v1, v2, v3......, vm is set of vertices in the graph, each vertex represents a unique noun term Ni of the document and E= e1, e2, e3, ..., en represents a set of edges symbolizing semantic relationship between vi and vj. To find semantic linkages between vi and vj, semantic relationships provided by Hindi Wordnet are used. There are many different semantic relationships defined in Hindi WordNet as Antonymy, Hypernymy-Hyponymy, Meronymy-Holonymy, Entailment, Troponymy, and Crosslink relationships as Ability link, Capability link, and Function link [33]. For this paper, two types of semantic relationships namely Hypernym-Hyponym and Meronym-Holonym are used. The reason for

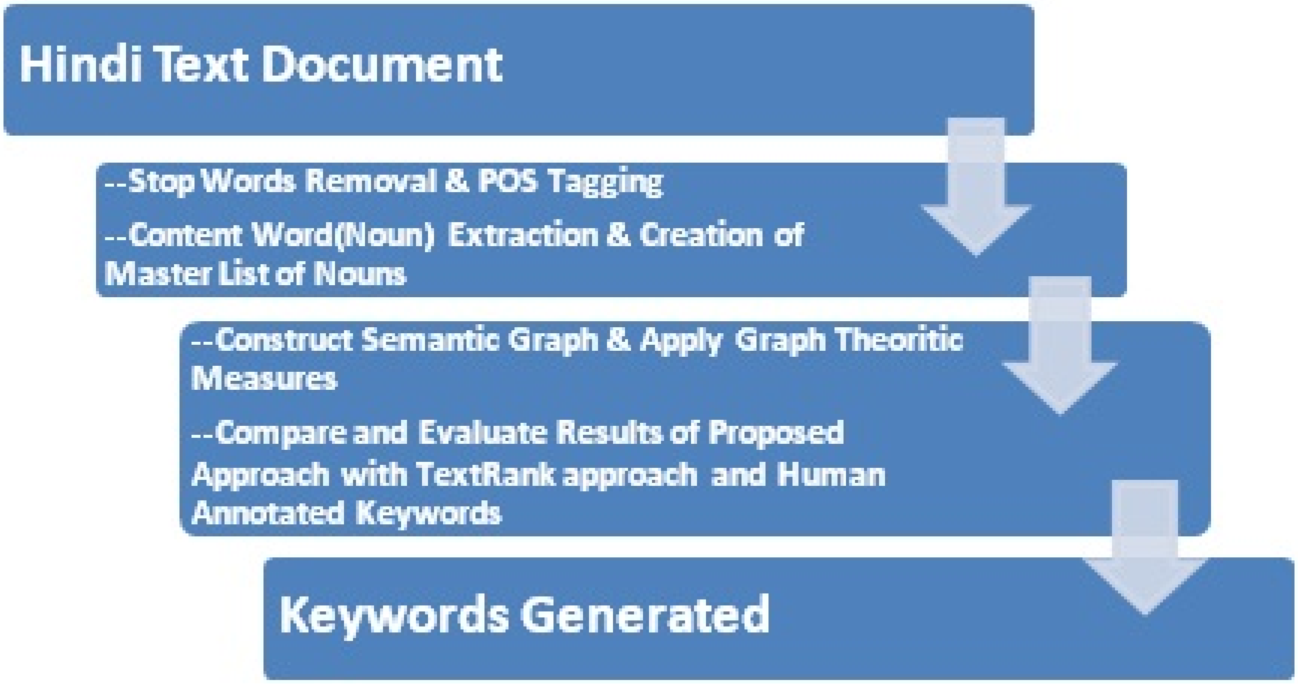
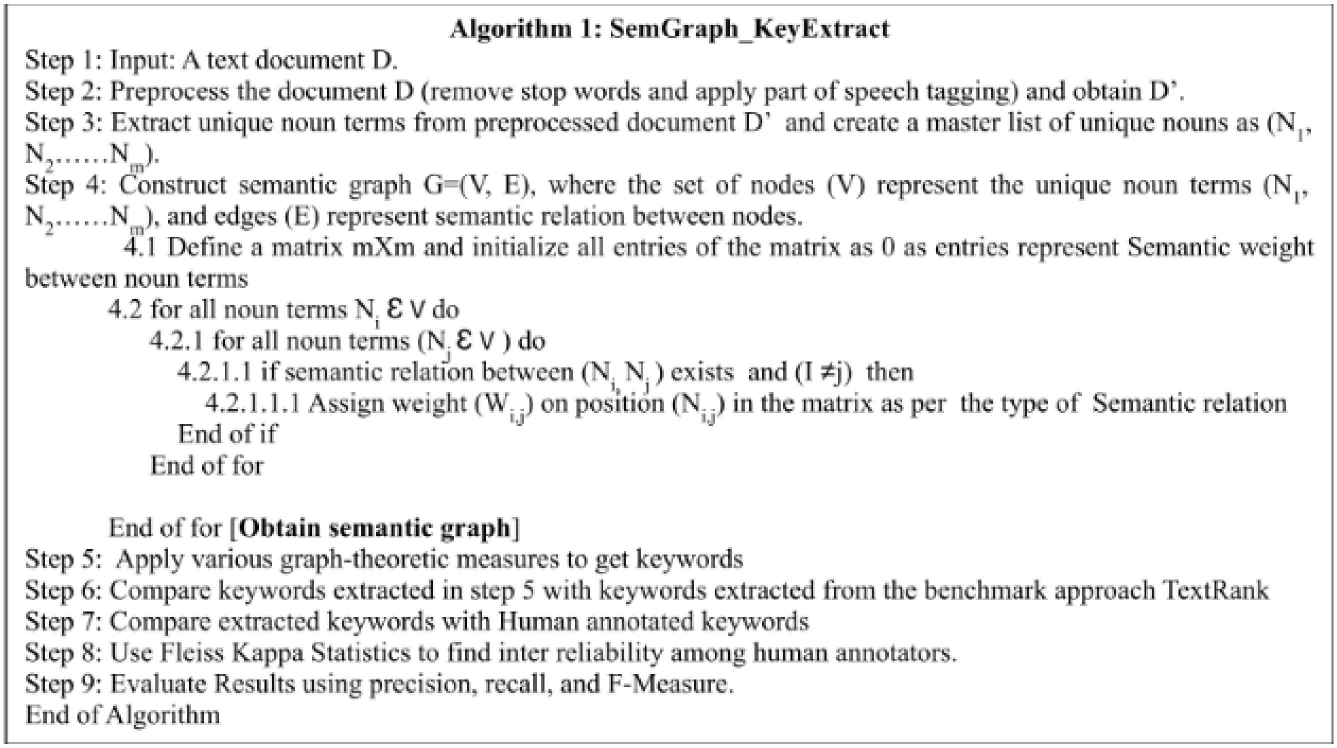

Hindi Text Document
--Stop Words Removal & POS Tagging
--Content Word(Noun) Extraction & Creation of Master List of Nouns
--Construct Semantic Graph & Apply Graph Theoritic Measures
--Compare and Evaluate Results of Proposed Approach with TextRank approach and Human Annotated Keywords
Keywords Generated

Figure 1. Process flow of proposed methodologyt



```
Algorithm 1: SemGraph_KeyExtract
Step 1: Input: A text document D.
Step 2: Preprocess the document D (remove stop words and apply part of speech tagging) and obtain D'.
Step 3: Extract unique noun terms from preprocessed document D' and create a master list of unique nouns as ($N_1$, $N_2$……$N_m$).
Step 4: Construct semantic graph G=(V, E), where the set of nodes (V) represent the unique noun terms ($N_1$, $N_2$……$N_m$), and edges (E) represent semantic relation between nodes.
        4.1 Define a matrix mXm and initialize all entries of the matrix as 0 as entries represent Semantic weight between noun terms
        4.2 for all noun terms $N_i$ Ɛ V do
            4.2.1 for all noun terms ($N_j$ Ɛ V ) do
                4.2.1.1 if semantic relation between ($N_{i,}$ $N_j$ ) exists and (I ≠j) then
                    4.2.1.1.1 Assign weight ($W_{i,j}$) on position ($N_{i,j}$) in the matrix as per the type of Semantic relation
                End of if
            End of for

        End of for [Obtain semantic graph]
Step 5: Apply various graph-theoretic measures to get keywords
Step 6: Compare keywords extracted in step 5 with keywords extracted from the benchmark approach TextRank
Step 7: Compare extracted keywords with Human annotated keywords
Step 8: Use Fleiss Kappa Statistics to find inter reliability among human annotators.
Step 9: Evaluate Results using precision, recall, and F-Measure.
End of Algorithm
```

Figure 2. Semgraph Key Extract

using only these two relationships is that as Antonymy is a lexical relation indicating ’opposites’, Gradation represents intermediate states between two antonyms, Entailment and Troponymy shows the relationship between verbs. Though Antonymy’s relation can be between any two nouns too it is observed that considering Antonymy for keyword extraction doesn’t seem much suitable as it conveys the opposite meaning. On the other side Cross-links establishes relationships between different parts of speeches like ’nouns’ and ’verbs’. Hence, as for this study, we are considering nouns, only hypernymy-hyponymy and matched with keywords produced by the proposed approach. In step 8 well known Fleiss’ Kappa statistics [14] is used to compute inter reliability among three human annotator In final step 10, the performance of the proposed approach is evaluated using precision, recall, and F-Measure.

