An Insight into Role of Wordnet and Language Network for Effective IR from Hindi Text Documents

Manju Lata Joshi
Banasthali University
Vanasthali, Rajasthan
manjulatajoshi@gmail.com

Namita Mittal
Malviya National Institute of Technology
Jaipur, Rajasthan
nmittal.cse@mnit.ac.in

Nisheeth Joshi
Banasthali University
Vanasthali, Rajasthan
jnisheeth@banasthali.in

ABSTRACT

This paper investigates the limitations of traditional Information Retrieval (IR) models and how the semantic based approaches overcomes these limitations. Further the paper analyzes a range of aspects of language network representation of text corpus and how different network properties can lead to improve the results for different applications of IR. The paper analyzes Hindi Wordnet to exploit its capabilities and applicability as knowledge source and then its limitation. The paper discusses various research issues yet to be explored in area of IR of Hindi text documents. This paper suggests that how application of fuzzy logic in semantics can improve the performance of IR outcomes. Our entire analysis is in relevance to Hindi language corpus.

CCS CONCEPTS:

• Analysis of Traditional IR models →Construction of Language Network using Hindi Wordnet as Background Knowledge → Exploring Graphical Properties of Language Network for different Applications of IR →Applying Fuzzy Logic on Semantics of Hindi Wordnet

KEYWORDS

Hindi Wordnet, Semantic Graph based IR, Fuzzy Logic based Semantics

ACM Reference Format:

Manju Lata Joshi, Namita Mittal and Nisheeth Joshi. 2017 An Insight into Role of Wordnet and Lexical Network for Effective IR from Hindi Text Documents. In Proceedings of Forum for Information Retrieval Evaluation (FIRE'17). ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1. INTRODUCTION

Since the emergence of web, Information Retrieval (IR) and Information Extraction (IE) are the topics of intensive research. Mostly two types of approaches are used for IR, one is Corpus based [20] and another is knowledge based [21]. In knowledge based approaches some background knowledge is used, by considering thesaurus, machine readable dictionaries or ontologies etc. For last few years there had been lack of ontology based application but now the scenario is changing since last few years. In past few years Researchers reported various applications of ontologies especially in area of IR and IE. However Most of the work done in this area is on English language corpus.

1.1 Traditional Approaches for IR & IE

In IR the entire idea of finding relevant documents depends on finding similarity between query and documents. Therefore an efficient information retrieval system is required to use proper similarity measures to match query and documents in order to satisfy needs of user. Thus the performance of IR system heavily depends on the similarity measure used by an IR system. There are several retrieval strategies available which assign a measure of similarity between a query and a document for the tasks such as document matching, ranking, clustering etc. Most of these strategies are based on frequency of the terms found in both the document and the query, more "relevant" the document is considered to be to the query. Traditional retrieval strategies: Boolean Model, Vector Space Model (VSM), and Probabilistic Model are based on keyword based similarity. Among all these VSM is most popular model. Although the VSM model is a very simple count model and may work well in many cases, still it has many limitations as long documents are poorly represented because they have poor similarity values, "False **positive match",** Semantic sensitivity leading **"False negative match",** Scoring Phrases of words difficult, does not support Boolean queries etc. Semantic similarities and Ontologies provide a way to overcome these limitations of traditional keyword based approaches.

1.2 Semantic Similarity and Ontologies

Lexicographic based similarity considers only keyword match and does not consider meaning of words to measure similarity therefore it is not an effective approach. We require an approach which considers keyword match and semantics of words as well. Apart from that it must also consider semantic relationship between the words. For example if we find similarity between 'car' and 'automobile' by using lexicographic based similarity they will not match but if we compare these words by using semantic similarity, we find that these are similar because users uses these two words alternatively. In other words, instead of dealing with words these measures deal with concepts of words. Ontologies provide a tool to find semantic similarity between the terms. Formally Gruber [2] defined ontology as a "Shared specification of conceptualization". More broader and understandable view of ontology is that *"Ontology is a data model that represents a set of concepts without or within a domain and the relationships between those concepts. It is used to reason about the objects without or within that domain"*. There are various ontologies available such as WordNet [3], HowNet, ConceptNet, Hindi Wordnet, Indo Wordnet etc. The basic model of ontologies comprises of concepts and relationship between them. Concepts can be taken as sets in form of nodes in ontology and edges symbolize relationship between these concepts. Most of the ontologies include various relationships to represent proximity among concepts. These

relationships include: Hypernym - Hyponym (car is-a vehicle), Meronym -Holonym (Earth is Part-of solar system) etc.

