International Journal of Engineering Applied Sciences and Technology, 2017
Vol. 2, Issue 4, ISSN No. 2455-2143, Pages 215-220
Published Online February-March 2017 in IJEAST (http://www.ijeast.com)

KEYWORD EXTRACTION: A REVIEW

Er. Tanya Gupta
Assistant Professor(CSE)
Gurukul Vidyapeeth
Punjab, INDIA

Abstract :- **In this paper different techniques for keyword extraction is presented. Keyword extraction is very useful as it help us to quickly find out the relevant text from a large amount of data. It is used in computer science field basically in information retrieval and can be used in text summarization, indexing etc. This paper also presents the keyword extraction used as to extract the keywords from social networking site such as twitter. Different parameters for the evaluation could be considered such as precision, entropy etc. Further literature of different keyword extraction is being done.In addition applications related to keyword are also discussed.**

Index terms— Keyword, indexing, summarization, twitter, precision, entropy etc.

I. INTRODUCTION TO NLP

Natural language processing is a language that is associated with communication between computer and human beings. Different challenges in NLP are to understand natural language, making computer efficient to provide meaning to human input language. NLP is a branch of artificial intelligence. Natural language processing is one of the field which is interesting as well as difficult to represent and develop. The main reasons that arise why we study natural language processing is:-

- When we want our system to communicate with the users in the way they want; in that case instead of forcing users to learn different languages the system is trained to learn that language and thus the concept of NLP takes place.
- There is a large amount of information which is recorded in many different natural language. The information could be produced in journals, books, research papers, reports etc which could be found online anywhere and at anytime. Therefore a computer that require a large amount of information must manipulate different natural languages which use information which is available on systems.
- The problems that come in artificial intelligence theories are being processed in natural language and thus NLP is used.

BASICS OF NLP:-

Token:- Before the text is processed, it must be broken into units such as words, number, punctuations etc.

Sentence:-Combination of token is called sentence.

Tokenization:-method of breaking a sentence into different tokens.

Corpus:-It is a structure that contains large amount of sentences.

POS Tag:-In a sentence word can be classified as noun, adjectives, verb, and articles.

II. KEYWORD EXTRACTION

Keyword extraction which is defined as the process of extracting the relevant information from a large amount of data.
Keywords extraction is one of the important task in many fields such as text classification, text clustering, tracking, topic detection, summarization and so on.
Social media analysis plays an important role in keywords extraction which is defined as the process of collecting information from social websites such as twitter, facebook and then to analyze that information to make final decisions.
Twitter is one the important micro blogging online social networking site that allow its user to share, broadcast information. The length of the twitter generated data is 140 characters which is known as tweets.
Therefore sometimes a large amount of data is received such as one is twitter which becomes quietly difficult for the user to read .so in this case it becomes important to extract the information that is useful. Thus the concept of keywords extraction takes place.

International Journal of Engineering Applied Sciences and Technology, 2017
Vol. 2, Issue 4, ISSN No. 2455-2143, Pages 215-220
Published Online February-March 2017 in IJEAST (http://www.ijeast.com)

III. METHODS OF KEYWORD EXTRACTION

SUPERVISED APPROACH:-

In supervised approach, keyphrase extraction algorithm and genex is used where is KEA is defined as the extracting (KEA) keyphrases from text document.KEA consists of two steps :-training and extraction .
TRAINING:-Make model to find out keyphrases from the document which is trained.
EXTRACTION:-Whereas in a extraction the keywords are choosed from a document which is not trained i.e. new document.
GenEx is a algorithm which automatically defined the key words from the documents. Uses quinlam's C 4.5 decision tree induction algorithm.