4. Experiments And Result Discussion

This section of the paper is organized into three parts, wherein the first part is about the data set and tools used to accomplish this study. The second part depicts the construction of a semantic graph for a document and extraction of keywords generated using the proposed approach. These keywords are further matched with keywords generated using TextRank, and then with human-annotated keywords. Further, results are evaluated and discussed using evaluation measures precision, recall, and F-Measure. In the next part of this section correlation analysis among eight different graph theoretical measures is carried out.

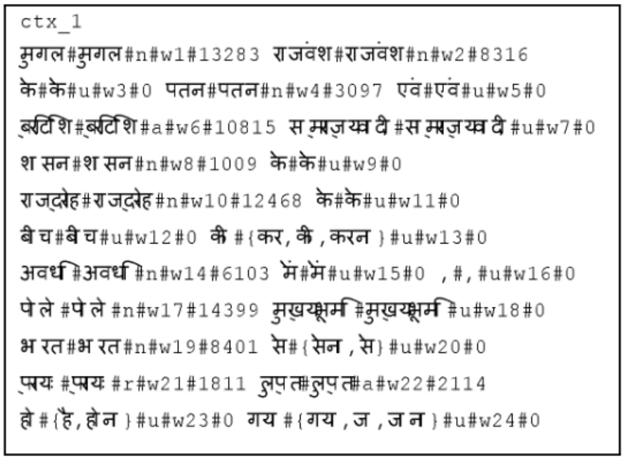
A. Data Set and Tools Used

The experiments are conducted on two different data sets Tourism data set and Health data set of Hindi language provided by the Centre for Indian Language Technology (CFILT), IIT Bombay [20]. Each data set consists of 152 documents. Each document of each data set is consisting 100 sentences; each word of each sentence is tagged with their respective part of speech, synset id as in Hindi Wordnet [46], and word id in that document. Though the data sets used are small but suitable for the proposed study. As Hindi is still a resource-poor language, still there is no benchmark data set available for keyword extraction; hence this data set is taken for this study. The experiments are implemented using Java as a high-level programming language on the meronymy-holonymy relationships are appropriate. Each relationship is assigned a weight to quantify the relationship between terms. Here weight Wi,j = 1 is assigned if (Ni, Nj) are related with hypernym-hyponym relationship and Wi,j = 2 is assigned if (Ni, Nj) are related with meronym-holonym relationship. By applying these relationships between terms (N1, N2. . . . . . Nm) the semantic graph is constructed. As the graph is constructed, in step 5 few graph-theoretic measures are applied to compute the importance of each node vi in the graph and sorted the terms in descending order of their semantic scores calculated by applying graph-theoretic measures. The top n terms are considered as candidate keywords. In step 6 the candidate keywords selected in step 5 are compared with keywords extracted by using the state-of-the-art approach TextRank. In step 7, the human-annotated keywords are collected by asking keywords from three different human annotators for each document. These human-annotated keywords are Linux platform. The Hindi Wordnet APIs are used to extract the semantic relationship between document terms. The Hindi WordNet 1.5 is a system for bringing together different lexical and semantic relations between the Hindi words. It organizes the lexical information in terms of word meanings and can be termed as a lexicon based on psycholinguistic principles

The design of the Hindi WordNet is inspired by the famous English WordNet. To construct and visualize the semantic graph the graph visualization tool GEPHI 0.8.2 is used [3]. Due to the unavailability of the gold standard set of keywords for tourism and health data sets, human-annotated keywords are taken as gold standard keywords.

B. Experimental Results and Analysis

This section is explaining step by step results as per the proposed approach on a single document. Here document 0009 hin tourism.txt is taken to illustrate our approach from the Tourism data set. As every document D comprises 100 sentences, one sample sentence is presented as a reference in Fig. 3, to show that how the input data is organized. The following sentence is from the Tourism domain.In the above Fig. 3, the ctx-1 is sentence id, each word is represented with some information with it, the word itself, it’s part of speech (POS), word id in the document, and its synset-id as per in Hindi WordNet. As per steps 2 and 3 of the proposed algorithm, stop words are removed to obtain D’ and only noun terms are extracted.



```
ctx_1
मुगल#मुगल#n#w1#13283 राजवंश#राजवंश#n#w2#8316
के#के#u#w3#0 पतन#पतन#n#w4#3097 एवं#एवं#u#w5#0
ब्रिटिश#ब्रिटिश#a#w6#10815 साम्राज्यवादी#साम्राज्यवादी#u#w7#0
श सन#श सन#n#w8#1009 के#के#u#w9#0
राजद्रोह#राजद्रोह#n#w10#12468 के#के#u#w11#0
बी च#बी च#u#w12#0 की #{कर, की ,करन }#u#w13#0
अवधि#अवधि#n#w14#6103 में#में#u#w15#0 ,#,#u#w16#0
पे ले #पे ले #n#w17#14399 मुख्यभूमि#मुख्यभूमि#u#w18#0
भ रत#भ रत#n#w19#8401 से#{सेन ,से}#u#w20#0
प्रायः #प्रायः #r#w21#1811 लुप्त#लुप्त#a#w22#2114
हो #{है,हो न }#u#w23#0 गय #{गय ,ज ,ज न }#u#w24#0
```

Figure 3.Excerpt from Document 009 hin tourism.txt

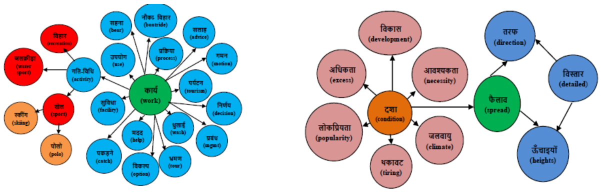
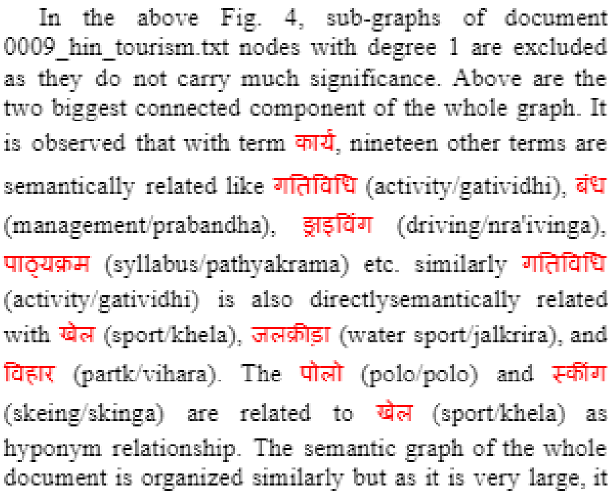


Figure 4. Giant components of the semantic graph constructed for document 0009 hin tourism.txt using Gephi



In the above Fig. 4, sub-graphs of document 0009_hin_tourism.txt nodes with degree 1 are excluded as they do not carry much significance. Above are the two biggest connected component of the whole graph. It is observed that with term कार्य, nineteen other terms are semantically related like गतिविधि (activity/gatividhi), बंध (management/prabandha), ड्राइविंग (driving/nra'ivinga), पाठ्यक्रम (syllabus/pathyakrama) etc. similarly गतिविधि (activity/gatividhi) is also directlysemantically related with खेल (sport/khela), जलक्रीड़ा (water sport/jalkrira), and विहार (partk/vihara). The पोलो (polo/polo) and स्कींग (skeing/skinga) are related to खेल (sport/khela) as hyponym relationship. The semantic graph of the whole document is organized similarly but as it is very large, it couldn't be included here as part of this research paper. Among 103 semantic linkages 100 Hypernym-hyponym relationships found and rest 03 are Meronym-holonyms relations.

