ijar
ISSN NO. 2320-5407

Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/3616
DOI URL: http://dx.doi.org/10.21474/IJAR01/3616

*RESEARCH ARTICLE*

# KEYWORD EXTRACTION: A COMPARATIVE STUDY USING GRAPH BASED MODEL AND RAKE.

**Lima Subramanian[1] and R.S Karthik [2] MCA MPHIL.**

1. Research Scholar CMS College of Science and Commerce Coimbatore.
2. Assistant Professor CMS College of Science and Commerce Coimbatore.

***Manuscript Info***

*Manuscript History*

Received: 06 January 2017
Final Accepted: 10 February 2017
Published: March 2017

***Key words:-***
Keyword Extraction, RAKE, Text Rank, HITS, PageRank, Positional Power function.

## *Abstract*

In this paper we introduce Rapid Automatic Keyword extraction an unsupervised, domain independent and language independent method for extracting keywords from individual documents and compare this model with a graph based ranking algorithm(TextRank).In general TextRank consist of two unsupervised methods for both keyword and sentence extraction. Also we conduct a simple study regarding TextRank with the previously published methods

## Introduction:-

Keyword extraction (KE) is termed as the process that automatically identifies a set of elements that best matches with the subject of document. To represent most relevant information contained in the document: key phrases, key segments, key terms or just keywords different approaches were used. Keyword extraction connects different areas of text mining, information retrieval and natural language processing. The process of keyword extraction can be classified into two categories; quantitative and qualitative.

Quantitative techniques are based on statistical relations in addition to formal linguistic processing. Methods like frequency,TF*IDF, co-occurrence are used as word statistics. The basic idea behind this approach is that important terms are most often referenced within the text. But sometimes the most frequent terms is not enough to represent a meaningful keyword. Certainly further modifications are needed for this approach.

Qualitative methods are based on sematic relations and analysis. Semantic analysis relies on semantic description of lexical terms. Such kind of extraction provides highly structured conceptual relations to the content of the text. This method is more reliable than quantitative approach because sometimes it is not compulsory to appear the most important keyword as a frequent item.

This paper presents a new extraction model RAKE operates on individual documents as well as multiple types of documents. The method start with the selection of candidate keywords and a graph of word co-occurrences is constructed. After that a score is calculated for each candidate keyword. Rather than that we include a discussion on an unsupervised keyword extraction.

**Corresponding Author:- Lima Subramanian.**
Address:- Research Scholar CMS College of Science and Commerce Coimbatore .

## Extraction Methods:-

Keyword assignment tasks can be divided into two subclasses: (1) Keyword assignment and (2) keyword extraction, both will finally results a set of keywords. In keyword assignment, keywords are chosen from a predefined class of terms, keyword extraction searches a document with keywords that are explicitly contained in the document.

According to Zahang, automatic keyword extraction methods can be classified into four.

1. Simple statistical approaches
2. Linguistics approaches
3. Machine learning approaches
4. Other approaches

**Simple Statistics Approaches:-**

Which is the simplest method of keyword extraction, which do not require training data. In addition such methods are language and domain independent. The disadvantage is that in information regarding health and medical, the most important keyword may appear only once.

**Linguistic Approaches:-**

In such types of methods which pay attention to linguistic features such as parts of speech, syntactic structure and semantic qualities tend to add value, Functioning sometimes are filters for bad words.

**Machine Learning Approaches:-**

This type of keyword extraction can be seen as supervised learning. In machine learning approach the keywords are extracted from training documents. This approach includes Naïve Bayes, Support vector machines.

**Other Approaches:-**

Other approaches combines the above mentioned methods with some heuristic knowledge such as position length, etc…

**Textrank:-**

TextRank is a graph based ranking model that can be used in text processing for the fruitfulness of natural language applications. It consist of 2 unsupervised methods for keyword and sentence extraction.

1. Graph based – Kleinberg’s HITS Algorithm
2. Ranking Algorithm Google’s PageRank

Let G = (V, E) be a directed graph with a set of V and set of edges E, where E is a subset of V x V. ln (Vi) be the set of vertices that point to it and Out (Vi) be the set of vertices that Vi points to. The score of a vertex is,

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

d is a damping factor and its value within the range O and 1. During the time of Web surfing, TextRank implements the “random surfer model”. When a user clicks on a link with a probability ‘d’ and the probability of jumping to a new page is 1-d. The value of d usually is 0.85. Starting from any arbitrary node, the computation iterates until convergence below a given threshold is achieved.

**Graph Representation of text:-**

Depending on the context, text units of various sizes / characteristics can be added as vertices. The important steps followed in the ranking algorithm can be as follows.