1.3 Hindi Wordnet

The Hindi Wordnet 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 Hindi Wordnet is inspired by the famous English WordNet. In the Hindi WordNet the words are grouped together according to their similarity of meanings. Two words that can be interchanged in a context are synonymous in that they represent same lexical concept. This is done to remove ambiguity in cases where a multiple words shares same meaning. Synsets are basic building blocks of Wordnet. The Hindi Wordnet contains content words, from different part of speeches as Noun, Verb, Adjective and Adverbs. Every entry in the Hindi Wordnet contains synset, gloss and position in Ontology. Relations in Hindi Wordnet are Hyponymy-Hypernymy, Meronymy-Holonymy, Entailment, Troponymy, Antonymy, Gradation and Causative. The relations between different POS provided by Hindi Wordnet are Linkages between nominal and verbal concepts, Linkages between nominal and adjectival concepts and Linkages between adverbial and verbal concepts. In comparison to English WordNet, Hindi has some inherent limitations. For eg. In Hindi Wordnet single term representation is present while in many cases one has to take the compound words to consideration. Apart from this, several other applications have been developed over English WordNet, which are not available for Hindi Wordnet. One such important application is tool for finding semantic similarity between two words using WordNet semantic similarity measures. These applications can be very useful for retrieving and exploring information from text documents.

1.4 IR and IE from English vs. Hindi Text Documents:

As most of the work in IR and IE has been done for English corpus, it is it is important to analyze whether it can be applied for Hindi language or not. From this point of view it is also important to find out some basic differences between Hindi and English language text.

The structural differences between English and Hindi are mostly attributed to the difference in their word orders. Language topologists' classify English as an SVO (Subject-Verb-Object) language and Hindi as an SOV (Subject-Object-Verb) language. Moreover, Hindi is a free word order language. Secondly, in Hindi the preposition comes after the noun or pronoun hence more appropriately named as postposition, in contrast to preposition in English. Apart from these differences, in English there is lot of significance of capital letters while no such distinction in Hindi. For e.g. In English nous always start with capital letter but in Hindi it is not so. In Hindi Consonant letters carry an inherent vowel which can be altered or muted by means of diacritics or *matra*.

1.5 Semantic Graphs or Language Networks:

In terms of computational processing, a text is an unstructured data with multiple units- words, phrases, lines, paragraphs or the entire document. All these text entities are connected to each other through semantic relations that contribute to the overall meaning, maintain cohesive structure and discourse unity of the text. [18]. Although, To represent texts developing a computationally practicable model is a difficult task. Considering these limitations a different approach can be considered, where text is represented as a graph where nodes represent word, sentences etc. and the edges corresponds to the relationship between these entities. Once the text corpus is converted into form of semantic graph, various language network properties can be analyzed for a variety of text mining applications using computational network analysis.

1.6 Motivation & Objective:

Most of the research and development in IR and IE has been done in the development of either CLIR (Cross Language Information Retrieval) system or purely for English language corpus. Considering knowledge based approaches for IE, WordNet [1] has been extensively used for IR and IE for English language corpus. There is little exploration for Hindi language corpus although Hindi is third most spoken language of world. Aim of this work is to target Hindi corpus for Effective IR and IE. The language constructs, query structure, general words etc. are entirely diverse in Hindi as compared to English. Hence to cope-up with the variations in Hindi, in addition to conventional search and NLP techniques, some novel strategy must be used to construct IR & IE system for Hindi corpus. These aspects provide a motivation for exploring the Hindi Wordnet for retrieving and extracting information from Hindi text documents. Hindi Wordnet is very much similar to English WordNet in features. Hindi being a resource poor language, external knowledge base such as Wordnet can prove to be useful for retrieving and exploring information from Hindi text documents.

