IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 8, October 2014.
www.ijiset.com
ISSN 2348 – 7968

Automatic Keyword Extraction from Dravidian Language

M. Hanumanthappa[1], M Narayana Swamy[2] and N M Jyothi[3]

[1] Dept of Computer Science, Bangalore University
Bangalore Karnataka, India

[2] Dept of Computer Science, Presidency College
Bangalore, India

Abstract

Keywords are significant words in a document that gives description of its content to the reader. They provide the summary of a document. Now a days the amount of electronic text increases rapidly in all the languages. So the text mining applications take the advantage of keywords for processing documents. There are few proposed methods for keyword extraction. But not much work has been done in keyword extraction for Indian languages. With exponential increase in the information in Indian languages on the web, automatic information processing and retrieval become an urgent need. Text Mining is essential for knowledge discovery from valuable texts available in many Indian languages. This paper introduces a method, which extracts the keywords from Dravidian languages of India like Tamil, Telugu, and Kannada. We made an attempt to extract the keywords by using words statistics of a document.

Keywords: *Vector Space Model, TFIDF, Recall, Dravidian Languages.*

1. Introduction

Keyword provides the summary of the document. Identifying keywords from a large amount of electronic text is very useful to produce a short summary of document. As electronic text documents rapidly increases in the size with the growth of internet, keyword extraction from text data is useful for many text mining applications such as automatic indexing, text summarization, information retrieval, classification, clustering, filtering, cataloging, topic detection and tracking, information visualization, report generation , web searches etc

2. Need for Indian Language keywords extraction methods

Text mining is a growing new field that attempts to gather meaningful information from natural language text. It may be loosely characterized as the process of analyzing text to extract information that is useful for particular purposes. Compared with the kind of data stored in databases, text is unstructured, and difficult to deal with algorithmically. In modern culture, text is the most common vehicle for the formal exchange of information.

With over **1.27 billion** people and more than one thousand languages, India is one of the multilingual nations in the world today. It is home to the Indo-Aryan and Dravidian language families, two of the world's largest. Indo-Aryan languages include Hindi-Urdu, Assamese, Bengali, Gujarati, Marathi, Punjabi, Rajasthani, Sindhi, Oriya etc. Dravidian languages include languages like Malayalam, Tamil, Telugu, Kannada etc.

In the Constitution of India, a provision is made for each of the Indian states to choose their own official language for communicating at the state level for official purpose. The availability of constantly increasing amount of textual data of various Indian regional languages in electronic form has accelerated [1].

With exponential increase in the information in Indian languages on the web, automatic information processing and retrieval become an urgent need. Due to this it is difficult for human beings to analyze data in the field of natural language processing in stipulated time. Huge number of available documents in digital media makes it difficult to obtain the necessary information related to the needs of a user. In order to solve this issue, text summarization systems can be used. The text summarization systems extract brief information from a given document while preserving important concepts of that document. By using the summary produced, a user can decide if a document is related to his/her needs without reading the whole document. Also other systems, such as search engines, news portals etc., can use document summaries to perform their jobs more efficiently.

To extract important information or sentences, high quality keyword plays crucial role as per user requirement. They help users to search information more efficiently. Due to growth of online information it is difficult for human beings to accomplish their task in the field of natural language processing in stipulated time. Extracting high quality keywords automatically are expensive and time consuming. This shows keyword extraction is challenging

ISSN 2348 – 7968

problem in the area of natural language processing especially in the context of global languages in acceptable time.

The Dravidian language content on web has also been increasing. There have been many portals, which host large amounts of the Dravidian language content. However, we argue that the data is being underutilized due to the unavailability of Dravidian language text mining methods.

Not much work has been done in Dravidian languages text processing .So the objective of this research work is to make an attempt to design an algorithm to extract keywords from Dravidian languages.

3. Preprocessing of Dravidian Language text documents

Before Keyword extraction, the text document should be preprocessed, that means the unstructured data should be converted to structured data.

In general, the task of text preprocessing can be divided into four stages namely:

- Selecting Candidate Terms
- Filtering
- Vector Space Model
- Ranking

3.1 Candidate Terms

There are some proposed methods for selecting candidate terms from textual Content, including:

Tokenization: The purpose of tokenization is to break down a document into tokens (terms) which are small units of meaningful text.

