Review of Data Mining Techniques used in Punjabi Text Classification

Sandeep Kaur[1] Ada Thour[2]
[1,2]Guru Granth Sahib World University Fatehgarh Sahib

Abstract— Since evolution of databases which gave rise to data mining, many kind of approaches are being researched and given a practical phase as well ,in order to get the desired results as well as desired reliability. However data mining is never ending process it will remain as an evolution towards its better results and outcomes. Text classification is one of the text mining technique which is used to manage the information by classifying the document into different classes using different classification algorithms. This paper surveys different data mining techniques and their usage, approaches and classification algorithms used to classify the documents into classes.

Key words: Data Mining; Association; Classification; Clustering; Rule Induction

General Terms: TECHNIQUES, APPROACHES, ALGORITHMS. DECISION TREES

I. INTRODUCTION

Data mining is one of the important step of knowledge discovery process[1]. Data mining is the process of extracting import information from large amount of data. Due to the rapid growth of digital data built in recent years, knowledge discovery and data mining have turn into a great deal of consideration with an immediate need for passing such data into useful information and knowledge. Knowledge discovery can be examine as the process of nontrivial derived information from extensive databases. Information that exists in the data is greatly useful for users. In the past years, a significant amount of data mining approaches have been proposed in order to achieve different knowledge tasks. These approaches include association rule mining, frequent itemset mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. Most of them are proposed for the purpose of developing effective mining algorithms to find particular patterns within a right and acceptable time limit. With a large number of patterns generated by using data mining approaches how to effectively use and update these patterns is still an open research issue. In this paper, we target on the development of a knowledge discovery model to effectively use and review the discovered patterns and apply it to the field of text mining.

II. TECHNIQUES IN DATA MINING

A. Association:

Association (or relation) is probably the better known and most familiar and important data mining technique. This make a simple correlation between two or more components, of the same type to identify patterns. For example, when tracking people's buying habits, you may find that a customer always buy a cream when they buy strawberries, and therefore suggest that the next time that they buy strawberries they might also want to buy cream. association or relation building based on data mining tools can be achieved simply with different tools.

B. Classification:

The classification to build up an idea of the type of customer, item, or object by describing multiple attributes to identify a particular class. For example, we can easily classify cars into different types (sedan, 4x4, convertible) by identifying different attributes. Given a new car, we might apply it into a particular class by comparing the attributes with our known sharpness. We can apply the same principles to customers, for example by classifying them by age and social group. We can use classification as the result of other techniques. For example, we can use decision trees to determine a classification. Clustering allows to use common attributes in different classifications to identify clusters.

C. Clustering:

By examining one or more attributes or classes, we can group individual pieces of data together to form a structure opinion. Clustering is the process of using one or more attributes as basis for identifying a cluster of correlating results. Clustering is useful to identify different information because it correlates with other examples so that we can see where the similarities agree.

D. Decision Trees:

The decision tree can also be used either as a part of the selection criteria, or to support the use and selection of specific data within the overall structure just as other techniques like classification and prediction are used. Decision trees are often used with classification systems to attribute type information, and with predictive systems, where different predictions might be based on past historical experience that helps to drive the structure of the decision tree and the output.

E. Combinations:

It's very rare that one of these are used exclusively. Classification and clustering are similar techniques. By using clustering to identify nearest neighbours, it can further refine classifications. We can use decision trees to build and identify classifications that we can record for a longer period to identify sequences and patterns.

III. DATA MINING APPROACHES

There are many different data mining approaches . Some among them which can be used in classification are given below.

A. Association Rule Mining Approach:

We use rule induction in data mining to obtain the accurate results with fast processing time. Through rule induction we can minimize the numbers of rules and maximize the coverage of data. If we use rule induction along with association rule mining then it can generates less numbers of rules with more accurate result. Association Rules form a most applied data mining approach. Association Rules are derived from frequent item-sets.

B. Decision List Induction Algorithm:

Decision list induction algorithm is used to make order and unordered list of rules to coverage maximum data from the data set. Using decision list induction we can generate number of rules for training dataset to achieve more accurate result with minimum error rate. The CN2 induction algorithm is a learning algorithm for rule induction. It is designed to work even when the training data is imperfect. It is based on ideas from the AQ algorithm and the ID3 algorithm. As a consequence it creates a rule set like that created by AQ but is able to handle noisy data like ID3. The algorithm must be given a set of examples, Training Set, which have already been classified in order to generate a list of classification rules. A set of conditions, Simple Condition Set, which can be applied, alone or in combination, to any set of examples is predefined to be used for the classification.