The goal of the work is to use Hindi Wordnet ontology as a knowledge source for retrieving and extracting documents from Hindi corpus. The proposed research is planned to convert the text document(s) in language network (Semantic Graph based network). Once the document is converted to semantic graph, various graphical measures can be explored for several application of IR such as clustering, summarization, WSD, Key word extraction, Query Expansion etc. This work further explores the limitations of Hindi Wordnet and then how these limitations can be overcome using Fuzzy Logic on semantics of the document.

In order to delimit the scope of this work and to identify specific objectives, various sub-areas of IR and IE where Hindi WordNet can be applied have been explored. A survey of the work already done in these areas has been done to identify the possibilities that can be explored for future work. Based on our study and observations, following are specific objectives:

1. To investigate the work done on English language using WordNet for various applications of IR and verify whether these measures are also pertinent to use to improve IR efficiency from Hindi text documents.
2. To explore the various applications of Hindi Wordnet for retrieving & extracting relevant information from Hindi text documents in general.
3. Explicit study of the role of Wordnet for Keyword Extraction & Concept Generation and Automatic Text Summarization.
4. Construct Semantic Graph for whole Hindi corpus(s)

5. Explore the different graphical properties of language network for certain IR applications.
6. Analyzing the limitations of Hindi Wordnet and beat them by applying Fuzzy Logic on semantics of document to gain better and more relevant results.
7. Comparative analysis among results produced after experimenting using Tf-Idf method, Semantic graph based IR and Fuzzy semantic graph based IR.

The discussed scenario can be presented using a framework as follows:

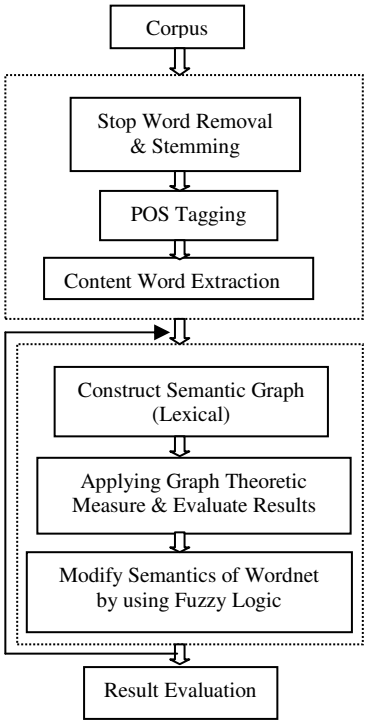

Figure 1: Proposed Framework for Lexical Network IR using Fuzzy Semantics

1.7 Research Gap

Hindi WordNet has been used by researchers for several purposes but limited work has been done for exploring exhaustive role of Hindi WordNet and its applications for different text mining aspects, the work is an effort to fill the existing research gap through detailed study of WordNet as a tool for different applications of IR & IE. Apart from this in regard with language network large numbers of graphical properties are still unexplored for retrieval and extraction of relevant documents in appropriate context. The target corpus is in Hindi Language (of open and closed domain both) where very less work has done**.**

This paper is tentatively divided in four sections. Second section represents the work done in area of Keyword Extraction & Concept Generation and Automatic Text Summarization. Third section is a theoretical description of our proposed work. Section four concludes our work with summary of the paper and discusses further applications of semantic network and how fuzzy logic can improve results of text mining.

2. LITERATURE SURVEY

This section investigates the work done so far in two key applications of IR namely Keyword Extraction & Concept Generation and Automatic Text Summarization.

1.1 Keyword Extraction and Concept Generation:

Automatic keyword extraction is a method by which representative terms are systematically extracted from a text with either minimal or no human intervention, based on the model. Keyword extraction techniques can either be Statistical, Linguistic based, Machine Learning based or Hybrid.