N-gram approach: This approach selects the candidate terms by means of extracting n-gram (phrase), where n equals one (uni), two (bi), or three (tri)[9].

Linguistically oriented: This method use natural language processing (NLP) methods such as NP-chunker and part-of-speech (PoS) to extract all candidate terms in the documents[8]

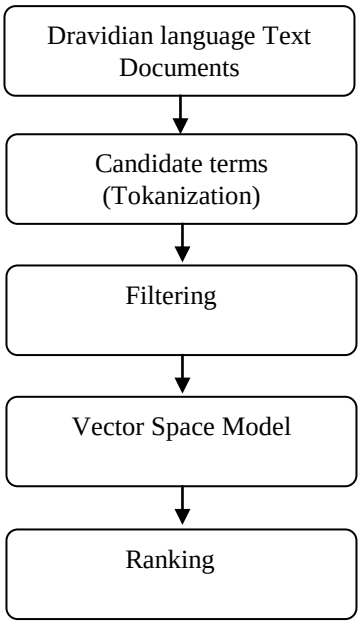

Fig .1 Text Preprocessing steps

3.2 Filtering

Filtering means vocabulary pruning. The purpose of vocabulary pruning is to minimize the candidate sets. The different methods used for vocabulary pruning are lemmatization, stemming, stop word removal. Lemmatization is to reduce all different forms of morphologically related words to their common root or lemma. Stemming is to reduce different forms to a common root or stem. Stop words removal means, predefined set of common terms are removed from the vocabulary.

3.3 Vector space model

One of the simplest ways to model texts is Vector Space Model [2].Vector space model or term vector model is an algebraic model for representing text documents as vectors. The principle behind the VSM is that a vector, with elements representing individual terms, may encode a document's meaning according to the relative weights of these term elements. Then one may encode a corpus of documents as a term-by-document matrix X of column vectors such that the rows represent terms and the columns represent documents. Each element x_{ij} tabulates the number of times term I occurs in document j. This matrix is sparse due to the Zipfian distribution of terms in a language.

ISSN 2348 – 7968

VSM representation scheme performs well in many text classification tasks. Weights are assigned to each term in a document that depends on the number of occurrences of the term in the document. By assigning a weight for each term in a document, a document may be viewed as a vector of weights.

3.4 Ranking

The ranking tries to maximize the informativeness of a candidate terms by assigning weights to each candidate and then keep the ones that are most relevant.

Statistical methods such as term frequency, document frequency, inverse document frequency, and co-occurrence are simple and practical methods to rank the candidate terms. These methods rank the candidate terms according to their scores. The candidate terms having highest scores are selected as keywords. There are also some ranking method using machine learning approach.

4. Existing Approaches

Keyword extraction methods can be divided into four categories [3]

- Simple statistics,
- Linguistics,
- Machine learning
- Mixed approaches

4.1 Simple Statistics Approaches

These methods are simple, have limited requirements and don't need the training data. They tend to focus on non-linguistic features of the text such as term frequency, inverse document frequency, and position of a keyword. The statistics information of the words can be used to identify the keywords in the document. Cohen uses N-Gram statistical information to automatic index the document [9]. Other statistics methods include word frequency, TF*IDF, word co-occurrences [4] etc. The benefits of purely statistical methods are their ease of use and the fact that they do generally produce good results.

4.2 Linguistics Approaches

These approaches use the linguistic features of the words, sentences and document. Methods which pay attention to linguistic features such as part-of-speech, syntactic structure and semantic qualities tend to add value, functioning sometimes as filters for bad keywords. Hulth[8] examines a few different methods of incorporating linguistics into keyword extraction. The candidate terms are considered as keywords based on three features: document frequency (TF), Inverse Document Frequency (IDF)[5], relative position of its first occurrence in a document and the term's part of speech tag. The results indicate that the use of linguistic features signify the remarkable improvement of the automatic keyword extraction. In fact, some of the linguistic methods are mixed methods, combining some linguistic methods with common statistical measures such as term frequency and inverse document frequency.

4.3 Machine Learning Approaches

Keyword extraction can be seen as supervised learning from the examples. The machine learning mechanism works as follows. First a set of training documents is provided to the system, each of which has a range of human-chosen keywords as well. Then they gained knowledge is applied to find keywords from new documents. The Key phrase Extraction Algorithm (KEA) [7] uses the machine learning techniques and naive Bayes formula for domain-based extraction of technical key phrases.

