ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE) Vol. 2,Special Issue 6, March 2016

Retrieval of Telugu Text using Big Data

Tharini Benarji*, Gopikrishna Chetlapally**,A Bhagya***
*Assoc.Professor, Indur Institute Of Engineering And Technology, tarini_benarji@yahoo.com
** Asst.Professor, Indur Institute Of Engineering And Technology, ch.gopikrishnaa@gmail.com
***Asst.Professor, Indur Institute Of Engineering And Technology, manolahari@gmail.com

Abstract

As more and more data is becoming available due to advances in information and communication technologies, gaining knowledge and insights from this data is replacing experience and intuition based decision making in organizations. Big data mining can be defined as the capability of extracting useful information from massive and complex datasets or data streams. In this paper, we propose a model to mine telugu text over large collection datasets from Big Data .Telugu is an official language derived from ancient Brahmi script and also the official language of the state of Telangana & Andhra Pradesh. Brahmi based script is noted for complex conjunct formations. The canonical structure is described as ((C) C) CV. The structure evolves any character from a set of basic syllables known as vowels and consonants where consonant, vowel (CV) core is the basic unit optionally preceded by one or two consonants. A huge set of characters that resemble the phonetic nature with an equivalent character shape are derived from the canonical structure. Words formed from this set evolved into a large corpus

***Index Terms*—** Big data, data mining, N-gram, bi-gram, tri-gram, classification techniques

I-Introduction

The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SG) slide deck by John Mashey with the title of "Big Data and the Next Wave of InfraStress". Big Data mining was very relevant from the beginning, as the first book mentioning 'Big Data' is a data mining book that appeared also in 1998 by Weiss and Indrukya. However, the first academic paper with the words 'Big Data' in the title appeared a bit later in 2000 in a paper by Diebold .The origin of the term 'Big Data' is due to the fact that we are creating a huge amount of data every day. Osama Fayyad in his invited talk at the KDD Big Mine 12 Workshop presented amazing data numbers about internet usage, among them the following: each day Google has more than 1 billion queries per day, Twitter has more than 250 million tweets per day, Face book has more than 800 million updates per day, and YouTube has more than 4 billion views per day. The data produced nowadays is estimated in the order of zeta bytes, and it is growing around 40% every year.

A new large source of data is going to be generated from mobile devices and big companies as Google, Apple , Face book , and Yahoo are starting to look carefully to this data to find useful patterns to improve user experience. “Big data” is pervasive, and yet still the notion engenders confusion. Big data has been used to convey all sorts of concepts, including: huge quantities of data, social media analytics, next generation data management capabilities, real -time data, and much more. Whatever the label, organizations are starting to understand and explore how to process and analyze a vast array of information in new ways. In doing so, a small, but growing group of pioneers is achieving breakthrough business outcomes. In industries throughout the world, executives recognize the need to learn more about how to exploit big data. But despite what seems like unrelenting media attention, it can be hard to find in-depth information on what organizations are really doing. So, we sought to better understand how organizations view big data, to what extent they are currently using it to benefit their businesses.

II. TELUGU TEXT MINING FRAMEWORK

Definition: Telugu Text Mining is the process of extracting interesting information or knowledge or patterns from the unstructured telugu text that are from different sources. As the telugu text is in unstructured form, it is quite difficult to deal with it. Finding "nuggets" of interesting information from the natural language text is the purpose of text mining. The Text Mining Process is shown in Fig. 1:

Stage-I: Pre-processing Text: Mining from a pre-processed text is easy as compare to natural languages documents. So, pre-processing of Text Documents from different sources