Statistical approaches are straightforward, don't need the training data and have very little requirements, emphasizing on non-linguistic features of the text such as tf, idf and location of a keyword in document.. Further for finding statistical relationship between words Co-occurrence based measures, Mutual information based measures, Lexical link based measures can be used. The previous research shows that VSM based on single word information does not provide sufficient statistics required for IR and IE. Therefore N-Gram based statistical approaches are required in order to generate required statistical information. It has also been found that a relationship exist between the importance of word and number of senses of word (which can be determined from WordNet). For eg. Frequently used stop words generally have less number of senses as compared to rarely used words. This aspect can be explored for finding importance of a word. The benefit of purely statistical methods is that they are easy to use and generally produces superior results**.** Linguistics Approaches use the linguistic features of the words, sentences and document. These approaches focuses on linguistic features such as, syntactic structure, part-of-speech and semantic similarities, functioning sometimes as filters for bad keywords. Majorly two lexical resources: The EDR electronic dictionary and Princeton University's WordNet [3] are being used by researchers. These sources provide robust lexicons including semantic relationships and linking. Experiments show that lexical resources produce better results as compared to statistical method. Hence, some of the linguistic methods are hybrid methods, combining few linguistic methods with widespread statistics based measures such as tf and idf.

In Machine Learning approaches system is trained through a set of training documents. Every document contains range of human preferred keywords as well. Then the achieved knowledge is applied to set of documents to be tested.

Frank, E. et. Al [4] uses the machine learning techniques and Naive Baye's method for extracting technical key phrases from documents of some specific domain. This extraction process is divided into two phases: term-weighting and keyword extraction. In first phase, a set of feature vectors is generated on a set of newspaper articles and then from different encyclopedia domains. Both vectors are compared using a similarity calculation so the news paper articles can be separate into different domains, then sorting is performed for producing the ultimate set of feature vectors. In the second phase, of keyword extraction, a segment is analyzed and the most relevant domain

is selected for it using the pre-existing feature vectors. **Hybrid Approaches** to extract keywords primarily merge the methods discussed or use some heuristic knowledge. The parameters commonly used are the position, layout feature of the words, length, html tags around of the words, etc.

WordNet has also been used for extracting concepts from documents. This Wordnet based approach has been found useful for QE, Document classification and Document clustering. Kang, B. et. al. [5] has explored the use of Wordnet for generating concept graphs. Further the graph theoretic techniques can be used to explore the derived concepts.

Graph based Approaches are based on exploration of network properties like degree, clustering coefficient, structural diversity index, strength, neighborhood size, page rank, HITS hub and authority score, betweenness, closeness and eigenvector centrality. Researchers suggest that centrality measures outperform the basic tf-idf model.

Mihalcea and Tarau [6] introduced state-of-the-art TextRank model. TextRank is derived from PageRank and intiated to graph based text processing for keyword and sentence extraction. The abstracts are modeled as directed or undirected and weighted co-occurrence networks using a co-occurrence window of variable sizes (2-10). The PageRank motivated score of the importance of the node derived from the importance of the neighboring nodes is used for keyword extraction. The attained TextRank performance compares favorably with the supervised machine learning n-gram based approach.

Litvak and Last [7] compared supervised and unsupervised approaches for keywords identification and then for effective summarization. These approaches are based on graph based syntactic representation of text and web documents. The results of HITS algorithm on a set of summarized documents performed comparable to supervised methods (Naïve Bayes, J48, SVM).

Tsatsaronis et al**.** in [8] present Semantic Rank algorithm which is based on network for keyword and sentence extraction from text. Boudin [9] compares several centrality measures for key phrase extraction and experiments on standard data sets of French and English proves that simple degree centrality achieves considerable outcomes comparable to the TextRank algorithm.

Zhou et al. [10] investigate weighted complex network based keyword extraction. On the basis of closeness centrality Importance of each node to become a keyword candidate is calculated. The experimental evaluation shows preferable effect on correctness, recall and F-value over the classic TF-IDF method.

Abilhoa and de Castro [11**]** propose a keyword extraction method representing tweets in form of graphs and applying centrality measures for finding the related keywords. They develop technique named Twitter Keyword Graph. Keywords are extracted by applying graph centrality measures – closeness and eccentricity. The performance of the algorithm is compared with the TF-IDF approach and KEA algorithm and gained good results in terms of computation.

2.2 Automatic Text Summarization:

Text summarization is a process to select the most "representative" sentences that can form the summary of a document. Formally, Text summarization is defined as "*to distill the most important information from a source or sources to produce an abridged version of it*" [12].