4.4 Mixed Approaches

Other approaches about keyword extraction mainly combine the methods mentioned above or use some heuristic knowledge in the task of keyword extraction, such as the position, length, layout feature of the words, html tags around of the words, etc[6]

4. Related work on Indian languages

Part of speech tagging plays a vital role in natural language processing. It presents a reasonably accurate POS tagger for Kannada language [10]. Oriya Language Text Mining Using C5.0 Algorithm [11]. Odia Text Summarization using Stemmer [12] Algorithm for Punjabi Text Classification.[13]

5. Proposed keyword extraction algorithm for Dravidian Indian languages text documents

The algorithm for keyword extraction:

(1) Dravidian language text document is tokenized
   (2) Stop words and frequent words elimination to get vocabulary words.
   (3) Vocabulary words are stored in the form of matrix called Vector space model
   (4) Term frequency, Inverse document frequency and TF*IDF for each word is calculated
   (5) Select the vocabulary words by fixing threshold value for TF*IDF.

ISSN 2348 – 7968

(6) Along with the keywords the corresponding line number or paragraph number or page number is also extracted

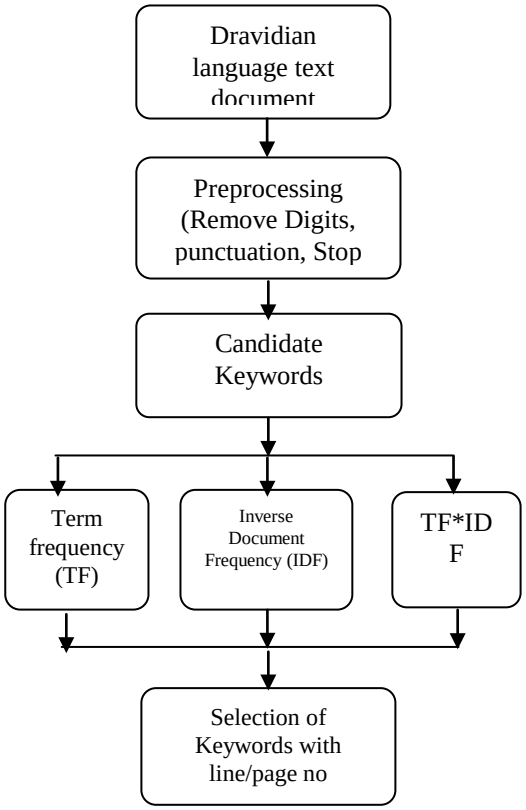

Fig 2. Keywords extraction algorithm

In our work, the term frequency – inverse document frequency also called TF*IDF, is a used to evaluate how important is a word in a document.

6. Experiments and Discussion

Experiments have been conducted in this paper to compare the performance and effectiveness of statistical techniques. On the other hand, we have measured the effect of some factors such as stemming, stop-words removal, and collection size on the performance and effectiveness of statistical techniques.

The first experiment aims at measuring the effect of digits, punctuation and stop-words removal on the effectiveness of statistical techniques on Dravidian language text document.

Table 1 Summary of the documents

Dravidian language	No of Tokens	No of Tokens after removing stop-words	No of Vocabulary
Kannada	531	262	177
Tamil	614	278	174
Telugu	547	262	190

7. Performance Evaluation and Discussion

For key word extraction Precision and recall are evaluation metric. Precision is the proportion of returned keywords that are targets, while recall is the proportion of target keywords returned.

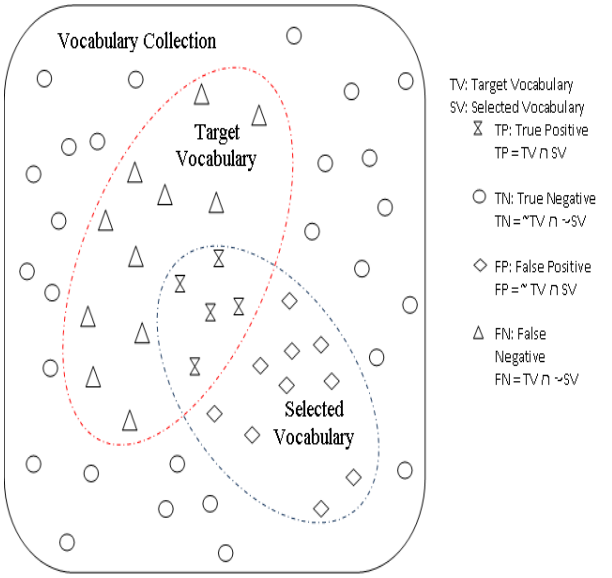

Fig 3 Target category and selected document

Recall = TP / (TP+FN) × 100 %
F-score = 2 × precision × recall / (precision + recall)