As this graph is constructed, in step 5 a range of graph theoretical measures applied and among resulting keyword (with the highest semantic scores) top 25 keywords are extracted produced by Out-Degree, Degree, Weighted Degree, Weighted in-Degree, Weighted Out Degree, Authority, Hub, PageRank, Clustering Coefficient and Eigenvector Centrality [25]. Table A depicts the resulting keywords extracted from the proposed approach. Here, only those measures' output has shown which is generating the best set of keywords for document 0009-hin-tourism.txt.

These keywords extracted using the proposed approach are compared with the keywords extracted from the state of the art keyword extraction approach TextRank and also compared with the Human annotated Keywords. For this paper, human-annotated keywords are considered as gold standard keywords. Three human annotator's suggested their own individual set of keywords for this study.

Table B exhibits the keywords extracted by applying the proposed approach and ranked using four different graph-theoretic measures. It is observed that graphical measure Authority is producing the best results matching the keywords generated from the baseline algorithm TextRank [24]. Authority is the measure to compute the importance of a node in the network. A vertex is considered an authority if it has many nodes linking to it. Authority is mathematically represented as:

$$x_i = \propto \sum_j A_{ij}\, y_j \qquad (1)$$

Here yj is the hub centrality score. The TextRank approach is an unsupervised approach that is based on the co-occurrence of the terms within a window of N words. If two lexical units co-occur within a window of N terms then an edge is drawn between these two. The score of a vertex Vi is defined as follows:

$$PR^{w}\left(V_i\right) = (1 - d) + d * \sum_{V_j \in In(V_i)} W_{ji} \frac{PR^{w}(V_j)}{\sum_{V_j \in Out(V_j)} W_{kj}}$$

(2)

In Table B, the keywords highlighted in bold are those which are commonly extracted by both approaches viz. TextRank as well as the proposed approach.

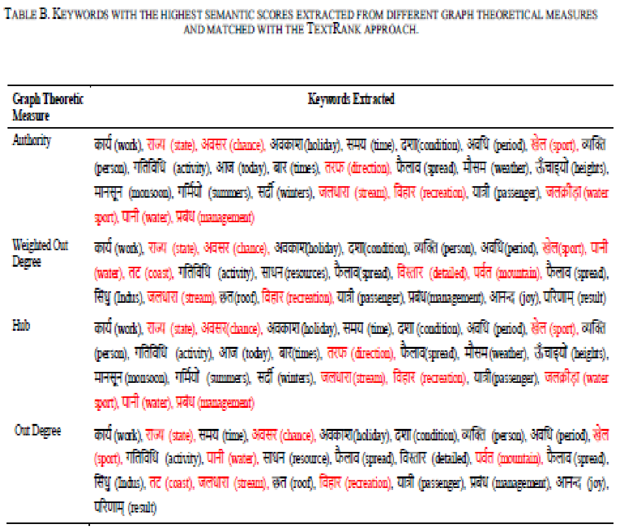


Table B. Keywords with the highest semantic scores extracted from different graph theoretical measures and matched with the TextRank approach.

Graph Theoretic Measure	Keywords Extracted
Authority	कार्य (work), राज्य (state), अवसर (chance), अवकाश(holiday), समय (time), दशा(condition), अवधि (period), खेल (sport), व्यक्ति (person), गतिविधि (activity), आज (today), बार (times), तरफ (direction), फैलाव (spread), मौसम (weather), ऊँचाइयों (heights), मानसून (monsoon), गर्मियों (summers), सर्दी (winters), जलधारा (stream), विहार (recreation), यात्री (passenger), जलक्रीड़ा (water sport), पानी (water), प्रबंध (management)
Weighted Out Degree	कार्य (work), राज्य (state), अवसर (chance), अवकाश(holiday), दशा(condition), व्यक्ति (person), अवधि(period), खेल(sport), पानी (water), तट (coast), गतिविधि (activity), साधन(resources), फैलाव(spread), विस्तार (detailed), पर्वत (mountain), फैलाव (spread), सिंधु (Indus), जलधारा (stream), छत(roof), विहार (recreation), यात्री (passenger), प्रबंध(management), आनन्द (joy), परिणाम (result)
Hub	कार्य (work), राज्य (state), अवसर(chance), अवकाश(holiday), समय (time), दशा (condition), अवधि (period), खेल (sport), व्यक्ति (person), गतिविधि (activity), आज (today), बार(times), तरफ (direction), फैलाव(spread), मौसम(weather), ऊँचाइयों (heights), मानसून (monsoon), गर्मियों (summers), सर्दी (winters), जलधारा(stream), विहार (recreation), यात्री(passenger), जलक्रीड़ा (water sport), पानी (water), प्रबंध (management)
Out Degree	कार्य (work), राज्य (state), समय (time), अवसर (chance), अवकाश(holiday), दशा (condition), व्यक्ति (person), अवधि (period), खेल (sport), गतिविधि (activity), पानी (water), साधन (resource), फैलाव (spread), विस्तार (detailed), पर्वत (mountain), फैलाव (spread), सिंधु (Indus), तट (coast), जलधारा (stream), छत (roof), विहार (recreation), यात्री (passenger), प्रबंध (management), आनन्द (joy), परिणाम (result)

TABLE A. KEYWORDS WITH THE HIGHEST SEMANTIC SCORES EXTRACTED FROM DIFFERENT GRAPH THEORETICAL MEASURES

S. No.	Noun	English meaning	Authority	Noun	English meaning	Degree	Noun	English meaning	Out Degree	Noun	English meaning	Closeness Centrality
1.	कार्य	work	0.06	कार्य	work	19	कार्य	work	19	स्कींग	Skiing	3.6
2.	राज्य	state	0.04	राज्य	state	13	राज्य	state	13	पोलो	Polo	3.6
3.	अवसर	chance	0.03	अवसर	chance	12	समय	time	9	जलक्रीड़ा	Water sport	3.52
4.	अवकाश	holiday	0.03	अवकाश	holiday	12	अवसर	chance	9	पाठ्यक्रम	syllabus	3.2
5.	समय	time	0.03	समय	time	10	अवकाश	holiday	9	विस्तार	detailed	3.08
6.	दशा	condition	0.03	दशा	condition	8	दशा	condition	8	फैलाव	Spread	3.08
7.	अवधि	period	0.02	अवधि	period	7	व्यक्ति	Person	4	विहार	recreation	2.8
8.	खेल	sport	0.01	खेल	sport	4	अवधि	period	4	खेल	Sport	2.64
9.	व्यक्ति	person	0.01	व्यक्ति	person	4	खेल	sport	3	तट	Coast	2.5
10.	गतिविधि	activity	0.01	गतिविधि	activity	3	गतिविधि	activity	2	जलप्रपातों	Water fall	2.5
11.	आज	today	0.01	आज	today	3	पानी	water	2	bIca		2.41
12.	बार	times	0.01	बार	times	3	साधन	resources	2	वर्ष	Year	2.41
13.	तरफ	direction	0.01	तरफ	direction	3	फैलाव	spread	2	आवश्यकता	necessity	2.41
14.	फैलाव	spread	0.01	फैलाव	spread	3	विस्तार	detailed	2	मौसम	weather	2.41
15.	मौसम	weather	0.01	मौसम	weather	3	पर्वत	mountain	2	मौसम	weather	2.41
16.	ऊँचाइयों	Heights	0.01	ऊँचाइयों	Heights	3	फैलाव	spread	2	विकास	development	2.41
17.	मानसून	monsoon	0.01	मानसून	monsoon	3	सिंधु	Indus	1	लोकप्रियता	popularity	2.41
18.	गर्मियों	summers	0.01	गर्मियों	summers	3	तट	coast	1	जलवायु	climate	2.41
19.	सर्दी	winter	0.01	सर्दी	winter	3	जलधारा	stream	1	थकावट	Tiring	2.41
20.	जलधारा	stream	0.01	जलधारा	stream	2	छत	roof	1	अधिकता	Excess	2.41
21.	विहार	recreation	0.01	विहार	recreation	2	विहार	recreation	1	पर्यटक	tourist	2.4
22.	यात्री	Passenger	0.01	यात्री	The traveler	2	यात्री	The traveler	1	तरफ	direction	2.33
23.	जलक्रीड़ा	Water sport	0.01	जलक्रीड़ा	Water sport	2	प्रबंध	management	1	ऊँचाइयों	Heights	2.33
24.	पानी	water	0.01	पानी	water	2	आनन्द	joy	1	गमन	motion	2.32
25.	प्रबंध	management	0.01	प्रबंध	management	2	coTI		1	ड्राइविंग	driving	2.32