Text Summarization methods can be classified into **Extractive** and **Abstractive** summarization. Extractive summary is set of important sentences; paragraphs etc. from the base document and combine these into shorter form. The significance of extracted sentences is decided based on linguistic and statistical features of sentences. Whereas in an abstractive summarization an attempt is made to develop understanding of the main concepts in a document and then express those concepts in clear natural language. The most common features used for extractive summarization are Sentence location feature, Title word feature, Length feature, Proper Noun feature, Content word (Keyword) feature, Upper-case word feature, Biased Word Feature, Font based feature, Pronouns, Sentence-to-Sentence Cohesion, Sentence-to-Centroid Cohesion, Cue-Phrase Feature, Occurrence of non-essential information, Discourse analysis. The methods used for extractive summarization are Cluster based method, Machine Learning approach, Graph theoretic approach, LSA Method, with neural networks, based on fuzzy logic, using regression for estimating feature weights etc.

The text summarization can be of two types based on span of text used for processing i.e. Single Document Summarization and Multiple Document Summarization. The most common and recent text summarization techniques use either Statistical approaches (based on word clustering, tf.idf, chi-squared) or Linguistic approaches (based on Lexicons and Dictionaries, Latent semantic analysis, WordNet, etc.), or some kind of linear combination of these. There has also been done a lot of work on text summarization using supervised and Semi-Supervised techniques for English language.

An algorithm proposed by Bellare K. et. al., 2004 is based on identifying semantic relations and is for generic text summarization. They use WordNet to understand the links between different parts of the document; subsequently extract the portion of the WordNet graph which is most relevant. The algorithm selects sentences on the basis of their semantic content and its relevance to the main ideas contained in the text. The algorithm was tested on DUC'2002 data sets and their reference summaries.

The results were also compared with the well known text summarizer: Copernicus [13]. Even though their average results are slightly worse than that of Copernicus', the algorithm is simple, repeatable and the results can be verified, unlike Copernicus'. Moreover, it is a novel approach and therefore, extending it can improve results significantly.

A number of interesting possibilities remain that can be explored in future. Firstly, the parameters used for generating summaries, eg, weightage given to different parts of speech, can be learned given a corpus of documents. Then, Resolution of pronouns can be used on top of the WordNet approach to get summaries which are more readable and have less dangling anaphors. Apart from this machine learning and soft computing techniques can be used on top of WordNet to learn parameters from documents for text summarization. One of such approach was also proposed by [14]. The idea of their approach is to find out key sentences by extracting keywords based on statistics and Synsets using

WordNet.. Semantic similarity analysis is conducted between candidates of key sentences to reduce the redundancy. Refining key sentences against WordNet semantic similarity comprehensively improved the correctness of automatic summary since redundancy is reduced to the minimum. The results show that the approach achieves reasonable performance compared with a commercial text summarizer (Microsoft Word Summarizer).

As far as Indian languages are concerned very little work has been done for automatic text summarization, Patel A. et. al., 2007 worked on a language independent approach to multilingual text summarization. Their paper presents a statistical approach to generate generic extractive summary. They have developed an algorithm for automatically generating a generic summary of a single document. The algorithm is highly flexible and requires only a stop words list (provided externally) and stemmer for corresponding language in which documents are to be summarized. A method is suggested to derive a vector representing the central idea (theme) of the document. Location feature complexity has been handled by partitioning the text and extracting 'best' sentences from each partition. Sentences, which are not complete by themselves, lead to inclusion of their corresponding preceding sentences to resolve the gaps in context and meaning. Summaries are generated at four fuzzy levels, viz. Normal, Short, Brief, Very Brief. Experimentation performed on standard data sets exhibits that the outcomes obtained are comparable with those of state-of-the-art systems for automatic summarization, while at the same time providing the benefits of a robust language independent algorithm. The quality of summary is tested w. r. t. its degree of representativeness for languages other than English. The results are encouraging. All the summaries tested by them include sentences of importance. However in some cases, it was found the flow of the summarized text not to be very smooth. Generally, different languages involve different complexity of their own semantics, making it harder to apply natural language processing.