In the case of Kannada text (fig 4), when TF*IDF is 3.0910 the recall is 100%. Therefore we are using 3.0910 as a TFIDF threshold value to extract the keywords. So we are considering those words whose TFIDF value is grater then 3.0910 as keywords

ISSN 2348 – 7968

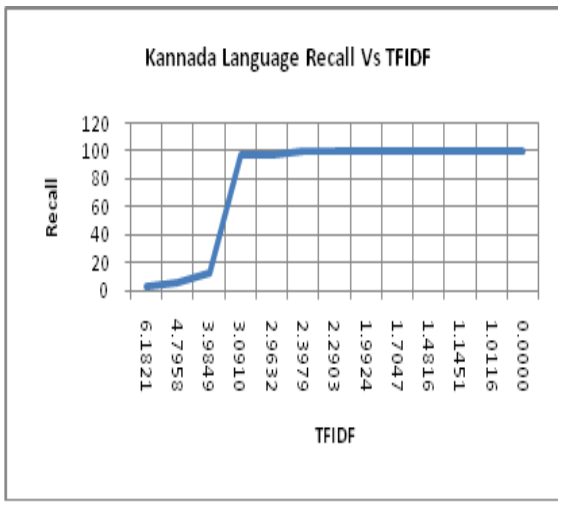

Fig . 4 : Kannada text Recall Vs TF*IDF

In the case of Tamil text(fig 4), when TFIDF is 2.3979 the recall is 100%. Therefore we are using 2.3979 as a TFIDF threshold value to extract the keywords. So we are considering those words whose TFIDF value is grater then 2.3979 as keywords

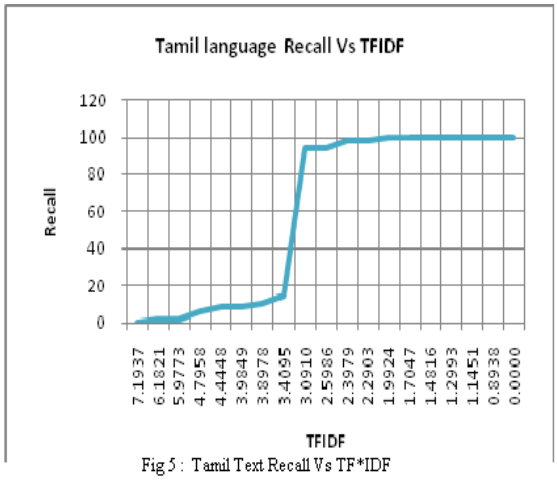

Fig 5 : Tamil Text Recall Vs TF*IDF

In the case of Telugu text,(Fig 6) when TFIDF is 3.4095 the recall is 100%. Therefore we are using 3.4095 as a TFIDF threshold value to extract the keywords. So we are considering those words whose TFIDF value is grater then 3.4095 as keywords

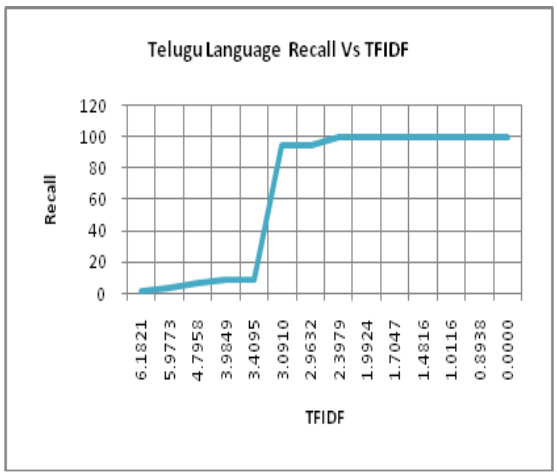

Fig 6 : Telugu Text Recall Vs TF*IDF

8. Conclusion and future work

The Keyword extraction algorithm using TFIDF feature selection is pro-posed for Dravidian languages like kannada, tamil and telugu to extract the keywords.