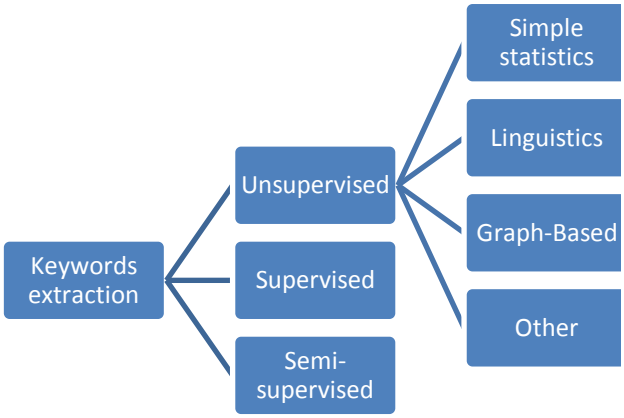

Fig 1:- Different Keyword Extraction method

UNSUPERVISED APPROACH:-

The unsupervised approach consists of:-
Simple Statistics Approaches:-This method is very easy and does not require train data. The words statistics can be used to define the keywords:-n gram statistics, word frequency, TFIDF, word co-occurrence and so on.
Linguistics Approaches:-This approach uses linguistics of sentences and document. Some of the linguistics are syntactic, lexical and so on.
Machine Learning Approaches:-This approach require the trained data to abstract keywords. Some of the approach include support vector machine naïve bayes and so on.

Other approaches:- This approach combines all the above discussed approach and uses some other heuristic knowledge such as length, tags, formatting etc to extract the keywords.

IV. KEYWORD EXTRACTION COMPONENTS:-

Candidate selection:-All the words, phrases are extracted that could possibly be considered as keywords.

Properties Calculation:-Based on the candidate keyword selected, the property is calculated that defines whether it could be a keyword.

Scoring and selecting keywords:-The candidates are scored by the formula or by using some technique.

V. OTHER APPROACH OF KEYWORD EXTRACTION:-

TFIDF (Term frequency inverse document function):

- This is a technique which is used for automatic extraction of keywords.
- used in information retrieval.
- Provides a weight to the words based on frequency.
- The tf part defines the most frequent occurrence of the word and idf provides with frequency to the word.Thus the word with highest frequency is considered first.
  TFIDF=TF*IDF

MAUI (Multipurpose automatic topic indexing)

- Automatically identifies main topic in text documents.
- Performs following tasks:- term assignment with controlled vocabulary, subject indexing automatic tagging.

VI. APPLICATIONS OF KEYWORD EXTRACTION

- **Indexing**
  Indexing is a term used in information retrieval which defines the topic of a document. Index terms make a vocabulary and can be used to extract the keywords from document. Index terms can be words, numeric or phrase.

- **Topic detection**

International Journal of Engineering Applied Sciences and Technology, 2017
Vol. 2, Issue 4, ISSN No. 2455-2143, Pages 215-220
Published Online February-March 2017 in IJEAST (http://www.ijeast.com)

Topic detection can be used in keyword extraction as it can help to automatically identify the main topic in a document

- **Summarization**
  It is a process which reduces the text to create a summary which defines the main important points of a documents.This could be used in keyword extraction to summarize the large of text into small by considering the relevant text.

- **Tracking**
  Keyword extraction could be used in tracking by finding out the relevant information and discarding the irrelevant information.

- **Text classification**
  It is a task which assigns a predefined categories to the text. For example paper could be classified as technical and non-technical. Another example of religion which could be hindu ,muslim, sikh etc.

VII. LITERATURE SURVEY

Zhao et al.[1] They extracted the keyword from the Chinese micro blog and to extract that keywords we use three features graph model, semantic space and location of words. In which they used the methodology In the first step we need to download the micro blog API of a user. In the second step we do pre processing by data cleaning, word segment, POS tagging and stop word removal. In the third step we create a graph model to extract the keywords on the basis of co occurrence between the words and we give a sequence number to the words according to the location and find a weight of the words by Score formula. In the fourth step we create a semantic space on the basis of topic detection and compute the statistical weight by TFIDF. In the fifth step we consider the another feature that is the location of words and compute the rank value in which we conclude that a number having smaller location will be ranked higher.

Hromic et al.[2] This paper focuses on the structure approach and graph generation.The approach used in this paper is structure based in which we create graph model and we identify the bursty topics and events.In the clustering of topic the tweets of twitter are separated to produce two graph i.e. homogeneous graph and heterogeneous graph.For homogeneous graph we use OSLOM algo to find the users interaction.For heterogeneous rank we use rankclus algo to construct a set of tweets ranked with number. At last from both the graph results,the meaning to tweet is done using python and then we join the tweets with the same name. In future different graph models can be used for different types of events and to construct a method that can define the events.

