Quest Journals
Journal of Software Engineering and Simulation
Volume 2 ~ Issue 10 (2015) pp: 01-04
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

INDIAN LANGUAGE TEXT MINING

[1]Dr. Hanumanthappa, Narayana Swamy.M[2]
[1]*Professor Department of Computer Applications Bangalore University Bangalore*
[2]*Research Scholar Bharatiyar University Coimbatore Tamil Nadu*

Received 15 April, 2015; **A**ccepted 08 May, 2015 © The author(s) 2015. Published with open access at **www.questjournals.org**

ABSTRACT:- India is the home of different languages, due to its cultural and geographical diversity. 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. In India, the growth in consumption of Indian language content started because of growth of electronic devices and technology. The availability of constantly increasing amount of textual data of various Indian regional languages in electronic form has accelerated. But not much work has been done in Indian languages text processing. So there is a huge gap from the stored data to the knowledge that could be constructed from the data. This transition won't occur automatically, that's where Text mining comes into picture. This research is concerned with the study and analyzes the text mining for Indian regional languages

Text mining refers to such a knowledge discovery process when the source data under consideration is text. Text mining is a new and exciting research area that tries to solve the information overload problem by using techniques from information retrieval, information extraction as well as natural language processing (NLP) and connects them with the algorithms and methods of KDD, data mining, machine learning and statistics. Some applications of text mining are: document classification, information retrieval, clustering documents, information extraction, and performance evaluation. In this paper we made an attempt to show the need of text mining for Indian languages.

Keywords:- Text Mining, Knowledge mining, Preprocessing, Text categorization, Keyword extraction, TFIDF

I. INTRODUCTION

India is one of the multilingual nations in the world today. India's growing focus on Internet services being provided in regional languages. The first step in this direction was the launch of TDIL (Technology Development for Indian Languages) Programme in 1991 by Ministry of Information Technology to develop information processing tools to facilitate human machine interaction in Indian Languages. [1] States that the internet users in India could increase by 24% if local language content is provided on the internet. Amongst current active internet users, local language usage penetration is around 42 per cent. Huge number of available documents in digital media makes it difficult to obtain the necessary knowledge related to the needs of a user. So with exponential increase in the information in Indian languages on the web, automatic information processing and retrieval become an urgent need.

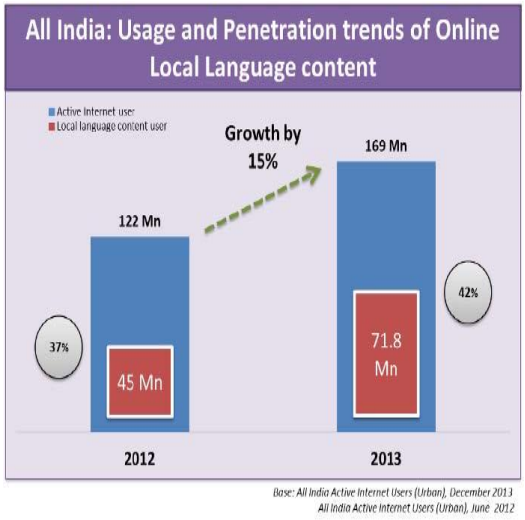
II. LITERATURE REVIEW

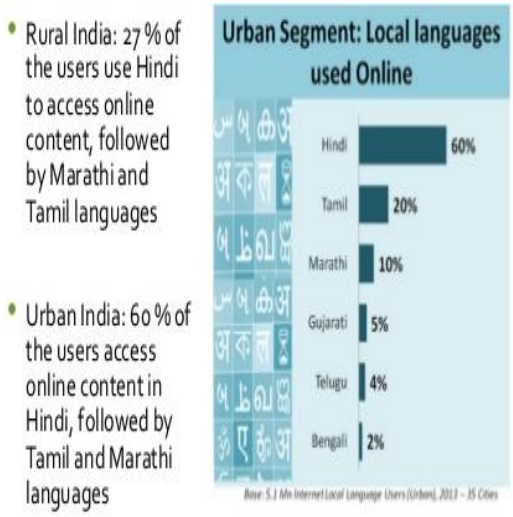
From the literature survey noticed that, not much work has been done in Indian languages text processing. Here an attempt is made to summarize the research work on Indian languages
[2] In this paper, reported work on keyword extraction and topic tracking for Punjabi language.
[3]Part of speech tagging plays a vital role in natural language processing. This paper presents a reasonably accurate POS tagger for Kannada language.
[4]Used Domain Based Ontology for the Classification of Punjabi Text Documents.
[5]They have worked on single- document opinion summarization for Bengali. The novelty of the proposed technique is the topic based document-level theme relational graphical representation. This is the first attempt on opinion summarization for Bengali.

[6] This is first time that these resources have been developed for Punjabi and these can be beneficial for developing other Natural language processing applications for Punjabi.
[7] It classifies a given document and then creates a summary. There is no standard stop word list for Kannada, or methods to do that. Hence a given procedure in this work can be used as a stop word removal method. The summarizer can be used as a tool in various organizations such as Kannada Development Authority, Kannada Sahitya Parishath etc.
[8] Used classification algorithms C5.0 to extract relevant data from Oriya language Oriya Language.
[9] Classified Telugu documents using Naive Bayes classifier. The base system on which a variety of further explorations can be carried out, both from the linguistic point of view and statistical point of view.

III. NEED OF TEXT MINING IN INDIAN LANGUAGES

With over 1.27 billion people and more than one thousand languages, India is one of the multilingual nations in the world today. India's growing focus on Internet services being provided in regional languages. The first step in this direction was the launch of TDIL (Technology Development for Indian Languages) Programme in 1991 by Ministry of Information Technology to develop information processing tools to facilitate human machine interaction in Indian Languages.





