Review Paper on Punjabi Text Mining Techniques

[1]Shruti Aggarwal, [2]Salloni Singla
[1,2]Dept. of CSE, Shri Guru Granth Sahib World University

Abstract
Text Mining is a field that extracts useful information from the text document according to users need which is not yet discovered. Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources .Text Classification is one of the text mining tasks to manage the information efficiently, by classifying the documents into classes using classification and clustering algorithms .Each text document is characterize by a set of features used in text classification method, where these features should be relevant to the task. This paper presents techniques for using text mining algorithm to identify the exact keyword of Punjabi newspaper and its text extraction.

Keywords
Text Mining, Text Extraction, Ontology Based Classification

I. Introduction

Researchers aim to use enhanced data mining techniques to extract text from Newspaper with reasonably high recall and precision. In recent years, along with development of Electronic media and electronic newspaper and information technology, technology grows rapidly. With the growth of the electronic media, enormous news paper databases are produced. It creates a need and challenge for data mining. Data mining is a process of the knowledge discovery in databases and the goal is to find out the hidden and interesting information [3]. The technology includes Association rules, classification, clustering, and evolution analysis etc. Clustering algorithms are used as the essential tools to group analogous patterns and separate outliers according to its principles that elements in the same cluster are more homogenous while elements in the different ones are more dissimilar [2]. Furthermore, data mining algorithms do not need to rely on the pre-defined classes and the training examples while classifying the classes and can produce the good quality of clustering, so they fit to extract the Newspaper text better. A major challenge for information retrieval in the life science domain is coping with its complex and inconsistent terminology. In this paper we try to devise algorithms which make word-based retrieval more robust. We will investigate how data mining algorithms based on keywords affects retrieval effectiveness in the newspaper domain. We will try to answer the following research question in this paper "How can the effectiveness of word-based in newspaper information retrieval be improved using data mining algorithm or text mining algorithm.

II. Method

Keywords are a set of significant words in an article that gives high-level description of its contents to readers. Identifying keywords from a large amount of on-line news data is very useful in that it can produce a short summary of news articles. As on-line text documents rapidly increase in size with the growth of WWW, keyword extraction [13] has become a basis of several text mining applications such as search engine, text categorization, summarization, and topic detection. Manual keyword extraction is an extremely difficult and time consuming task, in fact it is almost impossible to extract keywords manually in case of news articles are published in a single day due to their volume. For a rapid use of keywords, we need to establish an automated process that extracts keywords from news articles.

III. Common Techniques of Data Mining/Text Mining

There are many techniques of data mining. The most common techniques used in the field of data mining are followings.

A. Artificial Neural Networks [14]

This is Non-linear predictive models that learn through training and resemble biological neural networks in structure. This predictive model uses neural networks and finds the patterns from large databases.
The Artificial Neural Networks (ANNs) are information processing paradigms inspired by the way biological nervous systems such as the brain process information. A typical example of neural network in human brain's is shows in fig. 1. ANNs adopt this interconnected neuron network to perform complex computations. The typical ANN model is as shown below:

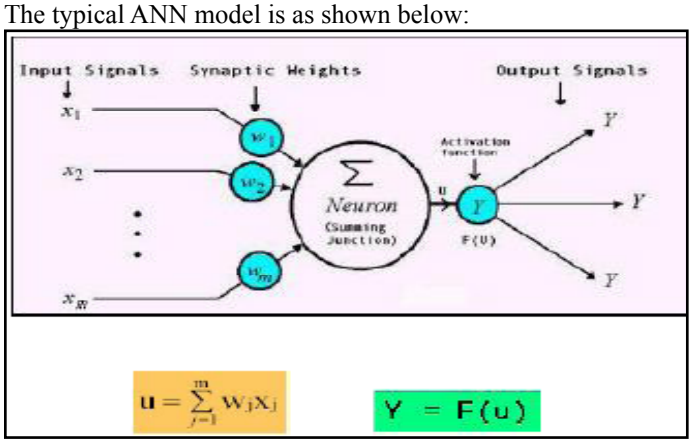

Fig. 1: Typical ANN Model [14]

1. Building the ANN Model

Our Multilayer Perceptron (MLP) is basic concept for the ANN Model. In it a network of processing elements or nodes arranged in layers. Principle Input pattern presented at the input layer causes network nodes to perform calculations in the successive layers until an output value is computed at each of the output nodes from which the most significant is selected.