Marujo et al.[3] They constructed a method for keyword extraction and to find the solution for the problem such as high variance and lexical variants and to compare the current technique with already existing technique. For the problem of lexical variant i.e the words although speeling different have same meaning but the technique which we use does not know about it.so to extract that word as keyword we used two methods Brown clustering:-in this we cluster the words with the same meaning such as ‘no’ ‘noo’ etc and then find the feature for the individual cluster.

Kim1 et al.[4] In this paper we implement a system that detect the popular keyword and the bursty keyword in which it detect the abbreviations, any typing or spacing error. The first step used in detecting trend and bursty keywords is to collect the candidate keywords i.e. the first word starting with the capital letter or the word enclosed in quotation mark is considered as candidate keyword.The second step is to merge the keywords and to merge that we consider acronyms,typo,and spacing and then we find the term frequency accordingly.The next step is to detect and select the popular keywords from the candidate keywords that were merged and to select the bursty keywords we use burst ratio. A prototype system can be build that can detect more bursty keywords.

Torres et al.[5] This paper proposes a technique called TOPOL which identifies the irrelevant noisy data from the useful data.The first step used is pre processing step in which elimination of hastags,URL,non textual symbols from the tweet is done.Second step consist of mapping in which a matrix is generated by applying SVD technique.Third and the last step used is the topic detection step in which the topic are selected based on the interest.Finally the results are computed based on the parameters such as topic recall,keyword precision,keyword recall.In future many other algorithms and techniques can be used for detecting the bursty topics.

Beliga et al.[6] This paper presents different methods and technique and approaches used in the

keyword extraction.Further it defines graph based method which are based on the extraction of nodes.This paper also discusses the extraction for the Croatian language.selectivity based keyword extraction method is used in which in future we can consider different length text,different languages based on different dataset,new techniques for evaluation,to find whether entities are extracted and in text summarization.

Beliga et al.[7] This paper presents different methods and approaches for keyword extraction.. paper also focuses on the graph types in which vertex and edge representation is considered..further selectivity based keyword extraction is used in which text is represented in the form of vertex and edges. The result is computed on the in degree, out degree, closeness ,selectivity.
In future the graph based method will attract the research community.

Abilhoa et al.[8] This paper proposes a keyword extraction method that represents the graph for the text and apply the centrality measure and find the relevant vertices.This paper proposes a technique called TKG(Twitter keyword graph) in which three steps are performed.The first is preprocessing in which stop words etc are removed.In the second step graph is represented in which nearest neighbor and all neighbors are considered.The results are computed based on the precision,recall,F-measure as well as scalability is also accessed.the future scope can be to use other centrality measure and defining more structure for graph using heuristics and elimination of noisy data.

Bennya et al.[9]This paper summarizes the data based on particular keyword.Two algorithms are used TDA(topic detection using AGF)and TCTR(topic clustering and tweet retrieval).The methodology used is first to extract tweets from twitter,then tfidf is applied which gives weight to the words along with the frequency.AGF is evaluated using keyword rating and concept for the imitation of the mental ability of word association.The results are calculated based on the class entropy,purity,cluster entropy.The future of this paper is to consider the sentiments and emotions of tweets and the number of retweets will also be taken.

Chen et al.[10] In this paper a technique is proposed in which user can search using search engine but without entering any keyword. In this a log is made in which user behavior and repository is saved. but in this paper a Google similarity distance is used to find the keywords. So the need for repository is abolished and everything is done online and real time. When the user search using a search engine a lot of information is provided but to extract the relevant and accurate information which the user need there is methods used such as keyword expansion and keyword extraction.

Lee et al.[11] A facility called as high relevance keyword extraction (HRKE) has been produced in Bayesian text classification to extract the keywords in the stage of classification without the use of pre classification process. the facility uses a posterior probability value to extract the keywords. The HRKE uses bayseian classification approach in which the first step is to extract words from the text which contains a list of words and this list is constructed by assuming the length of text to be n. and the the posterior probability is calculated. and then the TFIDF method is used which assign the weight to the words.

