ISSN: 2321-7782 (Online)
Volume 3, Issue 3, March 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Automatic Text Categorization Marathi Documents

Jaydeep Jalindar Patil[1]
Department of Computer Engineering
(Computer Networks)
K. J. College of Engineering & Management Research
Pune, India

Nagaraju Bogiri[2]
Professor
Department of Computer Engineering
(Computer Networks)
K. J. College of Engineering & Management Research
Pune, India

Abstract: ***Information technology generated huge data on the internet. Initially this data is mainly in English language so majority of data mining research work is on the English text documents. As the internet usage increased, data in other languages like Marathi, Tamil, Telugu and Punjabi etc. increased on the internet. This paper presents the retrieval system for Marathi language documents based on the user profile. User profile considers the user's interests, user's browsing history. The system shows the Marathi documents to the end user based on the user profile. Automatic text categorization is useful in better management and retrieval of these text documents and also makes document retrieval as simple task. This paper discusses the automatic text categorization of Marathi documents and literature survey of the related work done in automatic text categorization of Marathi documents. Various learning techniques exist for the classification of text documents like Naïve Bayes, Support Vector Machine and Decision Trees etc. There are different clustering techniques used for text categorization like Label Induction Grouping Algorithm, Suffix Tree Clustering, and K-means etc. Literature survey shows that for non-English documents VSM [Vector Space Model] gives the better results than any other models.***

The system provides text categorization of Marathi documents by using the LINGO [Label Induction Grouping] algorithm. LINGO is based on the VSM [Vector Space Model]. The system uses the dataset which contains 200 documents of 20 different categories. The result represents that for Marathi text documents LINGO clustering algorithm is efficient.

Keywords: Categorization, Clustering, Marathi Documents, LINGO, User profile

I. INTRODUCTION

Tremendous amount of data is available on the Internet but its retrieval is tedious. Irrelevant data needs to be eliminated from the result. There are various web-casting services which gives the customized documents to the end user based on the end user profile. In web casting system users have to create their interest profile which is tedious job and it's the manual process. Due to this manual nature profile creation is time consuming and complex process.

Users need to learn how to interact with Web engine and create the profile to retrieve the relevant information. Therefore, there is a need to solve mentioned problems using available Web-casting services.

The main purpose of Automatic Text Categorization is to design Marathi Search Engine which automatically generates user profiles based on the user browsing history. From browsing history system learns the types of content viewed by the end user. Unlike static registration information, user profiles will be updated automatically to reflect current interests of an end user. The user profile will include interest category and interest score to provide the level of interest in a particular category [1], [4].

a) Motivation

As the volume of information available on the Internet continues to increase, there is growing interest in helping user better find, filter and manage this information. In past this information is available mainly in English language. But as the Internet usage increased information in languages other than English increased. In Maharashtra (India) state regional language is Marathi. Marathi uses modified version of Devanagari script and phonetic. This system will help Marathi people for getting the required information in Marathi language. The system will perform the automatic document categorization, personalization based on users user's profile and retrieve the relevant documents which will reduce user's efforts. The system will generate user profile by user's profile which includes user's interest and browsing history.

b) Overview:

Section II of this paper deals with the survey done for different languages, Section III contains the techniques used for text document clustering, Section IV focuses on proposed work and overall system architecture, Section V deals with the experimental setup and results and Section VI deals with conclusion and future work.

II. Survey of Similar Work

Several research papers have worked on categorization and automatic content generation of text documents in various languages namely English, Tamil, Arabic etc.

El-Kourdi used Naïve Bayes classification method to perform automatic categorization of Arabic web documents with 62% accuracy. Saleh Alsaleem focused on automated text categorization of Arabic web documents using Support Vector Machine (SVM) categorization method with 78% accuracy. Kohilavani [1] and E. Iniya Nehru focused on delivering personalized contents in Tamil Language using Naïve Bayes Classification method with about 89% accuracy. The system used topic analyzer to identify user's interest and generated personalized content using intelligent evaluator system. Stanislaw Osinski used LINGO Algorithm with the help of Carrot^2 framework for categorization of English and Polish documents with 80-95% accuracy.