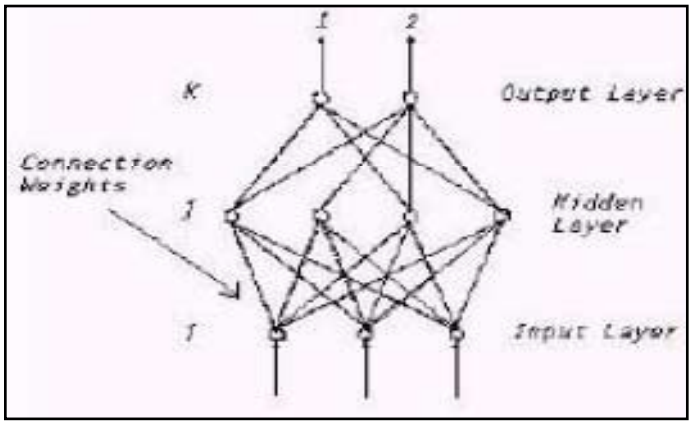

Fig. 2: Building ANN Model [14]

B. Decision Trees [16]

In this Set of decisions are represented by Tree-shaped structures. These decisions generate rules for the classification of a dataset under the large databases. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

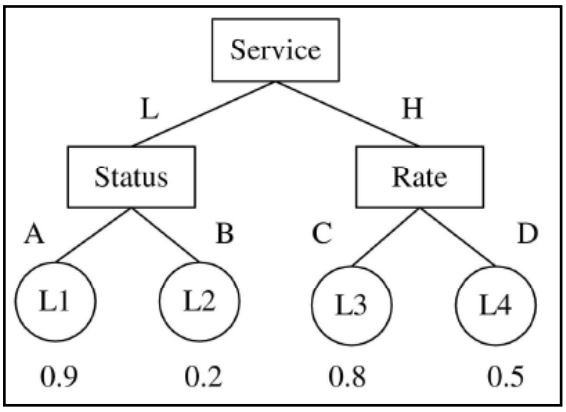

Fig. 3: Decision Tree [16]

Actions that are associated with attribute-value changes, in order to maximize the profit-based objective functions are presented by Decision tree. Large number of candidate actions to be considered, complicating the computation by this. More specifically, two broad cases are considered. One case corresponds to the unlimited resource situation, which is only an approximation to the real-world situations. Another more realistic case is the limited-resource situation, where the actions must be restricted to be below a certain cost level. Aim is to maximize the expected net profit of all the customers as well as the industry in both cases.

C. Genetic Algorithms [15]

Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

Holland in 1970 was developed Genetic Algorithm (GA) [15]. GA is stochastic search algorithm modeled on the process of natural selection, which underlines biological evolution. In many search, optimization, and machine learning problems GA has been successfully applied. GA process is iteration by generating new populations of strings from old ones. Every string is the encoded binary, real etc., version of a candidate solution. A fitness measure to every string indicating its fitness for the problem by evaluation function associated with it. General GA apply genetic operators such selection, crossover and mutation on an initially random population in order to compute a whole generation of new strings [15].

- Selection deals with the probabilistic survival of the fittest, in that more fit chromosomes are chosen to survive. Where fitness is a comparable measure of how well a chromosome solves the problem at hand.
- Crossover takes individual chromosomes from P combines them to form new ones.
- Mutation alters the new solutions so as to add stochasticity in the search for better solutions. In general the main motivation for using GAs in the discovery of high-level prediction rules is that they perform a global search and cope better with attribute interaction than the greedy rule induction algorithms often used in data mining. This section of the paper discusses several aspects of GAs for rule discovery.