The importance of the word is based on the value of TF*IDF. The number of keywords selection is based on the threshold value TF*IDF. Precision and recall tend to antagonize each other. This means that efforts to increase precision will generally compromise recall and efforts to increase recall will generally compromise precision.

As we seen in the graph there is a steep rise in the recall. The future development is to improve the result by removing the steep increase to gradual increase by considering the features of the text.

This algorithm shows satisfactory results for all dravidian language. This algorithm is Dravidian language independent. The results can be improved by considering vocabulary pruning methods like lemmatization and stemming. These two are language dependent.

9. References

[1] Narayana Swamy and, Hanumanthappa “Indian Language Text Representation and Categorization using Supervised Learning Algorithm” ICICA '14 Proceedings of the 2014 International Conference on Intelligent Computing Applications, IEEE Computer Society Washington, DC, USA , 2014 ISBN: 978-1-4799-3966-4

ISSN 2348 – 7968

[2] Gerard Salton and Michael J. McGill. “Introduction to Modern Information Retrieval” McGraw-Hill Book Company, New York, 1983.

[3] Michael J. Giarlo. A comparative analysis of keyword extraction techniques. Rutgers,The State University of New Jersey and Chengzhi Zhang, Huilin Wang, Yao Liu, Dan Wu, Yi Liao, Bo Wang.

[4] Y. Matsuo, M. Ishizuka. Keyword extraction from a single document using word co-ocuurrence statistical information. International Journal on Artificial Intelligence Tools, 2004],

[5] Kishore Papineni. Why inverse document frequency? IBM J.T Wattson research center, Yorktown heights, Y N10598, USA,

[6] J. B. Keith Humphreys. Phraserate: An HTML keyphrase extractor. Technical Report. 2002.

[7] I. Witten, G. Paynte, E. Frank, C. Gutwin, C. Nevill-Manning. KEA: practical automatic keyphrase extraction. In Proceedings of the 4th ACM Conference on Digital Library, 1999

[8] 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

[9] J. D. Cohen. Language and domain-independent automatic indexing terms for abstracting.Journal of the American Society for Information Science

[10] “POS Tagger for Kannada Sentence Translation” Mallamma V Reddy, Dr. M. Hanumanthappa, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Volume 1, Issue 1, May-June 2012 ISSN 2278-6856

[11] “Oriya Language Text Mining Using C5.0 Algorithm”, Sohag Sundar Nanda, Soumya Mishra, Sanghamitra Mohanty, Sohag Sundar Nanda et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (1) , 2011, 551-554551 ISSN:0975-9646 :

[12] “Odia Text Summarization using Stemmer” R. C. Balabantaray, B. Sahoo,D. K. Sahoo International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868

[13] “Algorithm for Punjabi Text Classification”, Nidhi Vishal Gupta, International Journal of Computer Applications (0975 – 8887) Volume 37– No.11, January 2012

First Author: Dr. M. Hanumanthappa is currently working as a faculty as well as chairman in the Dept. of Computer Science and Applications, Bangalore University, Bangalore. He has over 16 Years of teaching (Post Graduate) as well as Industry experience. His area of Interest includes mainly Data Mining, Information Retrieval and Programming Languages. Besides, He has conducted a number of training programmes and workshops for Computer Science students. He is also the Principle Investigator of UGC-Major Research Project; he has published nearly 50 Research Papers in National and International Journal and Conferences. Currently he is guiding students for Ph.D in Computer Science, under Bangalore University. He is also one of the member of Board of Studies as well as Board of Examiners for various Universities of Karnataka

Second Author: M.Narayana Swamy, received M Phil from MS University Tamil Nadu, India in 2002 and MCA degree from Madras University, India in 1998. Currently Pursuing PhD in Computer Science and applications, Bharatiyar University, Tamil Nadu, India under the guidance of Dr .M Hanumanthappa. He has over 12 years of teaching experience. He has published nearly 10 Research Papers in National and International Journal/ Conferences. The areas of interest are Data Base Management System, Data Mining, Text Mining