Gimpat et al.[12] In this paper POS problems are addressed from the twitter microblog.The results are obtained with 90 percent accuracy.The tools and techniques are used for the richer text of twitter.First we develop a pos tagset,manually tagging is done,features for the POS tagger is developed and the experiments are conducted and finally the annotated dataset is provided to research community.The hastags,URL and emotions are considered and must be occurred at the end.The system uses traditional tag dicitionary,distributional similarity,traditional tag dicitionary. Thus concluded that approach can be appleid to linguistic analysis as they arise in social media.also the annotated data can be used in semi supervised learning.

Carpena et al.[13] This paper solves the problem of statistical keyword extraction from text by introducing two approaches that is entropic and clustering approach.This paper we implement some changes to both the approaches and then find the results.and then the new approach is proposed which detect keywords based on the need of user.finally both the approaches work well for the long text.then we focus on the short text such as web pages,articles etc and we concluded that the clustering results better than the entropic. The main goal is to find and rank the word which is important in the text.The two approaches are used for the short text an the short text with glossary and generic short text is considered.the results were evaluated and the word clustering proves to be better both for the long as well as short text.while the entropic suits well for

long text and does not perform well for the text which is partitioned.

Yang et al.14] In this paper a metric is proposed which give rank to the words. This method uses the Shannon's entropy which is the difference between intrinsic and extrinsic mode which means the words define the author purpose and the words which are irrelevant are present random in the text. This method is well suited for single document of which no information is known in advance.The idea of intrinsic and extrinsic is that the words which are meaningful are grouped together.The words are extracted and ranked according to the entropy difference.Mean,mode and median is calculated based on entropy difference and the ED metric proved to be a good in ranking the words.In future this type of work can be used in many natural language processing in which words are defined without any knowledge of syntax.

Sun et al.[15] Twitter online social site which has a million of users which provide information daily.In this paper a technique for tweet segmentation is proposed called HybridSeg. In this paper first tweets are splitted into segments in this way meaning of the information is easily extracted. The HybridSeg finds the sum by segmenting the tweets. Two context are considered local and global. The highest accuracy is achieved by part of speech tagging. The segment based entity is better than word based entity. The future of this paper is to improve the quality by using more features. and another is to present task in tweet summarization, search etc.

Ventura et al.[16] The extraction of single and multi words expression using statistical language independent approach. Although different methods for single and multi words are present such as unigram extractor for single word and multi word extractor for multiple words.But in this paper an approach called conceptextractor approach is proposed which extract single and multi words from text.The results were evaluated based on three languages.Precision and recall were two parameters for the evaluation of results.This paper also defines a metric for the specificity of both single and multi word which can be used to test the other languages.In future approximation be made for n-gram and unigram concepts.

Hong et al.[17] Various Chinese keyword extraction method are considered. In this paper an extended term frequency method is defined which consider Chinese characteristics with TF method. This method also establishes model for classification support vector machine.The precision and recall rate improved much better.The TF method is used which is based on term frequency.The four improvement strategy were considered as grammar model in which it included noun, modifier, noun phrase and verb phrase.

Wong et al.[18]In this paper the discovery of terms that are "title like" are classified in document.the main idea is that the terms that are title and title like should behave same in document. The behavior can be found using distributional and linguistic features. The classifier is trained to find the behavior of terms. The recall rate of finding the title terms were high but precision was low because some of the words which were not title were also identified in title terms. The rating was calculated on three basis topical, thematic and title term. different features were considered such as location, frequency, document size etc after this the evaluation was done based on recall and precision.The main is to identify the title like term from title so the result from recall was much better than precision.

Usui et al.[19]Many researches are related to brain and have been used widely in the world. The resources should be used efficiently and effectively. Many neuroinformatics sites and centres are used which exchange the information and the resources related to brain among different countries. Many plateforms are developed which have their keywords that represents the main terms. The main advantage of keywords to be defined previously is that they help to classify the main text and the resources. So a tool is defined which automatically finds the most important keywords.

VIII. CONCLUSION AND FUTURE SCOPE

From the above discussed paper we can conclude that different methods and techniques could be used to extract keywords.Different approaches such as supervised and unsupervised are used in different scenarios.By keyword extraction we could easily find the words that are highly signified and of great importance. Some techniques resulted in a better precision rate while some has less results for this parameter. Our proposal is that the TFIDF approach could provide a better result and can be used effectively.