1. Identify text units, and add them as vertices in the graph.
2. Identify relationships between these text units and by using these relationships draw the edges
3. Iterate until convergence.
4. Score vertices based on their final score.

Text Rank Keyword Extraction is fully unsupervised. First the text is tokenized and decorated with parts of speech tags. We consider only single words as candidates. Now an edge is added those lexical units that co-occur within a window of n words. After the graph constructed, the initial score of all vertices are set to an initial value of 1, and the above ranking algorithm is applied for several iterations until it converges usually 20-30 iterations and a threshold of 0.0001.Once the final score is obtained for each vertex in the, vertices are sorted in reverse order of their score, and the top T vertices in the ranking are retained for post processing. T can be set to any fixed values usually ranging from 5 to 20 keywords.

## Rake:-

The input parameter of RAKE are a stop list, set of phrase delimiters, set of word delimiters. By using phrase and word delimiters the doc. Is partitioned into candidate keywords. To identify co-occurrences of words within this candidate keywords, no arbitrarily sized sliding window is needed.

E.g.: Compatibility of systems of linear constraints over the set of natural numbers.

Manually assigned keywords:
Linear constraints, set of natural numbers

Candidate keywords parsed:
Compatibility – systems - linear constraints – set - natural number

After identifying every candidate keyword a graph of word co-occurrences is build and a score is evaluated for each candidate keyword. For this several metrics were used. (1)Word frequency, freq(w),(2) word degree (deg (w)) (3) ratio of degree to frequency (deg (w) / freq (w))

After candidate keywords are scored, the top T scoring candidates are selected. We assume T as one-third the number of words in the graph as in Mihalcea and Tarau.

RAKE does not require a training set. TextRank applies syntactic fillers to a document text to identify content words and accumulates a graph of word co-occurrences in a window size of 2. A rank for each word in the graph is calculated through a series of iterations until convergence below a threshold is achieved.

TextRank damping factor 0.85.
Threshold 0.0001
Not have access to the syntactic filters

**Advantages:-**

- Total time needed for TextRank to extract keywords compared to RAKE is over to times the time of RAKE.
- RAKE If the number of content words increases, the performance also increases because it can score keywords in a single pass whereas TextRank requires repeated iterations to achieve convergence on word ranks.

## Related Works:-

According to Rada Mihalcea Dept of CS University of North Texas[1] ;their paper presents unsupervised method for automatic sentence extraction using graph based ranking algorithm in the context of text summarization task. They discussed a set of graph based Ranking algorithms.

HITS (Hyperlinked Induced Topical Search) Kileinberg 1999
Iterative algorithm for ranking web pages according to their degree of authority

**Positional power function:-**

Introduced by (Herings et al 2001) that determines the score of a vertex as a fun that combines both the number of its successors and the score of its successors.

1. Page Rank used for Web link analysis
2. Weighted graphs

Content of web surfing and citation analysis. TextRank model the graphs are build from natural language texts and may include multiple / partial links between the units that are extracted from text. Consequently, we introduce new formulae for graph based ranking that take into account edge weights when computing the scores associated with the vertex.

TextRank sentence extraction algorithm is evaluated in the context of a single document summarization task using 567 news articles provided during the document understanding evaluations 2002.

Martic Dostal and Karel Jezek [2] conducted a study on automatic keyword extraction based on NLP and statistical methods. Key phrase candidates are extracted by a combination of graph methods (TextRank) and statistical methods (TF * IDF). During the text preprocessing phase, the content is partitioned into meaningful tokens and non-significant characters are removed. The remaining tokens are considered as candidates keywords. Then calculate the TF * IDF score only for these candidate keywords.

Feifan Liu, Deana Pennel, Fei Liu and Yang Liu [3] approaches for automatic keyword extraction using meeting transcripts. The paper suggest that the important part of keyword extraction is to assign a score to a word depending on its importance. Also the article compare different methods for weight calculation the TF IDF framework and the graph model.

**TF – IDF FRAMEWORK:-**

**Basic TF IDF weighting:-**

The TF (term frequency) for a word w, in a doc is the no of times the word occurs. The IDF value is IDF I z log (N / Ni) where Ni denotes the no of documents containing word wi, and N is the total number of documents.

**Parts of speech filtering:-**

In addition to using a stop word list to remove words from consideration we also leverage POS information to filter unlikely keywords. Only verbs, nouns and adjectives are likely to be keywords.

**Integrating word clustering:-**

By using SRILM toolkit words with the same semantic meaning can be identified.

Guangyi Li, Houfeng Wang [4] suggest an improved automatic keyword extraction based on textrank using domain knowledge. They focused keyword extraction for Chinese scientific articles, they used a framework for selecting candidate keywords by Document Frequency Accessor Variety (DF AV) and a TextRank algorithm to improve the performance of keyword extraction, they considered keywords for a specific domain.