This work focuses on expanding the work for automatic personalization of Marathi documents which will help Marathi people in content generation. The present work uses LINGO [5] Clustering algorithm based on Vector Space Model (VSM).

III. Text Categorization Approach

Currently we have various techniques available for the classification of text documents like Decision Trees, Support Vector Machine, Naïve Bayes,[9][10] etc. Several clustering techniques are available for text categorization like K-means, Suffix Tree Clustering, Label Induction Grouping Algorithm, Semantic Online Hierarchical Clustering (SHOC),[4] etc. But our main focus is on LINGO Algorithm.

a) Clustering Text Documents

Clustering of documents is mainly used to minimize the amount of text by categorizing or grouping similar data items. This grouping is common way for human processing information, and one of the good techniques for clustering helps to build different varieties which provide automated tools. These clustering techniques can also be used to minimize the effects of user in the process. The main feature of high-quality clusters is that data items into the cluster is same but similar to each other and are distinct for two dissimilar clusters. Let us have brief introduction of some clustering approaches.

1. K-means algorithm

This algorithm is an iterative algorithm where the number of input clusters is needed to be mentioned. In this algorithm dataset is split into K clusters and the data points are arbitrarily assigned to the clusters resulting. That has roughly the similar number of data points. For each data point the difference from the data point to each cluster is evaluated. If the data point is

closer to its own cluster than keep it as it is. Suppose If the data point is not closer to its own cluster, copy it into the nearest cluster. The advantage is if the clusters are global than it produces tighter clusters than hierarchical clustering.[4]

2. Lingo Algorithm

To design a web based Search Engine using clustering algorithm, the importance must be given to both contents as well as the description (labels) of the resulting groups which should make sense to the user.

The most of the text clustering algorithms uses a Concept where first cluster content discovery is performed followed by the labels are determined based on the content. But often the Similarity among the documents does not correspond with the Human Understanding. To nullify such issues Lingo converses this process first makes sure that we can create human-perceivable cluster label and then allocate documents.

LINGO algorithm is having a lot of advantages over the other clustering algorithms as follows 1) it supports dynamic clustering according to the user query 2) it identify cluster label first then assigns the document to that cluster 3) it is based on vector space model 4) it can be work for multiple languages and supports multiple keyword based searching.

Lingo first extracts the user readable and frequent words/phrases from the input documents. Further by performing the Reduction of Original Term Document Matrix with Singular Value Decomposition (SVD) method to reduce the term document matrix, and then we find the labels of clusters and then assigns documents to that cluster labels based on the similarity value.[4]

b) LINGO Clustering Algorithm

Below is the pseudo code for the LINGO Clustering Algorithm [4].

Input: *D* = set of documents

Output: Clusters of documents.

1. D ← input documents (or snippets)

Step1. Preprocessing

2. for all d ϵ D do
3. Perform text segmentation of d; {Detect word boundaries etc.}
4. if language of d recognized then
5. apply stemming and mark stop-words in d;
6. end if
7. end for

Step2. Frequent Phrase Extraction

8. concatenate all documents;
9. P_c ← discover complete phrases; {See Section 3.2 for details}
10. P_f ← p : {p ϵ P_c ^ frequency(p) > Term Frequency Threshold};

Step3. Cluster Label Induction

11. A term-document matrix of terms not marked as stop-words and with frequency higher than the Term Frequency Threshold;
12. ∑,U, V SVD(A);