The Google launched a Hindi homepage in 2009, and now support Gujarati, Tamil, Marathi and Bengali; and others like Twitter started to support Hindi in 2011. Indiblogger aggregates links to Indian blogging sites, and you can choose between Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, Telugu and Urdu content on the site. [9] States that the internet users in India could increase by 24 per cent if local language content is provided on the internet.

Existing internet users, also long for content in the language of their choice. Amongst current active internet users, local language usage penetration is around 42 per cent. Huge number of available documents in digital media makes it difficult to obtain the necessary knowledge related to the needs of a user.

So with exponential increase in the information in Indian languages on the web, automatic information processing and retrieval become an urgent need. This motivated us to work on Text mining for Indian languages

IV. OBJECTIVES

India is multilingual nations. T*ext mining* is a growing *research area* in data mining. The most fundamental property of languages is the one known as Zipf's law. For any Indian language, if we plot the frequency of words versus their rank for a sufficiently large collection of textual data, we will see a clear trend, which resembles a power law distribution. So the aim is conduct a detailed study on text mining on Indian language. In this paper we have proposed techniques for Indian language text mining

1. To design a method for Indian language documents representation
2. To propose an algorithm to categorize documents based on language/domain.
3. To design a common algorithm to extract the information using keyword extraction methods from all the Indian language documents.

V. PROPOSED TECHNIQUES FOR INDIAN LANGUAGE TEXT MINING

In the proposed methods of Indian language text mining there are three phases.

a. Data preprocessing
b. Document categorization
c. Keywords extraction

a. Data Preprocessing

Data Preprocessing means converting unstructured data into structured data. Given a textual source containing different types of documents (different formats, language formatting) the first action that should text preprocessing. After preprocessing data mining algorithms can be applied .The different phases of text preprocessing is shown in the figure 1

b. Document Categorization

Categorization is the process of dividing the data into number of groups which are either dependent or independent of each other and each group acts as a class. The task of categorization can be done by using several methods using different types of classifiers.

Proposed algorithm for Indian language Document Categorization
Step 1: Identify specific language files.
Step 2: Associate a Language label with each of the files.
Step 3: Build a Corpus C
Step 4: Preprocess the Corpus C.
Step 5: Apply a Stemming algorithm to reduce all the words to their root form.
Step 6: Generate VSM or a Term Document matrix using Binary Term Occurrence D(i, j) (where i is the document i and j is the jth term of document i.)
Step7: Train the Classifier (kNN,j48 and NB) using C as training examples.

c. Keywords extraction

Keywords are widely used as a brief summary and index of documents. Keyword extraction is the task selecting a small set of words from the document that can describe the meaning of the document. So find the term frequency – inverse document frequency also called TF*IDF to evaluate how important is a word in a document. The proposed keyword extraction method is shown in the form or flowchart.

VI. CONCLUSIONS

An attempt is made to propose methods to preprocess the Indian language text documents and also to extract the keyword from the categorized Indian language text documents. The advantage of these methods is they are language independent. That means these are common methods which can be applied for any Indian languages text documents.

The biggest challenge here is to convert unstructured data into structured data. Vector space model (VSM) can be used to represent the text in the form of matrix. Once the data is represented in a structure format, the next step is to Categorization of text document based language. The attention is towards the popularly known k nearest neighbor approach (KNN), Decision tree and naive bayes classifier.

The next phase of work is keyword extraction. TF*IDF will be used to evaluate how important is a word in a document. The TF*IDF will be used as a threshold to select the important keyword.

REFERENCES

[1] "*Local Language Study 2013*" , published jointly by Internet and Mobile Association of India (IAMAI) and IMRB International
[2] "Topic Tracking For Punjabi Language", Kamaldeep Kaur and Vishal Gupta, Computer Science & Engineering: An International Journal (CSEIJ), Vol.1, No.3, August 2011 DOI: 10.5121/cseij.2011.1304 37
[3] "POS Tagger for Kannada Sentence Translation", Mallamma V Reddy and Dr. M. Hanumanthappa, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 1, Issue 1, May-June 2012 ISSN 2278-6856
[4] "Algorithm for Punjabi Text Classification", Nidhi and Vishal Gupta, International Journal of Computer Applications (0975 – 8887) Volume 37– No.11, January 2012

[5] "Topic-Based Bengali Opinion Summarization ",Amitava Das and Sivaji Bandyopadhyay, Coling 2010: Poster Volume, pages 232–240, Beijing, August 2010

[6] "Automatic Punjabi Text Extractive Summarization System", Vishal Gupta and Gurpreet Singh Leha, Proceedings of COLING 2012: Demonstration Papers, pages 191–198, COLING 2012, Mumbai, December 2012.

[7] "Document Summarization In Kannada Using Keyword Extraction" Jayashree.R, Srikanta Murthy.K and Sunny. K, Computer Science & Information Technology (CS & IT)

[8] "Oriya Language Text Mining Using C5.0" Algorithm Sohag Sundar Nanda, Soumya Mishra, Sanghamitra Mohanty ISSN: 0975-9646

[9] "*Automatic Categorization of Telugu News Articles*", Kavi Narayana Murthy, Department of Computer and Information Sciences, University of Hyderabad, Hyderabad, 500 046 http://202.41.85.68/knm-publications/il_text_cat.pdf.

[10] "*A Big Need for Indic-Language Solutions"*, Dr. Deepali Kamthania, Associate Professor, Bharati Vidyapeeth‟ s Institute of Computer Applications and Management, New DelhiCSI Communications March 2014

[11] Sannella, M. J. 1994 Constraint Satisfaction and Debugging for Interactive User Interfaces. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.