Basic Operations

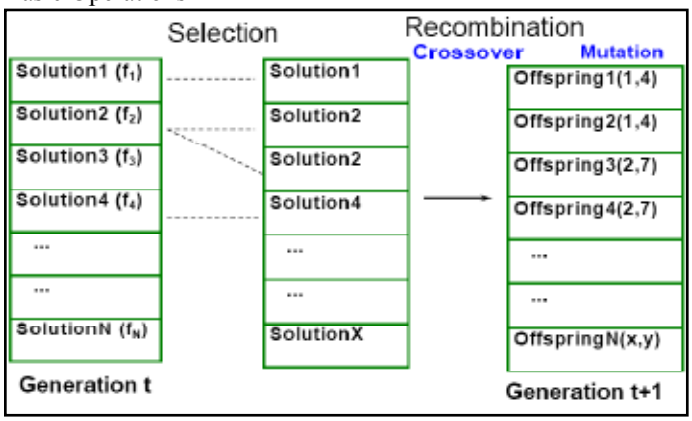

Fig. 4: Basic Operations of Genetic Algorithms [15]

D. Fuzzy C Means [17]

Data clustering is the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible. Depending on the nature of the data and the purpose for which clustering is being used, different measures of similarity may be used to place items into classes, where the similarity measure controls how the clusters are formed. Some examples of measures that can be used as in clustering include distance, connectivity, and intensity. In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In fuzzy clustering (also referred to as soft clustering), data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.

Topics that characterize a given knowledge domain are somehow associated with each other. Those topics may also be related to topics of other domains. Hence, documents may contain information that is relevant to different domains to some degree. With fuzzy clustering methods documents are attributed to several clusters simultaneously and thus, useful relationships between domains may be uncovered, which would otherwise be neglected by hard clustering methods. The modified the Fuzzy c-Means (FCM) algorithm for clustering text documents based on the cosine similarity coefficient rather than on the Euclidean distance [17]. The modified algorithm works with normalized k-dimensional data vectors that lie in hyper-sphere of unit radius and hence has been named Hyper-spherical Fuzzy c-Means (H-FCM). The objective function the H-FCM minimizes is similar to the FCM one [17], the difference being the replacement of the squared norm by a dissimilarity function Diα:

$$J_m(U,V)=\sum_{i=1}^{N}\sum_{\alpha=1}^{c}u_{\alpha i}{}^{m}D_{i\alpha}=\sum_{i=1}^{N}\sum_{\alpha=1}^{c}u_{\alpha i}{}^{m}\left(1-\sum_{j=1}^{k}x_{ij}\cdot v_{\alpha j}\right). \quad (1)$$

The cosine coefficient [17] ranges in the unit interval and when data vectors are normalized to unit length it is equivalent to the inner product. The dissimilarity function Diα in equation (1) consists of a simple transformation of the cosine similarity coefficient, i.e. Diα = 1– Siα.

$$u_{ki} = \left[\sum_{j=1}^{c}\left(\frac{D_{ki}}{D_{ji}}\right)^{\frac{1}{(m-1)}}\right]^{-1} = \left[\sum_{j=1}^{c}\left(\frac{1-\sum_{j=1}^{k} x_{ij}\cdot v_{kj}}{1-\sum_{j=1}^{k} x_{ij}\cdot v_{kj}}\right)^{\frac{1}{(m-1)}}\right]^{-1} \quad (2) \qquad v_{ik} = \sum_{i=1}^{N} u_{ki}{}^{m} x_{ij} \cdot \left[\sum_{j=1}^{k}\left(\sum_{i=1}^{N} u_{ki}{}^{m} x_{ij}\right)^{2}\right]^{-1/2}$$

(2)

E. Modified Algorithm for Punjabi Text Classification [18]

Classification is a set of significant words in an article that gives high-level description of its contents to readers. Classification of large amount of on-line news data is very useful in that it can produce a short summary of news articles. As on-line text documents rapidly increase in size with the growth of WWW, classification has become a basis of several text mining applications such as search engine, text categorization, summarization, and topic detection. Manual classification is an extremely difficult and time consuming task; in fact, it is almost impossible to classify manually in case of news articles published in a single day due to their volume. For a rapid use of classification, we need to establish an automated process that extracts data from news articles.

But for Punjabi Text Document, not much work has been done to classify the documents due to lack of resources, annotated corpora, name dictionaries, good morphological analyzers, POS taggers are not yet available in the required measure.

1. Punjabi Text Classification Process Divides into three Phases [18]

(i). Preprocessing Phase

This phase include process such as, removal of stop words, stemming, punctuation mark and special symbols removal.

(ii). Feature Extraction

This phase includes statistical approach and linguistic approach for the extraction of relevant features from the documents to perform classification.

(iii). Processing Phase

The last phase of the Punjabi text classification, apply text classification algorithms to the extracted features to classify the documents into classes.