{Product of SVD decomposition of A}

13. k ← 0; {Start with zero clusters}
14. n ← rank(A);
15. repeat
16. k ← k + 1;
17. q ← ($\sum_{i=1}^{k} \sum_{ii}$)/ ($\sum_{i=1}^{n} \sum_{ii}$);
18. until q < Candidate Label Threshold;
19. P phrase matrix for P_f ; {See section 3.3}
20. for all columns of U_k P do
21. find the largest component mi in the column;
22. add the corresponding phrase to the Cluster Label Candidates set;
23. labelScore ← mi;
24. end for
25. calculate cosine similarities between all pairs of candidate labels;
26. identify groups of labels that exceed the Label Similarity Threshold;
27. for all groups of similar labels do
28. select one label with the highest score;

end for

Step4. Cluster Content Discovery

30. for all L ϵ Cluster Label Candidates do
31. create cluster C described with L;
32. add to C all documents whose similarity

to C exceeds the Snippet Assignment Theshold;

33. end for
34. put all unassigned documents in the "Others" group;

Step5. Final Cluster Formation

35. for all clusters do
36. clusterScore ← labelScore × ||C||;
37. end for

First step of LINGO algorithm deals with applying pre-processing techniques to the documents which will help to improve the efficiency of clustering process. The main purpose of pre-processing step is to get rid of form of all input characters and the terms that can affect the quality of group descriptions

In the second step LINGO [4] we remove the "Frequent phrases" from the input documents which are the recurring ordered sequences of terms. "frequent phrase" is determined based on following conditions:

1. It should appear in the documents at least finite number of times.
2. It should be a "complete phrase".
3. It should not begin nor end with stop word.

Third step of INGO Discovers the Frequent Phrases that exceed the term frequency threshold and further they are used for cluster label induction. The *TF-IDF* (Term Frequency- Inverse Document Frequency) weighting scheme is used for this, this scheme calculates "Term Document" matrix and then it identifies labels of the clusters by applying Singular Value Decomposition technique to this matrix.

Single Value Decomposition decomposes a t × d matrix A into three matrices U, ∑ and V, such that M= U ∑V T. U is a t × t orthogonal matrix of M, V is a d × d orthogonal matrix of M, and ∑ is a t × d diagonal matrix of M arranged in decreasing order along its diagonal. The rank rk of matrix M is equal to the number of its non-zero distinctive values. The first rk columns of U form an orthogonal basis for the column space of M.

The phrase with the highest value in the vector is selected as the user graspable concept further by using the cosine ranking the value becomes the score of the candidate of the label of a cluster.

In the fourth step LINGO Searches the documents according to the Cluster. The Vector Space Model is used to allocate the given documents to the labels of clusters obtained from the previous phase. Let Q be the matrix, in which column vector represent the cluster label. Let C =Q^T A, where A shows the "Term Document" matrix of the given documents. Factor cij of the C matrix represents the relationship of the jth document with the ith cluster. A documents are put in to a cluster if cij is exceeds the threshold. And the unassigned documents are added to the separate cluster "Others".

In the last step of LINGO finds the final clusters having maximum scores. Clusters are classified to display depends on their score and is calculated using the following formula: Cluster-Score = *label-score* × $\|C\|$, where $\|C\|$ is the total number of documents of the cluster *C*.

c) An Overview of Processing Query and Clustering

In general a user submits a query to the middle ware via a specialized Web-based interface. The middle ware send request of query to a search engine via the search engine API and set of specified number (top K) of relevant web pages are retrieved. These k relevant documents are given to the clustering process. These k documents are applied to the pre-processing phase then pre processed documents are given to the clustering phase and at last documents in the form of cluster are displayed to the user as a result. The final result is a set of clusters of the documents with the target being to cluster pages based on related to real entity. A group of key words that display the web pages within a cluster is evaluated for each cluster. Every resulting cluster is then processed. The aim is that the user must be able to search the interested document.