CDAC (Centre for development of advance computing) Noida developed Automatic text summarization software for Hindi text [13]. It combines Statistics based technique, language oriented & heuristic technique for text summarization.

Chetana Thaokar and Latesh Malik [15] uses sentence extraction method to summarize Hindi text documents. To optimize the summary generated, genetic algorithm is applied. The summary generated cover maximum idea with a smaller amount redundancy.

M. Subramanium and Dalal [16] proposed a novel approach to create an abstractive summary for a single document. The approach creates a Rich Semantic Graph for the original document, reducing the generated semantic graph to more abstracted graph, and generating the abstractive summary from the reduced graph.

V. Dalal and L. Malik [17] proposed an approach for summarizing Hindi text document using semantic graph and particle swarm optimization algorithm. The approach proven ability in searching optimal solution, in spite of large dimensionality of the solution space. The approach is tested and results compared with other existing approaches on basis of Precision, Recall, F Measure and G Score.

3. PROPOSED WORK

Based on the related work discussed in section 2 some research gaps in various aspects of IR and IE have been identified. The research work will be delimited to following research issues, depending on the feasibility of resources (such as based on availability of corpus and other resources for Hindi language).

3.1 Exploring POS Tagger Information for IR & IE: Most of the work done so far in IR and IE is based on the Noun content of the document; the reason being is that most of the semantics of document are provided by the nouns only. But research shows that other POS such as verb, adverb etc. may be very significant for extracting information from documents**.** Number of approaches has been used for Part of Speech (POS) tagging as Rule Based approach, Statistical approach and Hybrid approaches for hindi text documents. Apart from this for Hindi corpus the importance of *postposition and conjectures* is very high. For eg. In case of WSD if we disambiguate the ambiguous postpositions the results are expected to be better. Further, we can explore the *cross part of speech linkages* (relationship between the synset of different POS) to get effective results in various sub-areas of IR & IE. It is observed that most of the research done using Wordnet is based on either hypernym-hyponym relationships or Meronym-holonym relationship, so there is a need of exploring other relationships such as Entailment, Troponymy, Antonymy, Gradation and Causative provided by Hindi Wordnet.

3.2 Identification of Semantic Features in Hindi for IR & IE: Hindi is morphologically a very rich language; the work can be carried out using different morphological variations of words. As compared to English, Hindi is a SOV (Subject-Object-Verb) language and apart from this, in Hindi there is lot of significance of diactrics (*matra*). We can also explore the orthographic features like capitalization in case of English, if there is any such feature in Hindi too.

3.3 Rule-Based IR & IE: In case of NLP there is lot of importance of Hand-crafted rules due to several kind of variations language wise. Further exploration of existing as well as new rules for Hindi language such as syntactical, semantical and orthographical can be helpful to gain better results.

3.4 Use of Spell Normalizer: Due to the need of suppressing the effect of morphological variations of words in Hindi language, the use of Spell Normalizer is effective. It uses in-built string processing module to stem the extension of word and convert it into root word.

3.5 Applying Fuzzy Logic in Semantics of Hindi WordNet: The limitation of Hindi Wordnet is that all the relations defined in Hindi Wordnet are crisp in nature i.e. the terms in document are completely related or not related at all. But in many real life scenarios articulated in NLP are matter of degree, there is gradual transition from not being related and being related therefore an association between terms by a mapping $T \times T \rightarrow [0, 1]$ i.e. by fuzzy relation instead of merely traditional relation is mandatory to consider enhancing the performance of IR systems [19].

4. CONCLUSION & FUTURE WORK

There can be few open research issues can be examined such as Wordnet can be explored for considering semantically related terms for QE and for finding nature of query which can further lead to be helpful in improving IR efficiency. Similarly, Machine learning and soft computing based approaches can be used on top of Wordnet to learn parameters for different applications of IR and IE, for e.g. QE, Text Summarization etc. WSD is such an aspect of IR & IE which is needed in many applications like for QE, NER etc. A limited work has been done for WSD specifically for Hindi language. The existing approaches can be improved by applying the proposed approach. Further, it can be explored that whether the disambiguation is beneficial at query level, document level or at the time of adding terms. In addition to this if morphology is handled exhaustively, results can be better for Hindi WSD. In NER, Rule-based NER, more rules can be identified for developing Hindi NER system like theta and thematic rules can be explored for NER.