2. DATASET

Documents are taken from the Punjabi news web sources such as likhari.org, jagbani.com, ajitweekly.com, punjabispectrum.com, europevichpunjabi.com, quamiekta.com, sahitkar.com, onlineindian.com, europesamachar.com, parvasi.com etc. As classification is a supervised learning, meaning we have predefined classes, so we have classes for these corpus, these are:

ਿਕਾਕਟ (krikaṭ), ਹਾਕੀ (hākī), ਕਬਡੀ (kabḍḍī), ਫੁਟਬਾਲ (phuṭbāl), ਟੈਿਨਸ (ṭainis), ਬੈਡਿਮੰਟਨ (baiḍmiṇṭan), ਉਲੰ ਿਪਕ (ōlmpik) and Others classes for the sports ontology.

3. Methodology

Step 1: Remove all special symbols e.g. <,>, :,{,},[,],^,&,*,(,), extra tabs, spaces, shifts from the
text documents.

Step 2: Remove stopwords e.g. ਦੇ (dē) (vicc), ਦੀ (dī), ਹੈ (hai), ਇਹ (ih) (valōṃ),, ਹਨ (han), ਨੰ (nūṃ) Stopwords List.

Step3: Extract names, places, dates, months name etc the text document using Gazetteer lists.

Step4: Calculate term frequency (TF) for each remaining word.

Step5: Eliminate terms whose term frequency is below the threshold value.

Step6: Calculate Inverse Document Frequency each word from the document after pre step.

Step7: Calculate TF *IDF of each word those words that are having TF less than threshold value. This step will further help in reducing dimensionality.

Step8: Create ontology for each class that consists of terms related to its classes. E.g. for Cricket Class Ontology for sports, we have terms such as ਗCਦਬਾਜੀ (gēndbāzī), (vikaṭ), ਸਿਪਨ (sapin (vikṭakīpar) etc. This results in Class wise list.

Step9: Remaining terms from step 7 is matched with each Class-wise list, and if maximum terms are matched with one class, assign that class to the unlabelled document.

4. Overview to the Process of Classification

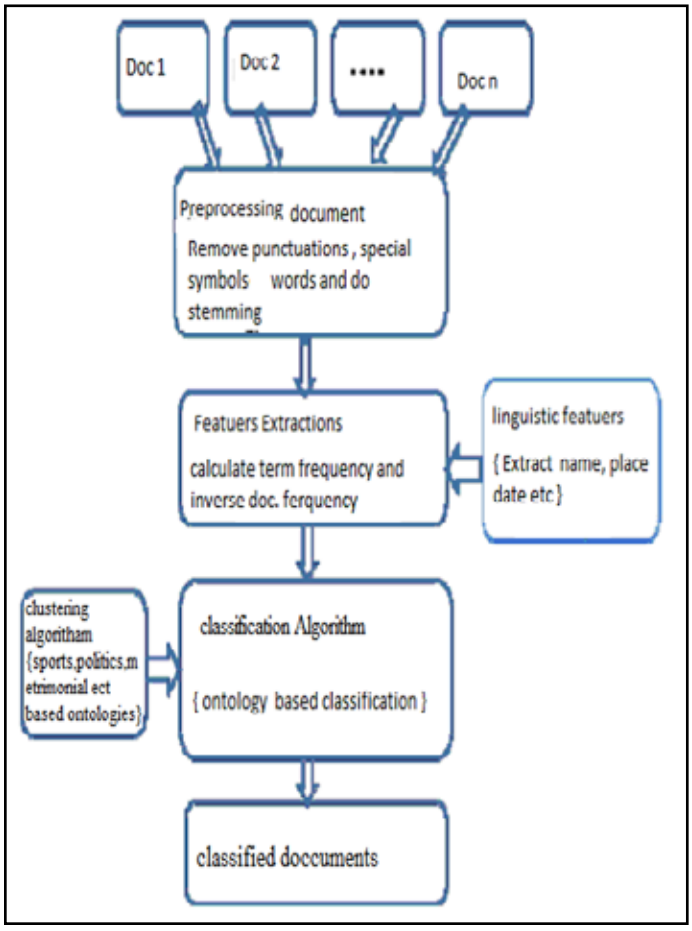

Fig. 5: An overview to the Process of Classification

IV. Conclusion

Extraction of text from Punjabi news paper literature is an essential operation. Given that there have been many text extraction methods developed; this paper presents a novel technique that employs keyword based article clustering to further get the text extraction process. The development of the proposed work is of practical significance; however it is challenging to design a unified approach of text extraction that retrieves the relevant text articles of Punjabi newspapers more efficiently. The proposed algorithm, using data mining technique called text-mining, seems to extract the text of seven keywords i.e sports section :- lawn tennis, cricket, hockey. Politics section- National and centre level, Matrimonial section. Seven tables will be maintained in a database and the relevant text will be displayed whenever the user enters the keyword related to this news section.