C. Rule Induction:

Rule induction is an area of machine learning in which formal rules are extracted from a set of observations. The rules extracted may represent a full scientific model of the data, or can also represent local patterns in the data.

IV. ALGORITHM

Following is the proposed algorithm using Ontology Based Classification to classify Punjabi text documents into eight predefined classes. These classes are: ਹਾਕੀ (hākī), ਕਬਡੀ (kabḍḍī), ਫੁਟਬਾਲ (ṭainis), ਬੈਡਿਮੰਟਨ (baiḍminṭan), ਓਲੰਪਿਕ Others. The Sports Based Ontology prepared for Punjabi Text Classification, contains sports related terms of each predefined classes [13].

1) Step1**:** Remove all special symbols e.g. <,>,:,{,},[,],^,&,*,(,), extra tabs, spaces, shifts from the text documents.

In step 1 tokenization is done. Tokenization is the task of chopping text into pieces, called tokens and throwing away special symbols,characters such as punctuations.

2) Step2**:** Remove stopwords e.g. ਦੇ (dē) (vicc), ਦੀ (dī), ਹੈ (hai), ਇਹ (ih) (valōṃ),,ਹਨ (han), ਨੰੂ (nūṃ) Stopwords List.

Stop words are the commom words from which we do not gain any information. Stop list can be determined by collection frequency. Collection frequency is the total number of times the term appears in document collection.

3) Step3**:** Extract names, places, dates, months name etc the text document using Gazetteer lists.
4) Step4**:** Calculate term frequency (TF) for each remaining word.
5) Step5**:** Eliminate terms whose term frequency is below the threshold value.
6) Step6**:** Calculate Inverse Document Frequency each word from the document after pre step.
7) 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.
8) Step8: Create ontology for each class that consists of terms. We have terms such as ਗ`ਦਬਾਜ਼ੀ (gēndbāzī),(vikaṭ), ਸਿਪਨ (sapin (vikṭakīpar) etc.
9) Step9: Remaining terms from Class-wise list, and if maximum terms are matched with one class, assign that class to the unlabelled document.

– Input: 150 Punjabi Text Documents (related to Sports only)
– Classes: ਿਕ1ਕਟ (krikaṭ), ਹਾਕੀਫੁਟਬਾਲ (phuṭbāl), ਟੈਿਨਸ (ṭainisਓਲੰਿਪਕ (ōlmpik) and Others.
– Output: This is the proposed results where e document is classified into its class.

Step7 is matched with each class maximum terms are matched with one class; assign that class to the unlabeled document

V. CONCLUSION

This paper deals with the classification of the Punjabi Text from Punjabi newspaper and detecting its particular category. The sport based ontology prepared for Punjabi text classification contains sport related terms of each predefined classes. This ontology can be further improved by defining the terms of other categories, hence results in better performance of the classifier.

REFERENCES

[1] J.H. Kroeze,M.C. Matthee and T.J.D.Bothma , "Differentiating between data-mining and text mining terminology",doi:10.1.1.95.7062, July 2007.
[2] Kao, Anne, Poleet, R. Steve, (Eds.), "Natural Language Processing and Text Mining",1st edition,XII, 265p, 655illus 2007.
[3] Vishal Gupta, Gurpreet S. Lehal, Journal of Emerging Technologies in Web Intelligence, VOL. 1, NO. 1, August 2009, "A Survey of Text Mining Techniques and Applications".
[4] Vidhya.K.A and G.Aghila,"Hybrid text mining model for document classification", IEEE volume 1, pp: 4244-5586, 2010.
[5] Anagha R Kulkarni, Vrinda Tokekar,Parag Kulkarani, Cummins College of engineering for women Karvenagar ,pune, "Identifying context of text document using naïve bayes classification and apriori association rule mining".
[6] K. Nithya, P C D.Kalaivaani and R.Thangarajan et al Kongu Engineering College,India, 2012 ,"An Enhanced Data Mining Model for Text Classification".
[7] Dharam Veer Sharma, Gurpreet Singh Lehal , "Form Field Frame Boundary Removal Processing System" , IEEE ,doi:10.1109/ICDAR.2009.179, 2009.
[8] Jingnian Chen, Houkuan Huang, Shengfeng Tian and Youli Qu, Expert Systems with Applications:An International Journal, Volume 36 Issue 3, and Elsevier,2009, "Feature selection for text classification with Naive Bayes",

[9] Nidhi and Vishal Gupta pp. 245–252, doi: 10.5121/csit.2012.2421, 2012," Punjabi Text Classification using Naïve Bayes, Centroid and Hybrid Approach".

