IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

Text Classification of CrowdSourced Data Using Parallel Computing Nested Baye's Classifier

Venkat Raman B[1], Rajesh Jakkula[2] and Mahesh Pathyam[3]

Lecturer, CSE department, RGUKT Basar, Hyderabad, India [1]

Student, CSE department, RGUKT Basar, Hyderabad, India [2, 3]

Abstract: Text classification has drawn attention due to the vast applicability in language identification, spam filtering, genre classification, sentiment analysis, readability assessment, article triage etc. With the goal of classifying crowdsourced data which is gathered from many websites into text classification, Crowdsourced data classification has challenges in getting data from various sources and whatever the data we are getting is huge in amount. Our approach will address challenges to classify the crowdsourced data in parallel computing in which many calculations are carried out simultaneously so that the crowdsourced data will be divided into tree format and calculations will be done simultaneously using nested baye's classifier. Our aim is to classify the algorithms which are obtained using web crawler and will be converted into programmes using nested baye's classifier technique.

Keywords: Web Crawler, Text Classification, Nested Bayes Classifier, Parallel Computing, CrowdSourcing.

INTRODUCTION

Crowds of people can solve some problems faster than individuals or small groups. A crowd can also rapidly generate data about circumstances affecting the crowd itself. This crowdsourced data can be leveraged to benefit the crowd by providing information or solutions faster than traditional means. However, the crowdsourced data can hardly be used directly to yield usable information. Intelligently analyzing and processing crowdsourced information can help prepare data to maximize the usable information, thus returning the benefit to the crowd. This article highlights challenges and investigates opportunities associated with mining crowdsourced data to yield useful information, as well as details how crowdsourced information and technologies can be used for response-coordination when needed, and finally suggests related areas for future research [6]. Text classification is classifying the text into fixed number of predefined categories. There is set of texts in each category which is already trained in the program. Whatever the crowdsourced data we are getting is in the form of text and it will be categorized into one of the category. What we are going to propose is the crowdsourced data will be in larger size and the calculations should be done very fast [2]. For that we are going to use multiple processes for parallel computing and for each processor we use a nested baye's classifier. For each category in text classification we may have sub categories. For that we use a nested baye's classifier.

RELATED WORK

Text classification can be performed within crowdsourced data. For instance, users can categorize documents or can assign labels, also known as classes (i.e., tags), to documents manually [5].

Crowdsourcing is often used to obtain solutions to a problem that are cheaper and superior in quality and quantity to those that are obtained from traditional professionals in the same industry.

Some websites such as Wikipedia, encyclopaedia and world map should not be developed by individual, they should developed by large number of users individually.

Facebook is another popular application that can be used for crowdsourcing data like business and market analysis urban planning, and product repository generation.

Facebook provides many applications such as designing surveys, opening forums for discussion, dropping a note to bring awareness to a topic, and creating interest groups. Facebook is a mechanism for building brands, calling people to action, or even introducing ideas.

Aweb crawler(also known as awebspider orwebrobot) is a program or automated script which browses the World WideWebin a methodical, automated manner. To visit files or websites in order to index them for searching.

Many legitimate sites, in particular search engines, Use spidering as a means of providing up-to-date data.

Parallel computing: crowdsourced data is divided into multiple processes for the purpose of doing the calculations faster. Each process is assigned with nested baye's classifier. This is because to do the calculations faster in nested baye's classifier. How many processes required will be based upon how much crowdsourced data we are getting through the web crawler for classification [1].

Baye's classifier: It is a simple probabilistic approach to categorize the given data into one of the predefined categories. This is based on baye's theorem. This is comes under supervised classification problem. This is because of the set of texts which is trained in the baye's classifier.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