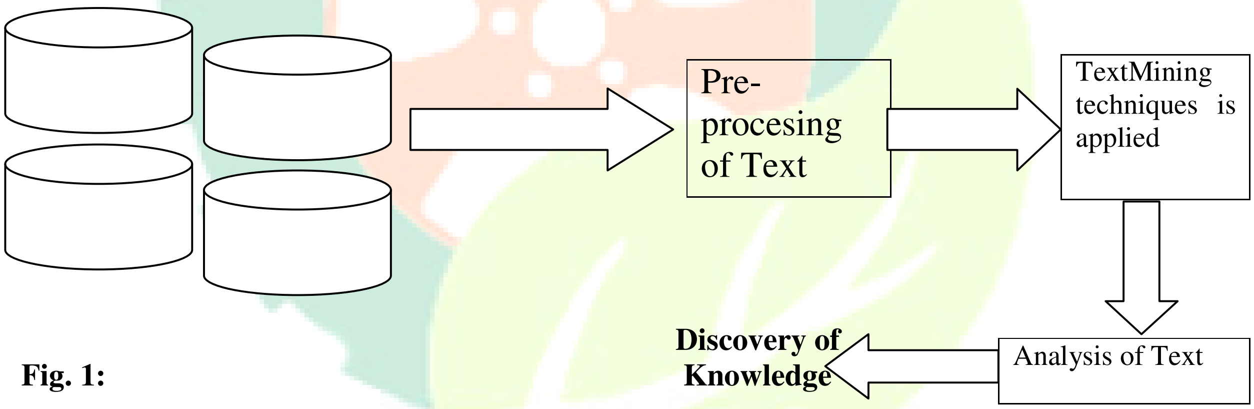

Fig. 1:

III. PROBLEMS OF TELUGU TEXT MINING

One main reason for applying data mining methods to text document collections is to structure them. A structure can significantly simplify the access to a document collection for a user. Well known access structures are library catalogues or book indexes.
However, the problem of manual designed indexes is the time required to maintain them. Therefore, they are very often not upto -date and thus not usable for recent publications or frequently changing information sources like the World Wide Web. The existing methods for structuring collections either try to assign keywords to Documents based on a given keyword set (classification or categorization methods) or Automatically structure document collections to find groups of similar documents (clustering methods).The problem of Text Mining is therefore Classification of data set
And Discovery of Associations among data.
In order to overcome from the problems of Data Mining the following algorithms have been designed.

IV. TASKS OF TELUGU TEXT MINING ALGORITHMS

- Text categorization: assigning the documents with pre-defined categories (e.g decision trees induction).
- Text clustering: descriptive activity, which groups similar documents together (e.g. Self- organizing maps).
- Concept mining: modeling and discovering of concepts, sometimes combines categorization and clustering approaches with concept/ logic based ideas in order to find concepts and their relations from text collections (e.g. formal concept analysis approach for building of concept hierarchy).
- Information retrieval: retrieving the documents relevant to the user's query.
- Information extraction: question answering.

V. BIGRAMS AND TRIGRAMS OF TELUGU WORDS

A general word string consisting of m letters leads to m+1 bi-grams, m+2 trigrams and m+n-1,

N-grams. The theoretical number of possible n- grams is very high. For an alphabet of 26 letters

there are 26^2 = 676 bi-grams and 26^3 = 17576 tri- grams possible. However, in English only 64% of these bi-grams and 16% of all tri-grams actually exist. It is this fact that allows for the detection of spelling errors. The n-grams that are possible from the Telugu word ధశరథరాముడు are explained as sliding a window of length n over this string, progressing one symbol at a time. Redundant coding is applied in this case. The bi-grams formed in this way from the Telugu word ధశరథరాముడు

are: explained as the first case. N-grams are then constructed by

ద దశ శర రథ థరా రాము ముడు డు .The word దశరథరాముడు with 7 letters results in 8 bi-grams. From the same word 9 tri-grams are extracted and listed below. **ద *దశ దశర శరథ రథరా థరాము రాముడు ముడు * డు **