In the above table, words highlighted in red color denote the common terms extracted as keywords from both systems generated as well as the Text rank approach. Similarly, the keywords generated from the proposed approach are compared with human-annotated keywords and it is observed that Closeness Centrality has come out as the best measure to generate prominent keywords. Closeness centrality is a measure to indicate how close a node is to all other nodes in the network. It is calculated as the avg. of the shortest path length from the node to every other node in the network. Mathematically, the closeness centrality of every vertex is formulated 3 as:

$$c_c(v) = \frac{N-1}{\sum v \neq u \; d_{vu}} \tag{3}$$

In the above equation (3), dvu is the shortest path distance between vertex v and u. Similarly, another measure Eccentricity also produced a good set of keywords having the second-highest precision, recall, and F-Measure when compared with Human Annotated Keywords. Eccentricity is defined as the maximum distance between a vertex to all other vertices in the graph. It is computed as:

$$E(v) = \max_{u \in v} d(u, v) \tag{4}$$

Table C shows those measures which produce the best keywords and has the highest precision and recall when compared with human-annotated keywords for document 0009-hin-tourism.txt. To measure the quality of generated summary we used well-accepted evaluation measures i.e. Precision, Recall, and F-measure which are depicted as following equations

In the above table, words highlighted in red color denote the common terms extracted as keywords from both systems generated as well as human-annotated. The performance evaluation of the proposed algorithm is done by applying precision, recall, and F-Measure. The following Table 1 exhibits the performance results for the Tourism data set while Table 2 represents performance results for the Health data set.

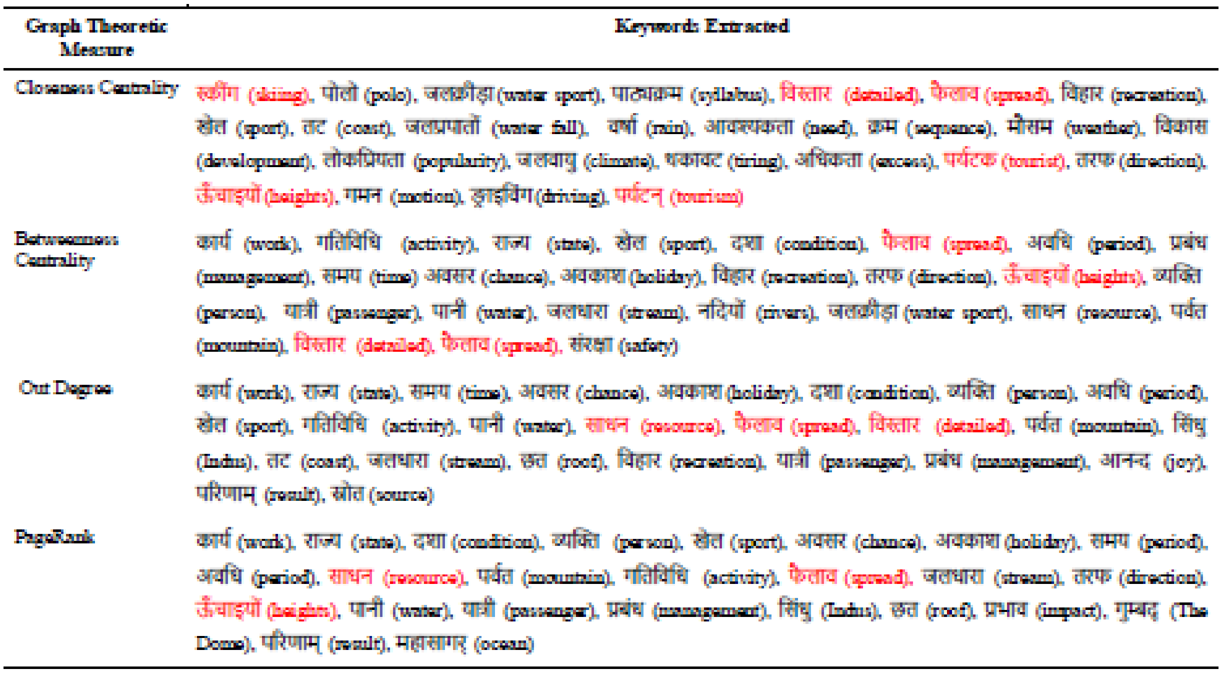
As the evaluation results are shown above, system-

$$precision = \frac{human\ annotated\ summary\ sentences \cap system-generated\ summary\ sentences}{total\ no.\ of\ summary\ sentences\ generated\ by\ syste0m} \quad (5)$$

$$recall = \frac{human\ annotated\ summary\ sentences \cap system-generated\ summary\ sentences}{total\ no.\ of\ sentences\ generated\ in\ human-annotated\ summary} \quad (6)$$

$$F - measure = 2\frac{precision{*}recall}{precision+recall} \quad (7)$$

TABLE C. KEYWORDS WITH THE HIGHEST SEMANTIC SCORES EXTRACTED FROM DIFFERENT GRAPH THEORETICAL MEASURES AND MATCHED WITH HUMAN ANNOTATED KEYWORDS.



Graph Theoretic Measure	Keywords Extracted
Closeness Centrality	स्कींग (skiing), पोलो (polo), जलक्रीड़ा (water sport), पाठ्यक्रम (syllabus), विस्तार (detailed), फैलाव (spread), विहार (recreation), खेल (sport), तट (coast), जलप्रपातों (water fall), वर्षा (rain), आवश्यकता (need), क्रम (sequence), मौसम (weather), विकास (development), लोकप्रियता (popularity), जलवायु (climate), थकावट (tiring), अधिकता (excess), पर्यटक (tourist), तरफ (direction), ऊँचाइयों (heights), गमन (motion), ड्राइविंग (driving), पर्यटन् (tourism)
Betweenness Centrality	कार्य (work), गतिविधि (activity), राज्य (state), खेल (sport), दशा (condition), फैलाव (spread), अवधि (period), प्रबंध (management), समय (time) अवसर (chance), अवकाश (holiday), विहार (recreation), तरफ (direction), ऊँचाइयों (heights), व्यक्ति (person), यात्री (passenger), पानी (water), जलधारा (stream), नदियों (rivers), जलक्रीड़ा (water sport), साधन (resource), पर्वत (mountain), विस्तार (detailed), फैलाव (spread), संरक्षा (safety)
Out Degree	कार्य (work), राज्य (state), समय (time), अवसर (chance), अवकाश (holiday), दशा (condition), व्यक्ति (person), अवधि (period), खेल (sport), गतिविधि (activity), पानी (water), साधन (resource), फैलाव (spread), विस्तार (detailed), पर्वत (mountain), सिंधु (Indus), तट (coast), जलधारा (stream), छत (roof), विहार (recreation), यात्री (passenger), प्रबंध (management), आनन्द (joy), परिणाम् (result), स्रोत (source)
PageRank	कार्य (work), राज्य (state), दशा (condition), व्यक्ति (person), खेल (sport), अवसर (chance), अवकाश (holiday), समय (period), अवधि (period), साधन (resource), पर्वत (mountain), गतिविधि (activity), फैलाव (spread), जलधारा (stream), तरफ (direction), ऊँचाइयों (heights), पानी (water), यात्री (passenger), प्रबंध (management), सिंधु (Indus), छत (roof), प्रभाव (impact), गुम्बद् (The Dome), परिणाम् (result), महासागर् (ocean)

TABLE I. RESULT ANALYSIS OF SEMANTIC GRAPH-BASED AUTOMATIC KEYWORD EXTRACTION FOR TOURISM DATA SET

S. no.	Algorithms Compared	Avg. Precision	Avg. Recall	Avg. F- Measure
1	Human Annotated vs. System Generated	0.4742	0.46274	0.4636
2	Human Annotated vs. TextRank	0.22751	0.23247	0.22924
3	TextRank vs System Generated	0.23142	0.21428	0.21703

TABLE II. Result Analysis of Semantic Graph-based Automatic Keyword Extraction for Health Data Set