Automatic keyword extraction from documents using conditional random fields. (Journal of Computational Information Systems 2008) [5] is another work published on a journal. In this paper keyword extraction based on a CRF model. Conditional Random Fields (CRF) is a state of the art sequence labeling method also their implementation results guarantee that the CRF model is more better than support vector machine, multiple linear regression model etc.

## Conclusion:-

In this paper we have proposed and compared two keyword extraction methods: TextRank and RAKE. RAKE gathers high performance when compared with existing keyword extraction techniques. RAKE automatically extracts keywords in a single pass. Also RAKE is high simple and efficient. But TextRank can be used for large number of natural language applications. TextRank consists of two unsupervised method for keyword and sentence extraction. Advantages of TextRank is that it doesn’t demand linguistic and domain knowledge.

## References:-

1. S. Agrawal, S. Chaudhuri, and G. Das. DBXplorer: A System for Keyword Based Search over Relational Databases. In ICDE, 2002
2. G. Bhalotia, A. Hulgeri, C. Nakhey, S. Chakrabarti, and S. Sudarshan. Keyword searching and browsing in databases using BANKS. In ICDE, 2002
3. P. Bouquet, H. Stoermer, D. Cordioli, G. Tummarello. An entity name system for linking semantic web data. 2008
4. J. D. Cohen. Language and domain-independent automatic indexing terms for abstracting. Journal of the American Society for Information Science, 1995
5. Yu Cong, H.V. Jagadish. Querying complex structured databases. In VLDB, 2007

6. Daniela Florescu, Donald Kossmann, Ioana Manolescu. Integrating keyword search into XML query processing. In WWW9, 2000
7. Michael J. Giarlo. A comparative analysis of keyword extraction techniques. Rutgers, The State University of New Jersey
8. Otis Gospodnetić, Erik Hatcher. Lucene in action, 2005 9. L. Guo, F. Shao, C.Botev, J. Shanmugandaram. XRANK: ranked keyword search over XML documents. SIGMOD, 2003 Hbase http://hadoop.apache.org/hbase/
9. V. Hristidis and Y. Papakonstantinou. DISCOVER: Keyword search in relational databases. In VLDB, 2002
10. V. Hristidis, Y. Papakonstantinou, and A. Balmin. Keyword proximity search on XML graphs. In ICDE, 2003
11. A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of the 2003 Conference on Emprical Methods in Natural Language Processing, Sapporo, Japan, 2003
12. Internet Movie Database http://www.imdb.com/
13. J. B. Keith Humphreys. Phraserate: An HTML keyphrase extractor. Technical Report. 2002
14. Y. Matsuo, M. Ishizuka. Keyword extraction from a single document using word co-ocuurrence statistical information. International Journal on Artificial Intelligence Tools, 2004
15. David Milne, Ian H. Witten. Learning to link with Wikipedia. In CIKM, 2008
16. David Milne, Olena Medelyan, Ian H. Witten. Mining domain-specific thesauri from Wikipedia : A case study. In WI, 2006
17. Okkam Project http://www.okkam.org/
18. T.Palpanas, J. Chaudhry, P. Andritsos, Y. Velegrakis. Entity Management in OKKAM. 2008
19. L. Plas, V.Pallotta, M.Rajman, H.Ghorbel. Automatic keyword extraction from spoken text. A comparison of two lexical resources: the EDR and WordNet. Proceedings of the 4th International Language Resources and Evaluation, European Language Resource Association, 2004
20. Peter Schonhofen. Identifying document topics using the Wikipedia category network.
21. In WI, 2006
22. Y. Suzuki, F. Fukumoto, Y. Sekiguchi. Keyword extraction of radio news using term weighting with an encyclopedia and newspaper articles. SIGIR, 1998.
23. Sandeep Tata, Guy M. Lohman. SQAK: Doing more with keywords. SIGMOD, June 9–12, 2008
24. Wikipedia http://www.wikipedia.org/
25. I. Witten, G. Paynte, E. Frank, C. Gutwin, C. Nevill-Manning. KEA: practical automatic keyphrase extraction. In Proceedings of the 4th ACM Conference on
26. Digital Library, 1999
27. Fei Wu, Raphael Hoffmann, Daniel S. Weld. Information extraction from Wikipedia: moving down the long tail. In KDD’08, 2008
28. Chengzhi Zhang, Huilin Wang, Yao Liu, Dan Wu, Yi Liao, Bo Wang. Automatic Keyword Extraction from Documents Using Conditional Random Fields. Journal of Computational Information Systems, 2008
29. Xuan Zhou, Gideon Zenz, Elena Demidova, Wolfgang Nejdl. SUITS: structuring user’s intent in search. In EDBT, 2009