Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 92 (2016) 18 – 22

Procedia Computer Science

2nd International Conference on Intelligent Computing, Communication & Convergence

(ICCC-2016)

Srikanta Patnaik, Editor in Chief

Conference Organized by Interscience Institute of Management and Technology

Bhubaneswar, Odisha, India

System for Marathi News Clustering

N Dangre , A Bodke , A Date, S Rungta#, S S Pathak*

Students, PICT, Pune

#SQA Engineer, VERITAS, Pune

*IT Dept, PICT, Pune

Abstract

An era of multi-lingual web contents has begun. Web users add, update and search contents in local languages. Automatic translator scripts present available contents in local languages. Therefore, web is attracting users from all levels of society. This advancement has initiated research for text retrieval techniques in local languages. Research community is working on many aspects of this area. In-depth survey reveals retrieval techniques for Marathi language are slightly explored. This work is a proposal towards better text retrieval in Marathi. It suggests system to cluster relevant Marathi news from multiple sources on web. Its application enables rich exploration of Marathi contents on web.

Peer-review under responsibility of the Organizing Committee of ICCC 2016

Keywords: Marathi text, clustering, news retrieval, Marathi retrieval

Introduction:

Internet is no more monolingual. People are interested to retrieve contents in various local regional languages. In India there are twenty two official regional languages. Growing use of these languages on internet has triggered multi-lingual research. Research community is exploring various alternatives in this area. Yet, research on Marathi language retrieval is still in preliminary phase.

Relevant information retrieval and its clustering is an important task in this era, as there is an enormous data on web. Precisely, even Marathi contents are used and updated by many users of web. Therefore, Marathi text clustering is a crucial task. Proposed system basically works on clustering relevant Marathi news from different sources on web. Also, scope includes comparing various clustering algorithms on Marathi text for their efficiency.

Rest of this paper is organized as – section 2 represents the literature survey, section 3 describes pre-processing phase i.e. stemming and stop word removal. Section 4 depicts proposed system and then conclusion.

Literature Survey:

In recent years, capacious amount of text documents in multitudinous Indian languages are up for grabs on internet. For better management and retrieval of such documents, automatic classification can be helpful. Till 2013, classification of Marathi text documents was unexplored, so in work [2] various classification methods are compared for Marathi Text. After testing Naive Bayes, Centroid Classifier, KNN, and modified KNN, results

Peer-review under responsibility of the Organizing Committee of ICCC 2016
doi:10.1016/j.procs.2016.07.317

concluded that Naive Bayes is most efficient considering time and accuracy. Here, Marathi text documents were pre-processed using rule based stemmer and Marathi word dictionary without removing stop words [2]. Fig. 1 shows various topics surveyed under this section, such as Marathi Text pre-processing, classification, clustering and retrieval of news data, and ranking of news.

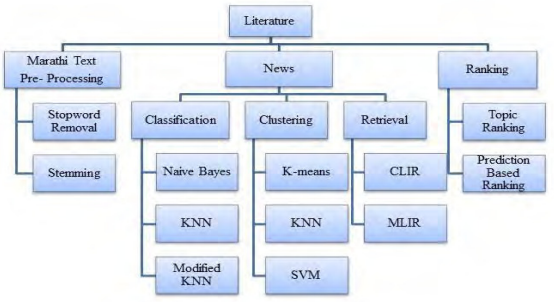

Fig. 1. Topics Explored under Marathi Text Retrieval

After viewing Marathi, Hindi and Bengali language with perspective of Information Retrieval (IR) [1] propounded light and aggressive stemming approaches. Improved retrieval effectiveness can be obtained by applying some aggressive stemmers. After comparison between no stemming and stemming indexing schemes significant performance differences were found. When an aggressive stemmer is applied, related improvements calculated were approximately 28% for Hindi language, approximately 42% for Marathi language and approximately 18% for Bengali language as compared to a no stemming approach. In evaluation of these stemming technologies, FIRE 2008 Test collection and two language independent indexing methods i.e. n-gram and trunc-n are used. To expedite IR operation, two algorithmic stemmers were exhorted. First one is to remove inflectional suffixes and another is to remove frequently occurring derivational suffixes [1].

Automatic Text categorization is a supervised learning task to automatically assign document to predefined classes of documents. Statistical theory methods such as Naive Bayes, KNN, SVM, decision tree have been applied to text categorization. In [11] Novel separability measure is defined based on Support Vector Domain Description (SVDD) and to solve multi-class problems of text categorization, improved decision tree is provided. For text categorization, method based on K-means clustering feature selection is discussed in [12]. K-means is used for clustering as well as for feature selection. For the text data, good features are those words that express correct semantic in a class. For each class, cluster centroids are calculated using K-means method and then text feature for categorization is chosen from high frequency words in a centroid.

To interpret similarities between two documents, a new approach based on fuzzy system, with knowledge base that tries to incorporate human knowledge about significance of named entities categories in the news. There are two approaches for bilingual news clustering i.e. feature selection and document similarity calculation. Paper [3] has represented the documents only by means of Named Entities (NE). Named Entities play an important role in measuring the document similarity. Method to identify NEs present in under resourced Indian Languages is proposed by [4] and presented language independent Multilingual Document Clustering (MDC) approach.

Event registry is a system that can analyse news articles and identify world events in them. The system is able to identify group of articles that describes the same events and identifies group of articles in different languages that describes the same event [5]. Today, various systems that can provide news updates on daily basis, crawls news sites, filters out news and non-news contents (advertisements), group's news into stories on same events and generates summary are available [6].

Retrieval system on Marathi text document based on user profile is useful for better management and retrieval of text document and also makes the document retrieval a simple task. LINGO clustering algorithm based on vector spaced model has been tested on news documents [7]. Some systems mainly focuses on providing personalized documents to the end users by analysing browsing history and user profile of user in Marathi language and presented automatic personalization of Marathi language [8]. System is available for cross language information retrieval which deals with asking query in one language and retrieving the query in other language [9].