S. no.	Algorithms Compared	Avg. Precision	Avg. Recall#	Avg. F- Measure
1	Human Annotated vs. System Generated	0.404444	0.405409	0.404603
2	Human Annotated vs. TextRank	0.259259	0.266431	0.262554
3	TextRank vs System Generated	0.191111	0.181069	0.188034

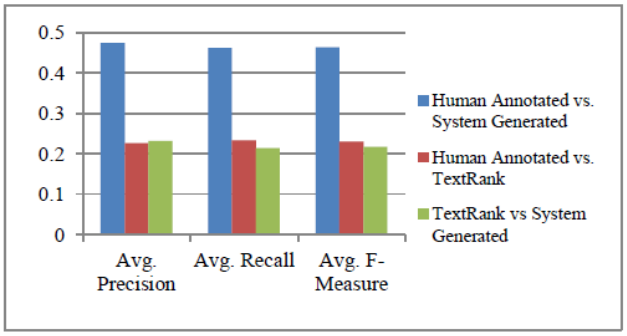


Figure 5. Comparison between different Algorithms for Tourism Data Set

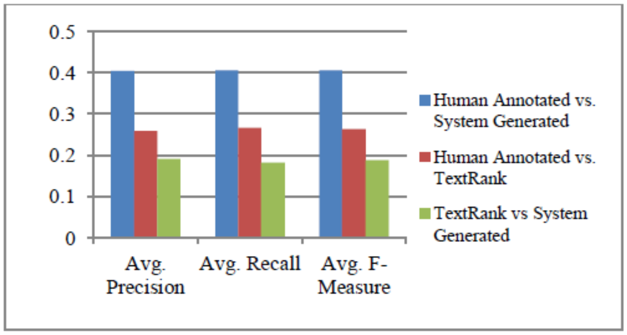


Figure 6. Comparison between different Algorithms for Health Data Set

generated keywords are the keyword generated through the proposed approach. The set of keywords generated by the proposed approach are better matched with the gold standard keywords as compared to, match of keywords generated through the TextRank approach with gold standard keywords. The following graph shows the results in graphical form.

C. Time Complexity

In this section, time complexity of proposed approach is discussed. The major part of proposed approach is construction of semantic graph using Hindi WordNet relations. As the semantic relations are extracted for each noun term and total number of noun terms are considered as n then searching for semantic relation for all noun terms to all other noun terms takes nXn times i.e. n2 times. So the time complexity to construct semantic graph is o(n2). Other statements in algorithm for comparisons, assignments and computing values for all graphical measures takes constant time, hence ignored to compute time complexity.

D. Result Implications

In the Tourism data set, each document of tourism data set explains some tourist places and the famous monuments or things about that place. These things or monuments are an integral part of that tourist place but as Hindi WordNet is a general/open domain ontology focusing more on the semantic relationship, in most cases it does not consider such a relationship. For example, there is no relationship found between राजस्थान (Rajasthan/rajasthana) and ऊँट (camel/umta)and in आगरा (Agra/agara) and ताजमहल (Tajmahal/tajmahala). Similarly, the country वेनेजुएला (Venezuela/ veneju'ela) found in Hindi wordnet but एंजल फॉल्स (Angel falls/ enjala pholsa) do not found in Hindi wordnet and so no relationship between वेनेजुएला (Venezuela/ veneju'ela) and एंजल फॉल्स (Angel falls/ enjala pholsa) while these both terms are sharing significant relationship as far as tourism is a concern. Similar instances are with the Health domain as well. Here, the need for domain ontology can be felt. If there would have tourism ontology available and the semantic relationships are extracted from that ontology then there are maximum chances that the above-mentioned terms like वेनेजुएला (Venezuela/ veneju'ela) and एंजल फॉल्स (Angel falls/ enjala pholsa) and other such terms shares some semantic relation. As Hindi WordNet is lacking such a relationship, to some extent it also affects the semantic graph constructed for an experiment in the proposed study. The semantic graph doesn't show any link between these, so no graphical measure assigns a good value to such terms, and hence, these terms do not take place in the list of the top keywords. While a human annotator even not from the same domain assigns a higher priority to these relations. Consecutively, it influences the precision, recall, and F-Measure values.

It is also observed, that the human-annotated keywords are subjective, due to different preferences of annotators in the selection of keywords. Depending on each annotator's preference, the selection may lead to a set of keywords varying from each other. It may decline in precision and recall for human annotators vs. system-generated keywords. The study in [13] focuses on the benchmarks and suggested three different types of annotators; author of the document, reader, and professional indexer to extract keywords. As there are no defined guidelines for annotating keywords, it is difficult to generate a reliable gold standard of keywords. In this paper, Fleiss' Kappa statistics are also used to compute inter-reliability among three human annotators which outcome as fair inter-reliability amongst three human annotators Authors and Affiliations

E. Correlation Analysis among different Graphical Measures

In this sub-section, the study finds the correlation between different graph theoretical measures used for experiments. Since different graphical measures produce a different set of keywords we are not analyzing keywords word to word but using correlation here shows that how the keywords generated through these measures are correlated. As in statistics, correlation coefficients are used to measure the relationship between two given variables. The correlation coefficient produces values ranging from -1 to +1 wherein -1 represents a negative relationship, 0 represents no relationship and +1 represents the positive relationship between variables. In this study, three different methods are used to compute correlation coefficients i.e. Pearson's correlation, Spearman's correlation, and Kendall's correlation. The major difference among these methods is that Pearson's correlation coefficient method measures the linear correlation between two variables while Spearman's and Kendall's correlation coefficient methods are rank-based [17].

Pearson's correlation is also known as Pearson Product Moment Correlation (PPMC). This method denotes the linear relationship between two sets of data. It can be mathematically represented as follows:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2] \times [n\sum y^2 - (\sum y)^2]]}} \qquad (8)$$

Where x and y are two classes of observation and n is cardinality. Tables 3 and 4 show the results of Correlation between eight different graphical measures using Pearson's Correlation coefficient method for Tourism and Health data sets respectively.

Spearman's correlation coefficient determines the strength and direction of the monotonic relationship between two variables unlike the strength and direction of the linear relationship between two variables determined by Pearson's Correlation Coefficient. Mathematically it can be represented as:

$$r = \frac{\sum_i (x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \, \sum_i (y_i - \bar{y})^2}} \qquad (9)$$

Where xi and yi are different observations of two variables x and y while x and y denotes the mean of x and y variables.

Tables 5 and 6 shows correlation coefficient results using Spearman's correlation method for Tourism and Health data sets respectively.

TABLE III. PEARSON'S CORRELATION COEFFICIENT MATRIX FOR DIFFERENT GRAPHICAL MEASURES FOR THE TOURISM DATA SET

	Closeness Centrality	In-Degree	Betweenness Centrality	Degree#	Out-Degree	Authority	Hub	PageRank
Closeness Centrality	1	0.743958709	0.874572787	0.908057989	0.906753968	0.90247858	0.90247858	0.864293884
In-Degree	0.743958709	1	0.725562307	0.719862983	0.713238148	0.709095088	0.709095088	0.660493418
Betweenness Centrality	0.874572787	0.725562307	1	0.949229563 0.943813908	0.946936927	0.946936927	0.93318685	
Degree	0.908057989	0.719862983	0.949229563	1	0.987385742	0.996392069	0.996392069	0.955432294
Out-Degree	0.906753968	0.713238148	0.943813908	0.987385742	1	0.989798596	0.989798596	0.954589998
Authority	0.90247858	0.709095088	0.946936927	0.996392069	0.989798596	1	1	0.959047028
Hub	0.90247858	0.709095088	0.946936927	0.996392069	0.989798596	1	1	0.959047028
PageRank	0.864293884	0.660493418	0.93318685	0.955432294	0.954589998	0.959047028	0.959047028	1

TABLE IV. PEARSON'S CORRELATION COEFFICIENT MATRIX FOR DIFFERENT GRAPHICAL MEASURES FOR HEALTH DATA SET

	Closeness Centrality	In-Degree	Betweenness Centrality	Degree#	Out-Degree	Authority	Hub	PageRank
Closeness Centrality	1	0.846362	0.863758317	0.76734	0.84337	0.784257	0.784257	0.77225112
In-Degree	0.846362	1	0.945616262	0.902645	0.970988	0.9103246	0.910325	0.89737434
Betweenness Centrality	0.8637583	0.945616	1	0.868074	0.941313	0.8760474	0.876047	0.89273689
Degree	0.7673395	0.902645	0.868073767	1	0.918721	0.9906427	0.990643	0.9717754
Out-Degree	0.8433701	0.970988	0.941312605	0.918721	1	0.9149031	0.914903	0.89049391
Authority	0.784257	0.910325	0.876047418	0.990643	0.914903	1	1	0.9715039
Hub	0.784257	0.910325	0.876047418	0.990643	0.914903	1	1	0.9715039
PageRank	0.7722511	0.897374	0.892736887	0.971775	0.890494	0.9715039	0.971504	1