V. References

[1] "Efficient Retrieval of Text for Biomedical Domain using Data Mining Algorithm", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 4, 2011.

[2] Gupta Vishal, Lehal Gurpreet S,"A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol 1, No. 1, August 2009.

[3] Clifton, Christopher,"Encyclopedia Britannica: Definition of Data Mining", 2010.

[4] Han Jiawei, Kamber Michelin,"Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 70-181, 2001.

[5] Nidhi M.E, Gupta Vishal,"Recent Trends in Text Classification Techniques", International. Journal of computer Application, Vol. 35, No. 6, December 2011.

[6] Lehal G S, Singh Chandan,"Feature extraction and classification for OCR of Gurmukhi script", Vivek, Vol. 12, No. 2, pp. 2-12, 1999.

[7] Kantardzic, Mehmed, John Wiley & Sons,"Data Mining: Concepts, Models, Methods, and Algorithms", 2003.

[8] Miller, H., Han, J,"Geographic Data Mining and Knowledge Discovery", London: Taylor & Francis, 2001.

[9] Manu Aery, Naveen Ramamurthy, Y. Alp Aslandogan "Topic Identification of Textual Data", Technical report, The University of Texas at Arlington, 2003.

[10] Pavel Berkhin,"Survey of Clustering Data Mining Techniques", Technical report, Accrue Software, San Jose, CA, 2002.

[11] Cecil Chua, Roger H.L. Chiang, Ee-Peng Lim,"An integrated data mining system to automate discovery of measures of association", In Proceedings of the 33rd Hawaii International Conference on System Sciences, 2000.

[12] George Forman, J. Mach,"An extensive empirical study of feature selection metrics for text classification", 2003.

[13] Rayid Ghani,"Combining labeled and unlabeled data for text classification with a large number of categories", In IEEE Conference on Data Mining, 2001.

[14] Karan Kamdar, Amit Mathapati,"The Artificial Neural Networks for Cancer Research in Prediction & Survival (ANNCRIPS)", Vivekanand Institute of Technology, India

[15] Sufal Das, Banani Saha,"Data Quality Mining using Genetic Algorithm", International Journal of Computer Science and Security, (IJCSS) Vol. 3, Issue 2.

[16] Mrs. Swati, V. Kulkarni,"Mining knowledge using Decision Tree Algorithm", International Journal of Scientific & Engineering Research, Vol. 2, Issue 5, May-2011.

[17] Ahmad Shahi, Rodziah Binti Atan, Nasir Sulaiman ,"An Effective Fuzzy C-Mean and Type-2 Fuzzy Logic for Weather Forecasting", Journal of Theoretical and Applied Information Technology , 2009 .

[18] Salloni Singla, Shruti Aggarwal,"Punjabi Text Using Clustering and Classification Techniques", IJCST Vol. 4, Issue 1, Jan - March 2013.

Shruti Aggarwal did her Master's in Engineering in Computer Science from U.I.E.T., Panjab University, Chandigarh in 2011. She did her B.Tech (Computer Science & Engineering) from Ambala College of Engineering and Applied Research, Kurukshetra University in 2008. She has teaching experience of one year as lecturer at Emax Group of Institutes, Ambala. She has also taught at Dr. B.R.Ambedkar National Institute of Technology, Jalandhar for one semester. From last one year, she is working as Assistant Professor at Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India. Her main research area is Data Mining and she has published many research papers in International Journals and Conferences on Audio Mining and other Data Mining Techniques.

Salloni Singla did her B.Tech (Information Technology) from Baba Banda Singh Bahadur Engineering College Fatehgarh Sahib, Punjab Technical University in 2011. She is doing her Master's in Engineering in Computer Science from Shri Guru Granth Sahib World University, Fatehgarh Sahib.