Baye's Theorem:

$$P(A/B) = \frac{(P(B/A).P(A))}{P(B)}$$

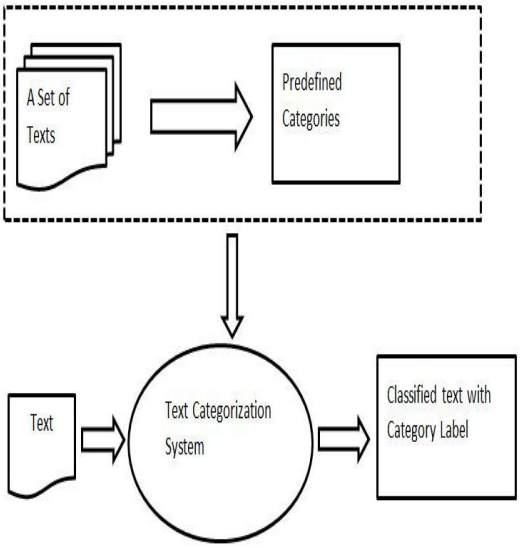

Fig: Process of Baye's classifier

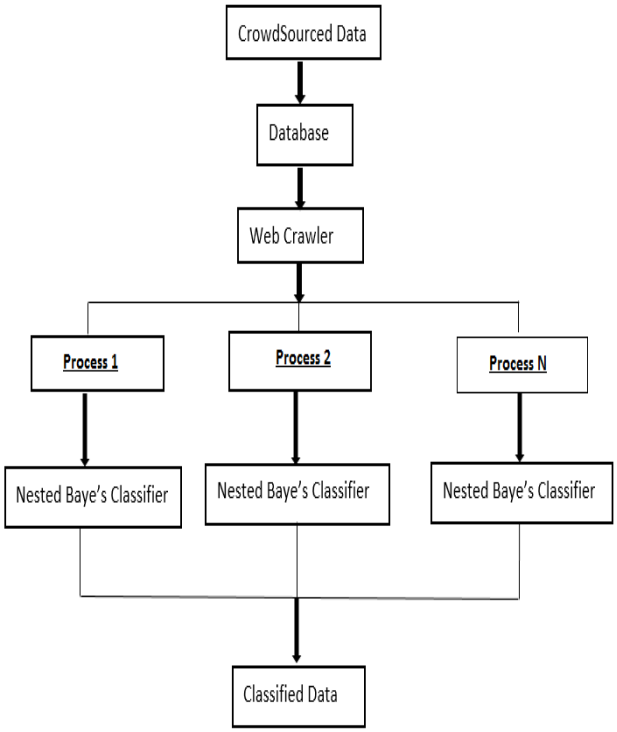

Fig: Text Classification of Crowdsourced data using Parallel computing Nested Baye's classifier.

PROPOSED METHOD

There are so many users who wants their algorithm to be converted into program. There are websites which ask users to enter algorithms into their websites so that all of the users can know others people views on how one problem can be viewed in different ways. We want to convert those algorithms into programs so that every individual algorithm will be converted into program and shared with all the users and the users will be benefited with those different types of programs. For this purpose we will be using a web crawler for getting the algorithms from different websites and we will be maintaining a database for the set of texts for the dictionary purpose. We don't use any dictionary externally. It is required to obtain samples with sufficient information so that the hypotheses devised from the sample information can be easily generalized to larger datasets. However, the question that needs to be answered here is how the sampling distribution should be selected so that the information obtained from the sample is maximized. We will be storing the set of texts in predefined categories in database and that will be acted as a dictionary for classification. There is huge amount of data we are getting through web crawler and to make it simple and not to take much time for classification, we will be using multiple processes for the classification and for each process we will be assigning a nested baye's classifier. Why we use nested baye's classifier is there may be sub categories which also have to categorize into one of the class. Suppose we have an algorithm, we have to convert that into code. We need not use nested baye's classifier for taking input from the user because there are no sub categories in taking inputs. Whenever it comes to print output we require nested baye's classifier because user may ask for printing just a message or printing values as an output.

IMPLEMENTATION

Whatever the data is crawled from web crawler will be stored in the database. It will be checked for the classification whether the text is already exists or not. If the text is already exists in the database no need to go for classification. If it doesn't exists then it will have to go for classification. We will be maintaining two databases. One for the crowdsourced data and another is for the web crawler. Whatever the data is entering from multiple users will be stored in database. In our database the data is algorithm which is in the English language. Web crawler crawl the text and will be check for the classification.