TABLE V. SPEARMAN'S CORRELATION COEFFICIENT MATRIX FOR DIFFERENT GRAPHICAL MEASURES FOR THE TOURISM DATA SET

	Closeness Centrality	In-Degree	Betweenness Centrality	Degree#	Out-Degree	Authority	Hub	PageRank
Closeness Centrality	1	0.7605282	0.961213445	0.94368	0.941025	0.939574	0.93957402	0.96413
In-Degree	0.7605282	1	0.761842979	0.74856	0.754439	0.73528	0.73527994	0.747215
Betweenness Centrality	0.9612134	0.761843	1	0.96008	0.963302	0.954522	0.95452217	0.98905
Degree	0.9436769	0.7485589	0.960084695	1	0.946545	0.986462	0.98646235	0.963072
Out-Degree	0.9410248	0.7544393	0.963302206	0.94654	1	0.944702	0.94470229	0.966529
Authority	0.939574	0.7352799	0.954522166	0.98646	0.944702	1	1	0.963581
Hub	0.939574	0.7352799	0.954522166	0.98646	0.944702	1	1	0.963581
PageRank	0.9641301	0.7472152	0.989050084	0.96307	0.966529	0.963581	0.9635814	1

TABLE VI. SPEARMAN'S CORRELATION COEFFICIENT MATRIX FOR DIFFERENT GRAPHICAL MEASURES FOR HEALTH DATA SET

	Closeness Centrality	In-Degree	Betweenness Centrality	Degree#	Out-Degree	Authority	Hub	PageRank
Closeness Centrality	1	0.909266486	0.955776186	0.875444804	0.939190063	0.87013499	0.87013499	0.884639039
In-Degree	0.909266486	1	0.956111526	0.883566802	0.942763698	0.88537658	0.88537658	0.88344732
Betweenness Centrality	0.955776186	0.956111526	1	0.91409234	0.973446688	0.91130559	0.91130559	0.922475211
Degree	0.875444804	0.883566802	0.91409234	1	0.894207097	0.96554629	0.96554629	0.986466538
Out-Degree	0.939190063	0.942763698	0.973446688	0.894207097	1	0.90411615	0.90411615	0.900307487
Authority	0.870134988	0.885376578	0.91130559	0.965546293	0.90411615	1	1	0.97378105
Hub	0.870134988	0.885376578	0.91130559	0.965546293	0.90411615	1	1	0.97378105
PageRank	0.884639039	0.88344732	0.922475211	0.986466538	0.900307487	0.97378105	0.97378105	1

Kendall's correlation coefficient, not only provides the relationship between variables but also provides a distribution-free set of independence. It can be computed as:

$$r = \frac{n_c - n_d}{n(n-1)/1}$$

(10)

Kendall's rank correlation method depends on concordant and discordant. Concordant is defined as ($x_i > y_i$) else (x_i, y_i) pair is Discordant. Here n_c denotes a total number of concordant while n_d represents a total number of discordant and n represents a total number of observations. Tables 7 and 8 show correlation coefficient results using Kendall's correlation method for Tourism and Health data sets respectively.

TABLE VII. KENDALL'S CORRELATION COEFFICIENT MATRIX FOR DIFFERENT GRAPHICAL MEASURES FOR THE TOURISM DATA SET

	Closeness Centrality	In-Degree	Betweenness Centrality	Degree#	Out-Degree	Authority	Hub	PageRank
Closeness Centrality	1	0.67886972	0.913412546	0.89499822	0.888640662	0.8889581	0.888958	0.916315878
In-Degree	0.67887	1	0.660457174	0.67482918	0.680847861	0.6660882	0.666088	0.649413204
Betweenness Centrality	0.913413	0.66045717	1	0.91147801	0.910703855	0.900823	0.900823	0.967643569
Degree	0.894998	0.67482918	0.91147801	1	0.912330518	0.9732006	0.973201	0.917279484
Out-Degree	0.888641	0.68084786	0.910703855	0.91233052	1	0.9113097	0.91131	0.919390226
Authority0.888958	0.66608821	0.900823041	0.97320058	0.911309721	1	1	0.916939153	
Hub	0.888958	0.66608821	0.900823041	0.97320058	0.911309721	1	1	0.916939153
PageRank	0.916316	0.6494132	0.967643569	0.91727948	0.919390226	0.9169392	0.916939	1

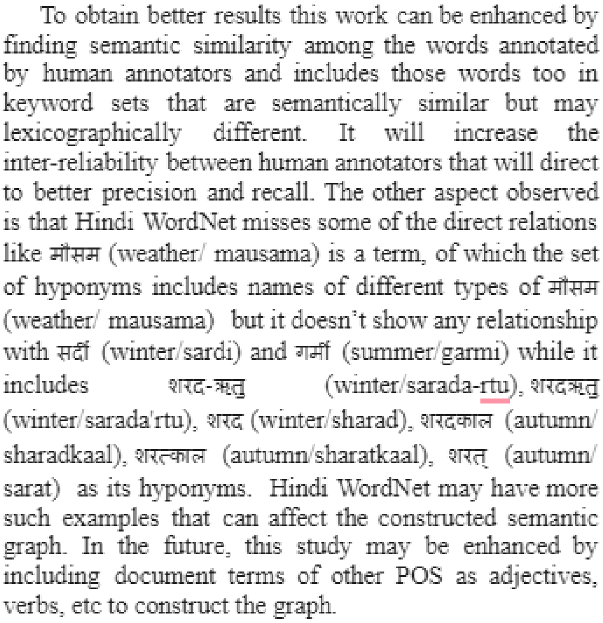
TABLE VIII. Kendall's correlation coefficient matrix for different graphical measures for Health data set

	Closeness Centrality	In-Degree	Betweenness Centrality	Degree#	Out-Degree	Authority	Hub	PageRank
Closeness Centrality	1	0.8485335	0.929894323	0.8382129	0.886818	0.822930	0.82293	0.866740
In-Degree	0.8485335	1	0.8930163	0.8458952	0.908041	0.852857	0.85285	0.826180
Betweenness Centrality	0.9298943	0.8930163	1	0.8787961	0.922135	0.866273	0.86627	0.914390
Degree	0.8382129	0.8458951	0.878796053	1	0.864611	0.932273	0.93227	0.947777
Out-Degree	0.8868183	0.9080418	0.922135679	0.8646116	1	0.878695	0.87869	0.854590
Authority	0.8229301	0.8528569	0.866273518	0.9322738	0.878694	1	1	0.924484
Hub	0.8229301	0.8528569	0.866273518	0.9322738	0.878694	1	1	0.924484
PageRank	0.8667400	0.8261800	0.914390541	0.9477773	0.854590	0.924484	0.92448	1

The above-computed correlation measures (Pearson's, Spearman's, and Kendall's correlation coefficient) exhibits that most of the graphical measures discussed are highly correlated. It is observed that measure Closeness Centrality and Authority are highly correlated with other measures, while In-degree is comparatively less correlated with other measures. The authority and hub are the measures that are circularly related as Hub point good authority vertices and similarly, Authority points to good Hub vertices, hence they are completely correlated to each other and score as 1 through all correlation coefficient measures. It is also being experiential that In-Degree has the least correlation with PageRank measure while on the other side Degree has the highest correlation with Authority and Hub in all the correlation coefficient methods. The reason behind their strong correlation is the fact that Hub/Authority rankings are 'Degree biased' i.e. they are strongly correlated within/out degrees of corresponding nodes. To compute the correlation coefficient we have excluded those measures which have shown very low precision, recall, and F-Measure. Such measures are not shown in the above methods of the correlation coefficient.