5. REFERENCES

[1] Fellbaum, C., 1998, WordNet: An Electronic Lexical Database, *the MIT Press, Cambridge, MA*.

[2] Gruber, T. R. 1993. A translation approach to portable ontologies, *Knowledge Acquisition,* 5(2): (1993), 199-220.

[3] Plas, L., Pallotta, V., Ghorbel, H., 2004. Automatic keyword extraction from spoken text. A comparison of two lexical resources: the EDR and WordNet, *Proceedings of the 4th International Conference on Language Resources and Evaluation, European Language Resource Association*, *2004*, 2205-2208.

[4] Frank, E., 1999, Domain-specific Learning Algorithms for Keyphrase Extraction, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99),* pp. 668-673.

[5] Kang, B., Lee, S. 2005, Exploiting Concept Clusters for Content-based Information Retrieval, *Information Sciences,* Volume 170, Issues 2-4, pages. 443–462.

[6] Mihalcea, R., Tarau, P., 2004. TextRank -- bringing order into texts.

[7] Litvak, M., Last, M., Friedman, M. 2010 A new approach to improving multilingual summarization using a genetic algorithm, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p.927-936, July 11-16, 2010, Uppsala, Sweden.

[8] Tsatsaronis, G., Varlamis, I. and Vazirgiannis, M. 2010 Text relatedness based on a word thesaurus, *J. Artif. Intell.* Res., vol. 37, pp. 1–38, 2010.

[9] Boudin, F. 2013. A comparison of centrality measures for graph-based keyphrase extraction. *In Proc. of IJCNLP*, pages 834–838, Nagoya, Japan.

[10] Zhou, Z., Zou,X., Lv, X., Hu, J. 2013 Research on Weighted Complex Network Based Keywords Extraction, in *Lecture Notes in Computer Science* Volume 8229, 2013, pp. 442-452.

[11] Abilhoa, W. , Castro, D. 2014. A keyword extraction method from twitter messages represented as graph. *Applied Mathematics and Computation* v. 240, pp. 308-325, 2014.

[12] Niggemeyer, B. 1998. Summarizing Information, *Springer, New York, NY*.

[13] Copernicus Summarizer: *http://www.copernic.com/en/products/summarizer/*.

[14] Chenghua D., Luo, X. 2008, WordNet Based Document Summarization', *Proceedings of 7 WSEAS Conf. On Applied Computer and Applied Computational Science ACACOS'08)*, HangZhou, China, Apr. 2008.

[15] Thaokar, C., Malik, L. 2013. Test Model for Summarizing Hindi Text using Extraction Method, *Proceedings of 2013 IEEE International Conference on Information and Communication Technologies*,11-12 April 2013, Thuckalay, TN.

[16] Subramaniam, M., Dalal, V. 2015. Test Model for Rich Semantic Graph Representation for Hindi Text using Abstractive Method in Volume: 02 Issue: 02 May-2015, e-ISSN: 2395-0056

[17] Dalal, V., Malik, L., 2017. Semantic Graph Based Automatic Text Summarization for Hindi Documents Using Particle Swarm Optimization, *International Conference on Information and Communication Technology for Intelligent Systems,* pp 284-289.

[18] Yadav, C., Sharan, A., Joshi, M.L., 2014. Semantic Graph Based Approach for Text Mining, *International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)* DOI: 10.1109/ICICICT.2014.6781348.

[19] Jain, A., Lobiyal, D.K., 2016. Fuzzy Hindi WordNet and Word Sense Disambiguation Using Fuzzy Graph , Journal ACM Transaction on Asian and Low Resource Language Information Processing (TALIIP), Vol. 15, Issue 2, Article No. 8, USA, doi>10.1145/2790079.

[20] Jiang, Jay J., David W. Conrath. 1997 Semantic similarity based on corpus statistics and lexical taxonomy , *arXiv preprint cmp-lg/9709008, 1997.*

[21] Porter B., Souther A., 1999. Knowledge-based Information Retrieval, AAAI Technical Report FS-99-02. 1999, AAAI (www.aaai.org).