1. Pre-processing of documents

The purpose of the pre-processing phase is to remove all stop word characters form input documents and terms that can may be influence the quality of group representation. It involves the two steps:

A. Remove the stop words

B. Stemming

A. Stop Word Removal Process

To remove the stop words from the documents System would maintain the list of stop words in Marathi and these words are compared with each other and if the corresponding match is found then removed it from the document. For Marathi language the list of the stop words is created Manually by using the Marathi dictionary and Marathi Documents and have been given as input to the system .there are about 900 to 1000 Marathi stop words such as “Aani”, “Athava”, “Mhanun”, “Kinva” etc., that does not offers information which is useful about the document’s label which reduces the size of the indexing to the great extent.

B. Stemming

In response to the user’s query, search engines return a list of documents which is ranked but if the query is general then it is becomes difficult for the user to identify the relevant document in which the user is interested. In Stemming techniques the text words are converted into their root words form to improve the efficiency of the search engine in the information retrieval system. After removing the affixes and prefixes the portion of the word left within is called as “Stem” or “Root Word” For example the word “Sachincha” must be converted to the word “Sachin”.

For the classification of the Polish and English Web documents Stanislaw Osinski used the LINGO clustering algorithm in which they found that about 70–80% clusters are useful to the users and 80–95% of snippets into these clusters are relating to their query.

Hence from the literature survey performed, it is concluded that the Vector space model supports the probabilistic model and so decided to use LINGO clustering algorithm for the present work.

IV. Overall System Architecture

Most of the work on categorization of text documents dealt with English, Tamil, Polish, Arabic and Spanish Language. The present work focuses on helping Marathi people by providing Marathi content generation using LINGO Clustering algorithm.[1],[2],[3],[4].

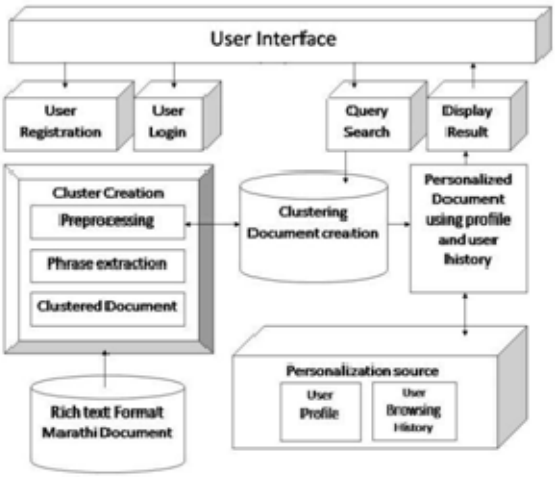

Figure 1: System Architecture Diagram

The detail architecture diagram of the system is shown in Figure 1 which showcases various blocks present in the system along with their functionalities.

The system is divided into three main modules namely:

i) Marathi Data Set

ii) Cluster creation

iii) Personalization of cluster documents.

The first module focuses on the creation of Marathi data set. It creates the Marathi data set from the already downloaded Marathi content.

The second module focuses on creation of cluster base using the clustering techniques of text categorization. This module depends on the query fired by the user. LINGO algorithm will form clusters from the data set. All the preprocessing, phrase extraction of text is carried out in this module.

The third module focuses on the personalization of clusters. The user interface will provide a Registration form to the user. The system will capture user's interest information during the registration of the user. When the user logins to the system the system will analyze the user's interest and deliver personalized information to the user. As the user uses the system the following events are handled namely No. of hits, Time Session.

i) No. of hit's involved

The users have different interest categories and thus it becomes difficult to give most relevant category to the user. So depending on No. of hits for particular category, the desired category is retrieved as first preference when found in the cluster base.

ii) Time Session

Sometimes some users may just click on a certain category but doesn't read the whole document, and then system has to figure out such situations.

Time session is calculated from user input. It's the time difference between open document and close document time.

This reading session time is recorded in the table for further processing. These values are used to calculate Tscore. Tscore formula is given below:

Tscore = Average of (No. of hits * 100/ Max(No. Of hits)) and (Time Session * 100/Max(Time Session))

So in this way all the values are calculated after user performs logout operation. After user performs login operation and fires query, the system will directly retrieve the most viewed document depending on maximum Tscore.