5. CONCLUSION AND FUTURE WORK

The study proposed a novel semantic network-based approach for automatic keyword extraction. The approach represents the document in the form of a graph where vertices represent noun terms of the document and edges symbolize the semantic relationship between noun terms. To get semantic relationship Hindi WordNet is used as a background knowledge source. A set of graph-theoretic measures applied to extract keywords based on their semantic scores. The results are compared with the state-of-the-art keyword extraction method TextRank as well as with gold standard keywords. The results are evaluated based on precision, recall, and F-measure. The analysis shows that the set of keywords generated by the proposed approach are better matched with the gold standard keywords as compared to the match of keywords generated through the TextRank approach with gold standard keywords. It is noticed that the closeness centrality measure produces better precision and recall as compared to other graphical measures in the case of matching with gold standard keywords. Another graphical measure that came up as a good measure to extract keywords is Eccentricity. On the other hand, Authority proved as a fine graphical measure to produce keywords through the proposed approach matching with TextRank. The proposed approach is independent of language and domain. It can be applied to the data set of any language and any domain. The application of keywords extracted can be constructive for Topic Detection of a document, Title Construction of the document, Text Summarization, Semantic Indexing, semantic labeling and similarly to comment on Nature of document.



To obtain better results this work can be enhanced by finding semantic similarity among the words annotated by human annotators and includes those words too in keyword sets that are semantically similar but may lexicographically different. It will increase the inter-reliability between human annotators that will direct to better precision and recall. The other aspect observed is that Hindi WordNet misses some of the direct relations like मौसम (weather/ mausama) is a term, of which the set of hyponyms includes names of different types of मौसम (weather/ mausama) but it doesn't show any relationship with सर्दी (winter/sardi) and गर्मी (summer/garmi) while it includes शरद-ऋतु (winter/sarada-rtu), शरदऋतु (winter/sarada'rtu), शरद (winter/sharad), शरदकाल (autumn/ sharadkaal), शरत्काल (autumn/sharatkaal), शरत् (autumn/ sarat) as its hyponyms. Hindi WordNet may have more such examples that can affect the constructed semantic graph. In the future, this study may be enhanced by including document terms of other POS as adjectives, verbs, etc to construct the graph.

In this study, we have also computed correlation coefficients between different graphical measures using Pearson's, Spearman's, and Kendall's correlation coefficient methods. The correlation coefficient scores show that most of the graphical measures are highly correlated. The proposed work can also be applied to other languages and data set of a closed and open domain as well.

6. REFERENCES

[1] W. D. Abilhoa, and L. N. De Castro, "A keyword extraction method from twitter messages represented as graphs," Applied Mathematics and Computation, vol. 240, pp. 308-325, 2014.

[2] R. Baeza-Yates and B. Ribeiro-Neto, Modern information retrieval Vol. 463, New York: ACM Press, 1999.

[3] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open-source software for exploring and manipulating networks." Icwsm, pp. 361-362, August 2009.

[4] S. Beliga, A. Meštrović, and S. Martinčić-Ipšić, "Toward selectivity based keyword extraction for Croatian news." arXiv preprint arXiv:1407.4723, 2014.

[5] S. K. Bharti, K. S. Babu, A. Pradhan, S. Devi, T. E. Priya, P. Konwar et al., "Automatic keyword extraction for text summarization in multi-document e-newspapers." articles. European Journal of Advances in Engineering and Technology, vol. 4(6), pp. 410-427, 2017.

[6] F. Boudin, "A comparison of centrality measures for graph-based keyphrase extraction." In Proceedings of the sixth international joint conference on natural language processing pp. 834-838, October 2013.

[7] P. L. Chen and S. J. Lin, "Automatic keyword prediction using Google similarity distance." Expert Systems with Applications, vol. 37(3), pp. 1928-1938, 2010.

[8] H. Cohen-Kerner, Automatic extraction of keyword from abstracts." lecture notes in computer science, 2773, 843-849, 2003.

[9] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. "A practical part-of-speech tagger." In Third Conference on Applied Natural Language Processing pp. 133-140, March 1992.

[10] S. Deepthi, B. Rajesh, N. Vyshnavi and K. M. Sushma, "Automatic Key Term Extraction from Research Article using Hybrid Approach." International Journal of Computer Applications, 975, 8887, May 2017.

[11] R. Devika and V. Subramaniyaswamy. "A semantic graph-based keyword extraction model using ranking method on big social data." Wireless Networks, 1-13, 2019.

[12] G. M. de Schryver, "Lexicographers' dreams in the electronic[U+2010]dictionary age." International Journal of Lexicography, 16(2), 143-199, 2003.

[13] N. Firoozeh, A. Nazarenko, F. Alizon, and B. Daille, "Keyword extraction: Issues and methods". Natural Language Engineering, vol. 26(3), pp. 259-291, 2020.

[14]J. L. Fleiss, B. Levin and M. C. Paik. Statistical methods for rates and proportions. john wiley sons, 2013.

[15]E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin and C. G. Nevill-Manning, "Domain-Specific Keyphrase Extraction", to appear in: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, 1999.

[16]T. R. Gruber, "A translation approach to portable ontology specifications." Knowledge acquisition, 5(2), 199-220, 1993.

[17]J. Hauke and T. Kossowski, "Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data." Quaestiones geographicae, vol. 30(2), pp. 87-93, 2011.

[18]A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge," In Proceedings of the 2003 conference on Empirical methods in natural language processing, pp. 216-223, 2003.

[19]M .T. Khan, Y. Ma and J. J. Kim, "Term Ranker: A Graph-Based Re-Ranking Approach." In FLAIRS Conference, pp. 310-315, March 2016.

[20]M. M. Khapra, A. Kulkarni, S. Sohoney and P. Bhattacharyya, "All words domain adapted WSD: Finding a middle ground between supervision and unsupervision." In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1532-1541, July 2010.

[21]M. Litvak, M. Last, H. Aizenman, I. Gobits and

A. Kandel, “DegExt—A language-independent graph-based keyphrase extractor.” In Advances in intelligent web mastering–3, pp. 121-130. Springer, Berlin, Heidelberg, 2011.

[22]J. Martinez[U+2010]Romo, L. Araujo and A. Duque Fernandez, “Sem Graph: Extracting keyphrases following a novel semantic graph[U+2010]based approach.” Journal of the Association for Information Science and Technology, vol. 67(1), pp. 71-82, 2016.

[23]Y. Matsu and M. Ishizuka, “Keyword extraction from a single document using word co-occurrence statistical information.” International Journal on Artificial Intelligence Tools, vol. 13(01), pp. 157-169, 2004.

[24]R. Mihalcea and P. Tarau, “Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing, pp. 404-411, July 2004.

[25]R. Mihalcea and D. Radev, Graph-based natural language processing and information retrieval. Cambridge university press, 2011.

[26]G. A. Miller, “WordNet: a lexical database for English.” Communications of the ACM, vol. 38(11), pp. 39-41, 1995.

[27] J. Mothe, F. Ramiandrisoa and M. Rasolomanana. Automatic keyphrase extraction using graph-based methods. In Proceedings of the 33rd Annual ACM Symposium on Applied Computing, pp. 728-730, April 2018.

[28] R. Nagarajan, S. Nair, P. Aruna, and N. Puviarasan, “Keyword extraction using graph-based approach.” International Journal of Advanced Research in Computer Science and Software Engineering, vol. 10(10), 2016.

[29] D. Narayan, D. Chakrabarti, P. Pande, and P. Bhattacharyya, “An experience in building the indo wordnet-a wordnet for Hindi.” In First International Conference on Global WordNet, Mysore, India, Vol. 24, January 2002.

[30] T. D. Nguyen and M. Y. Kan, “Keyphrase extraction in scientific publications,” In International conference on Asian digital libraries pp. 317-326, Springer, Berlin, Heidelberg, December 2007.

[31] Y. Ohsawa, N. E. Benson and M. Yachida, “KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor.” In Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries-ADL’98- (pp. 12-18). IEEE, April 1998.

[32] D. Paranyushkin, “Identifying the pathways for meaning circulation using text network analysis.” Nodus Labs, vol. 26, 2011.

[33] H. Redkar, R. Shukla, S. Singh, J. Saraswati, L. Kashyap, D. Kanojia, P. Bhattacharyya et al. “ Hindi Wordnet for Language Teaching: Experiences and Lessons Learnt.” In Proceedings of the 9th Global WordNet Conference (GWC 2018) (p. 317), January 2018.

[34] K. Sarkar, “A hybrid approach to extract keyphrases from medical documents.” arXiv preprint arXiv:1303.1441, 2013.

[35] C. Sharma, M. Jain and A. Aggarwal, “Keyword Extraction Using Graph Centrality and WordNet.” In Towards Extensible and Adaptable Methods in Computing, pp. 363-372, Springer, Singapore, 2018.