Further in the future other parameters such as purity, entropy could be considered of great importance and sentiments can be used to extract the keywords from the twitter. More in future the sentiments and

emotions of the tweet can also be considered as well as some heuristic approach can be used that eliminate the noisy data which could further reduce the computational time of the algorithm.

IX. REFERENCES

1.Zhao, H., & Zeng, Q. (2013). Micro-blog keyword extraction method based on graph model and semantic space. *Journal of Multimedia*, *8*(5), 611-617.

2.Hromic, H., Prangnawarat, N., Hulpuș, I., Karnstedt, M., & Hayes, C. (2015). Graph-based methods for clustering topics of interest in twitter. In*Engineering the Web in the Big Data Era* (pp. 701-704). Springer International Publishing.

3.Marujo, L., Ling, W., Trancoso, I., Dyer, C., Black, A. W., Gershman, A., ... & Carbonell, J. Automatic Keyword Extraction on Twitter. *Volume 2: Short Papers*, 637.

4. Kim, D., Kim, D., Rho, S., & Hwang, E. J. (2013). Detecting trend and bursty keywords using characteristics of Twitter stream data. *International Journal of Smart Home*, *7*(1), 209-220.

5. Torres-Tramón, P., Hromic, H., & Heravi, B. R. (2015). Topic Detection in Twitter Using Topology Data Analysis. In *Current Trends in Web Engineering*(pp. 186-197). Springer International Publishing.

6. Beliga, S. (2014). Keyword extraction: a review of methods and approaches.*University of Rijeka, Department of Informatics, Rijeka*.

7.Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2015). An Overview of Graph-Based Keyword Extraction Methods and Approaches. *Journal of Information and Organizational Sciences*, *39*(1), 1-20.

8.Abilhoa, W. D., & de Castro, L. N. (2014). A keyword extraction method from twitter messages represented as graphs. *Applied Mathematics and Computation*, *240*, 308-325.

9. Benny, A., & Philip, M. (2015). Keyword Based Tweet Extraction and Detection of Related Topics. *Procedia Computer Science*, *46*, 364-371.

10.Chen, P. I., & Lin, S. J. (2010). Automatic keyword prediction using Google similarity distance. *Expert Systems with Applications*, *37*(3), 1928-1938.

11.Lee, L. H., Isa, D., Choo, W. O., & Chue, W. Y. (2012). High Relevance Keyword Extraction facility for Bayesian text classification on different domains of varying characteristic. *Expert Systems with Applications*, *39*(1), 1147-1155.

12.Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., ... & Smith, N. A. (2011, June). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2* (pp. 42-47). Association for Computational Linguistics.

13.Carretero-Campos, C., Bernaola-Galván, P., Coronado, A. V., & Carpena, P. (2013). Improving statistical keyword detection in short texts: Entropic and clustering approaches. *Physica A: Statistical Mechanics and its Applications*, *392*(6), 1481-1492.

14.Yang, Z., Lei, J., Fan, K., & Lai, Y. (2013). Keyword extraction by entropy difference between the intrinsic and extrinsic mode. *Physica A: Statistical Mechanics and its Applications*, *392*(19), 4523-4531.

15..Li, C., Sun, A., Weng, J., & He, Q. (2015). Tweet segmentation and its application to named entity recognition. *Knowledge and Data Engineering, IEEE Transactions on*, *27*(2), 558-570.

16.Ventura, J., & Silva, J. (2012). Mining concepts from texts. *Procedia Computer Science*, *9*, 27-36.

17.Hong, B., & Zhen, D. (2012). An extended keyword extraction method.*Physics Procedia*, *24*, 1120-1127.

18.Wong, C. W., Luk, R. W., & Ho, E. K. (2005). Discovering "title-like" terms.*Information processing & management*, *41*(4), 789-800.

19.Usui, S., Palmes, P., Nagata, K., Taniguchi, T., & Ueda, N. (2007). Keyword extraction, ranking, and organization for the neuroinformatics platform.*Biosystems*, *88*(3), 334-342.

20..https://www.airpair.com/nlp/keyword-extraction-tutorial.

21.Y. Matsuo andM. Ishizuka(2004). Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools,13:200

International Journal of Engineering Applied Sciences and Technology, 2017
Vol. 2, Issue 4, ISSN No. 2455-2143, Pages 215-220
Published Online February-March 2017 in IJEAST (http://www.ijeast.com)