Above all techniques are used for retrieving personalized document in Marathi language. So the result is more efficient, more intensive and accurate for user query.

V. Experimental Setup And Results

A. Dataset

Manual data set of 200 documents for News category is created. The document is in the Rich Text Format for Marathi language. We have considered News categories like Sports, Politics, Crime, Economics, Education, Entertainment, Social, Health, etc.

B. Performance Metrics:

We have used precision and recall as parameters for measuring the performance of the system on any given document set.

Consider D is the set of documents with respect to end user's fired query, M is the set of the returned resultant documents. Also let T denotes the set of documents in D that are relevant to the user's fired query. Finally let T_M denote the intersection of T and M. Now Precision and Recall are calculated.

Now Precision is calculated using formula:

Precision = $| M_T | / | T |$.

Now Recall is calculated using formula: Recall = $| M_T | / | M |$.

C. Evaluation and Results:

The recall and precision parameters have been calculated for each category for analysis purpose. Entertainment data set is tested for user queries namely "Movie", "Actor", "Director", and "Bollywood".

Table 2 shows average of all queries calculated for different categories.

Thus it shows that LINGO Clustering Algorithm is better for categorization of Marathi documents and its quality of performance is very efficient in calculating the user interest

TABLE 1 SYSTEM RESULTS

Sr. No	1	2	4	5	6	7
Category	Entertainment		Sports		Economics	
Query	Cinema	Actor	Umpire	Coach	Funds	Share
Average Precision in %	76	76	81	80	78.6	77.6
Average Recall in %	80.33	77	80.33	88	77	79

VI. Conclusion And Future Work

The system provides automatic text categorization of Marathi documents based on the User's profile which includes User's browsing history. Vector Space Model gives far better results than other Probabilistic Models. The accuracy of results in case of this system is better than as compared to work done in Tamil language. LINGO algorithm supports better cluster quality than other clustering techniques. Lot of Government websites are in Marathi language, so the present work will be useful for the Government websites.

References

1. Kohilavani, S., Mala, T., Geetha, "Automatic Tamil Content Generation", IAMA2009,IEEE International Conference, Sep 2009.
2. El-Kourdi M., Bensaid A. and Rachidi T., "Automatic Arabic Document Categorization Based on the Naive Bayes Algorithm", Proceedings of COLING 20th Workshop on Computational Approaches to Arabic Script-based Languages, pp. 51-58, August 2004.
3. E.Iniya Nehru, T.Mala, " Automatic e-content generation ", International Tamil Internet Conference, Sep 2009
4. Stanislaw Osinski, " An algorithm for clustering of web search results ", Masters thesis , Poznan University of Technology, Poland, 2003.
5. Xiaoou Tang, Ke Liu, Jingyu Cui, Fang Wen and Xiaogang Wang , "Intent Search: Capturing User Intention for One-Click Internet Image Search ", IEEE Year 2012.
6. G. Tascini1, P. Puliti, L. Lella, M. Pallotta, A. Montesanto , "An Automated User Profile System Generator ", IEEE.
7. Matthias Eichstaedt, Qi Lu, San Jose, Shang-Hua Teng , "Automatic user interest profile other publications Generation from structured Document access information.", Year 2002
8. Feng-Hsu Wanga,*, Hsiu-Mei Shaob, "Effective personalized recommendation based on time-framed navigation clustering and association mining", Expert Systems with Applications 27 (2004) 365–377.
9. Susan Dumais, John Platt, David Heckerman, "Inductive Learning Algorithms and Representations for Text Categorization", Microsoft Research One Microsoft Way Redmond, WA 98052.
10. Alex A. Freitas, "A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery", Postgraduate Program in Computer Science, Pontificia Universidade Catolica do Parana Rua Imaculada Conceicao, 1155. Curitiba PR. 80215-901. Brazil.