Non-redundant coding uses word fragments with no overlaps. Then the word ధశరథరాముడు yields: * ధ శర థరా ముడు with bi grams and * ధ శర థరా ముడు * with trigrams. Wrongly spelled words usually have a large similarity with their correct version. There are some variations while attempting with bigrams, the last space missed in the above example. We consider a trigram index providing us better result while searching with a Telugu string. Telugu script is syllabic, in the sense that vowels are represented differently in different contexts; the syllabic (primary) context and the intra-syllabic (secondary) context. That is, vowels have one form when they appear in a stand-alone form and in a different form when they appear in conjunction with consonants. Any text contains many variant word forms, such as: "అక్కడ " "అ□□□డె" " అ□□□డాॊ " " అ□□□డకూడ " " అక్కడక్కడ" and so on. A conflation algorithm is a program that brings all these variants together into one word class. Clearly, words belonging to the same class have a very large bi-gram similarity. Similarity between two text strings can be measured indifferent ways. Using bi-grams we see that the similarity is observed between "అక్కడ " and " అక్కడకూడ " and presented as * అ అక్క క్కడ డ * * అ అక్క క్కడ డకూ కూడ డ *

The entire index is to keep a duplicate copy of the index in which each of the index terms is spelled backwards. When a trigram based index is used, word suffix retrieval is simply a word fragment retrieval in which the last character of the trigram is a blank. In general a word is rated as a sequence of characters delimited by blanks. Word prefix retrieval is simply a word fragment retrieval in which the first character of the trigram is a blank.

Classification Algorithm

The Classification problem can be stated as a training data set consisting of records. Each record is identified by an unique record id, and consist of fields corresponding to the attributes. An attribute with a continuous domain is called a continuous attribute. An attribute with a finite domain of discrete values is called a categorical attribute. One of the categorical attribute is the classifying attribute or class and the value in its domain are called class labels.

Objective

Classification is the process of discovering a model for the class in terms of the remaining attributes. The objective is to use the training data set to build a model of the class label based on the other attributes such that the model can be used to classify new data not from the training data set.

Classification Models:

The different type of classification models are as follows:

1. Decision Tree
2. Neural Network
3. Genetic Algorithm

VI. CONCLUSION

Big Data is term for collection of complex datasets. Now way days it will be very popular because it is used for social sites frequently in order to get the information about text, videos, and audios. Among those text data Telugu Text is also stored in the large datasets (in Big Data). Many technologies are developed for the extraction of information from Big Data using different Text Mining Techniques. In this paper we proposed a model to mine Telugu text over large collection of data sets using N-gram, bi-gram and tri gram methods.

REFERENCES:

[1].Ariho ohsato, Izumi Suzuki, Yoshi mikami., A language and character set determination method based on N-gram statistics, ACM Transactions on Asian language information processing, Sep 2002

[2].Cavnar, W. B., Trenkle, J. M., N-gram-Based Text Categorization, Symposium on Document Analysis and Information Retrieval, April 1994.

[3].De Heer, T., Experiments with Syntactic Traces in Information Retrieval, Information torage Retrieval, Volume 10, January 1974.

[4].Adams, E. S., Meltzer, A. C., Trigrams as Index Elements in Full Text Retrieval Observations and Experimental Results

,ACM Computer Science Conference, February 1993

[5].Cavnar, W. B., N-gram-Based Text Filtering for TREC-2, The Second Text Retrieval Conference (TREC-2), February 1994.

[6].Cavnar, W. B., Using an N-gram-Based Document Representation with a Vector Processing Retrieval Model, The Fourth Text Retrieval Conference (TREC-3), April 1995.

Alex Berson and Stephen J.Smith Data Warehousing,Data Mining and OLAP edition 2010.

[7]. Department of Finance and Deregulation Australian Government Big Data Strategy-Issue Paper March 2013 NASSCOM Big Data Report 2012

[8]. Wei Fan and Albert Bifet "Mining Big Data: Current Status and Forecast to the Future", Vol 14,Issue 2, 2013.

[9]. Algorithm and approaches to handle large Data-A Survey, IJCSN Vol 2,Issue 3,2013

[10]. Xindong Wu , Gong-Quing Wu and Wei Ding " Data Mining with Big data ", IEEE Transactions on Knoweledge and Data Enginnering Vol 26 No1 Jan 2014