[36] A. Tixier, F. Malliaros and M. Vazirgiannis, “A graph degeneracy-based approach to keyword extraction.” In Proceedings of the 2016 conference on empirical methods in natural language processing pp. 1860-1870, November 2016 Complex Processes, pp. 575-585. Springer, Berlin, Heidelberg, 2008.

[37] D. A. Vega-Oliveros, P. S. Gomes, E. E. Milios and L. Berton, “A multi-centrality index for graph-based keyword extraction.” Information Processing Management, vol. 56(6), 102063, 2019.

[38] X. Wan and J. Xiao (2008, July). “Single Document Keyphrase Extraction Using Neighborhood Knowledge.” In AAAI Vol. 8, pp. 855-860 , July 2008.

[39] M. S. Tran-L, T. T. Vo-Dang, Q. Ho-Van and T. K. Dang, “Automatic Information Extraction from the Web: An HMM-Based Approach.” In Modeling, Simulation and Optimization of

[40] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin and C. G. Nevill-Manning, “Kea: Practical automated keyphrase extraction.” In Design and Usability of Digital Libraries: Case Studies in the Asia Pacific, pp. 129-152. IGI global, 2005.

[41] C. Wu, M. Marchese, J. Jiang, A. Ivanyukovich and Y. Liang, “Machine Learning-Based Keywords Extraction for Scientific Literature.” J. UCS, vol. 13(10), pp. 1471-1483, 2007.

[42] C. S. Yadav, A. Sharan and M. L. Joshi, “Semantic graph based approach for text mining.” In 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), pp. 596-601, IEEE, February 2014.

[43] C. Zhang, “Automatic keyword extraction from documents using conditional random fields.” Journal of Computational Information Systems, vol. 4(3), pp. 1169-1180, 2008.

[44] K. Zhang, H. Xu, J. Tang, and J. Li, “Keyword extraction using support vector machine.” In international conference on web-age information management, pp. 85-96, Springer, Berlin, Heidelberg, June 2006.

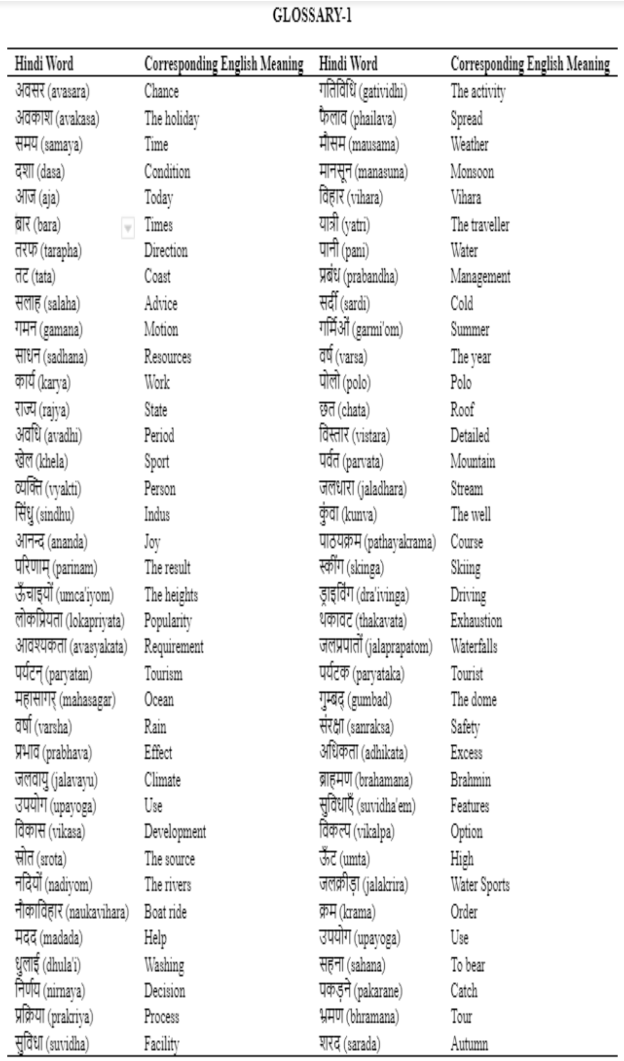
[45] Z. Zhou, X. Zou, X. Lv and J. Hu, “Research on weighted complex network based keywords extraction.”

In Workshop on Chinese Lexical Semantics, pp. 442-452, Springer, Berlin, Heidelberg, May 2013.
[46] “ ” (Hindi Wordnet),” Hindi Wordnet. [Online]. Available: http://www.cfilt.iitb.ac.in/wordnet/webhwn/ Accessed 23 August 2020.

Ms. Manju Lata Joshi
Ms. Manju Lata Joshi is a research scholar in the Department of Computer Science at Banasthali Vidyapith, Rajasthan, India. Her areas of interests include Natural language Processing, Information Retrieval and Text Mining.

Dr. Namita Mittal
Dr. Namita Mittal is working as Associate Professor in Department of Computer Science and Engineering at Malviya National Institute and Technology, Jaipur, India. She is a recipient of Career Award for Young Teachers (CAYT) by AICTE. Her Current research areas are Data Science, Information Retrieval, Data Mining and NLP.

Dr. Nisheeth Joshi
Dr. Nisheeth Joshi is an Associate Professor in the Department of Computer Science at Banasthali Vidyapith, Rajasthan, India. He primarily works in Machine Translation, Information Retrieval and Cognitive Computing. He has over 12 years of experience.



GLOSSARY-1

Hindi Word	Corresponding English Meaning	Hindi Word	Corresponding English Meaning
अवसर (avasara)	Chance	गतिविधि (gatividhi)	The activity
अवकाश (avakasa)	The holiday	फैलाव (phailava)	Spread
समय (samaya)	Time	मौसम (mausama)	Weather
दशा (dasa)	Condition	मानसून (manasuna)	Monsoon
आज (aja)	Today	विहार (vihara)	Vihara
बार (bara)	Times	यात्री (yatri)	The traveller
तरफ (tarapha)	Direction	पानी (pani)	Water
तट (tata)	Coast	प्रबंध (prabandha)	Management
सलाह (salaha)	Advice	सर्दी (sardi)	Cold
गमन (gamana)	Motion	गर्मिओं (garmi'om)	Summer
साधन (sadhana)	Resources	वर्ष (varsa)	The year
कार्य (karya)	Work	पोलो (polo)	Polo
राज्य (rajya)	State	छत (chata)	Roof
अवधि (avadhi)	Period	विस्तार (vistara)	Detailed
खेल (khela)	Sport	पर्वत (parvata)	Mountain
व्यक्ति (vyakti)	Person	जलधारा (jaladhara)	Stream
सिंधु (sindhu)	Indus	कुंवा (kunva)	The well
आनन्द (ananda)	Joy	पाठयक्रम (pathayakrama)	Course
परिणाम् (parinam)	The result	स्कींग (skinga)	Skiing
ऊँचाइयों (umca'iyom)	The heights	ड्राइविंग (dra'ivinga)	Driving
लोकप्रियता (lokapriyata)	Popularity	थकावट (thakavata)	Exhaustion
आवश्यकता (avasyakata)	Requirement	जलप्रपातों (jalaprapatom)	Waterfalls
पर्यटन् (paryatan)	Tourism	पर्यटक (paryataka)	Tourist
महासागर् (mahasagar)	Ocean	गुम्बद् (gumbad)	The dome
वर्षा (varsha)	Rain	संरक्षा (sanraksa)	Safety
प्रभाव (prabhava)	Effect	अधिकता (adhikata)	Excess
जलवायु (jalavayu)	Climate	ब्राह्मण (brahamana)	Brahmin
उपयोग (upayoga)	Use	सुविधाएँ (suvidha'em)	Features
विकास (vikasa)	Development	विकल्प (vikalpa)	Option
स्रोत (srota)	The source	ऊँट (umta)	High
नदियों (nadiyom)	The rivers	जलक्रीड़ा (jalakrira)	Water Sports
नौकाविहार (naukavihara)	Boat ride	क्रम (krama)	Order
मदद (madada)	Help	उपयोग (upayoga)	Use
धुलाई (dhula'i)	Washing	सहना (sahana)	To bear
निर्णय (nirnaya)	Decision	पकड़ने (pakarane)	Catch
प्रक्रिया (prakriya)	Process	भ्रमण (bhramana)	Tour
सुविधा (suvidha)	Facility	शरद (sarada)	Autumn