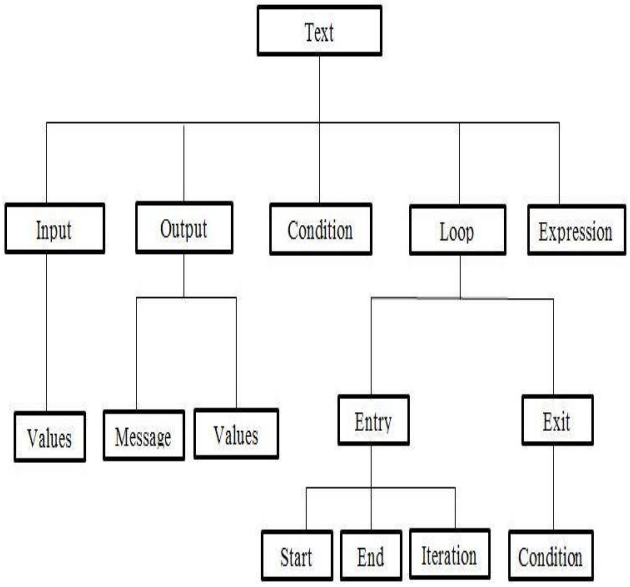

Fig: Predefined Categories and sub categories

DOI 10.17148/IJARCCE.2016.53181

Predefined categories and Set of Texts:

Input: read, take, accept, user, keyboard, input
Output: write, display, print, show, monitor

Loop: repeat, iteration, continue, till, goto, as long as

Condition: if, condition

Expression

Input, Output, Loop, Condition, Expression are all predefined categories in which we have set texts as shown above. Now the text which is obtained from web crawler will be classified into one of the category using nested baye's classifier and eventually the algorithm will be converted into the desired program.

CONCLUSION

Basically the current contribution of this project are proposing how the crowdsourced data which is obtained from multiple sources through the web crawler will be classified into categories and the use of the web crawler to crawl the data from many websites and storing the information in database. After this we came across parallel computing to divide the data into multiple processes for the ease of doing calculations in efficient manner.

Set of texts are trained in the predefined categories and the text which is obtained from database of a web crawler will be classified into any one of the category. This will be help for the people who want to see other people algorithms and other people codes for their problem. The users will see how other people write algorithms and they can share their ideas through this and for all of them they can get the code whatever the algorithm they write.

REFERENCES

[1] Text: Akl, Parallel Computation, Model and Methods Prentice Hall, 1997.
[2] Manas Joglekar, Adam Marcus, Aditya Parameswaran, and Vasilis Verroios, "Challenges in Data Crowdsourcing", IEEE Transactions on knowledge and data engineering, Vol.28, NO.4 Hector Garcia-Molina, Member, IEEE,
[3] D. R. Karger, S. Oh, and D. Shah, "Efficient crowdsourcing for multi-class labeling," in Proc. ACM SIGMETRICS/Int. Conf. Meas. Model. Computer. Syst., 2013, pp. 81–92.
[4] Parallel Computing Systems and Applications J.Hollingsworth Dept. of computer science, University of Maryland, USA
[5] An empirical study of the naïve baye's classifier. I.Rish, T.J Watson research centre.
[6] Maximizing benefits from crowdsourced data, Geoffrey Barbier · Reza Zafarani · Huiji Gao Gabriel Fung · Huan Liu) Published online: 25 June 2012 © Springer Science + Business Media, LLC 2012 2012

BIOGRAPHIES

Venkat Raman B, is working as teaching faculty in the department of Computer Science and Engineering, Rajiv gandhi University of Knowledge Technologies-Basar, India. His passions include teaching, seminars and conducting workshops. His research interest is in the area of data mining and machine learning.

Jakkula Rajesh is an undergraduate student at Rajiv gandhi University of Knowledge Technologies-Basar, India. His passions include developing innovative software. His research interest is in the area of data mining and machine learning.

Pathyam Mahesh is an undergraduate student at Rajiv gandhi University of Knowledge Technologies-Basar, India. His passions include developing innovative software. His research interest is in the area of data mining and machine learning.

DOI 10.17148/IJARCCE.2016.53181