WEB PAGE CLASSIFICATION USING ANT COLONY OPTIMIZATION

Submitted By:

Bhanu Prakash (RIT 2012035)

Vamsi Krishna (RIT 2012061)

Kulshreshtha Singh (RIT 2012070)

Cyril Santosh (RIT 2012087)

Shiva Reddy (RIT 2012088)

Aditya Shama (RIT 2012090)

Under Supervision of:

Dr. Sonali Agarwal

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY ALLAHABAD

DECEMBER 2014

INTRODUCTION

The amount of information available on the web is huge and growing each year. At present Google searches more than 4.2 billion pages. As the web has grown, the ability to mine for specific information has become almost important as the web itself.

Data mining consists of a set of techniques used to find useful patterns within a set of data and to express these patterns in a way which can be used for intelligent decision making. In this project the knowledge is represented as classification rules. A rule consists of an antecedent (a set of attribute values) and a consequent (class):

```
IF <attrib = value> AND ... AND <attrib = value> THEN <class>.
```

The class part of the rule (consequent) is the class predicted by the rule for the records where the predictor attributes hold. This is important, because the general goal of data mining is to discover knowledge that is not only accurate, but also comprehensible to the user.

In this project, the goal is to discover a good set of classification rules to classify web pages based on their subject. The main classification algorithm to be used in this paper is Ant-Miner, the first Ant Colony Optimisation (ACO) algorithm for discovering classification rules.

PROBLEM DEFINITION AND SCOPE

Web mining is the application of data mining techniques to discover usage patterns from large Web data repositories. Objective of this project is twofold:

1. To categorize web pages based on their subject by implementing Ant Miner Algorithm.
2. Calculate and discuss classification accuracies based on the proposed and actual subject of web pages.

LITERATURE SURVEY

Reviews are very important, especially when someone is starting on a new research area. For a good review on Ant Colony Optimization we referred to a research paper by Nicholas Holden. In that paper, he used Ant-Miner – the first Ant Colony algorithm for discovering classification rules – in the field of web content mining, and showed that it is more effective than C5.0 in two sets of BBC and Yahoo web pages used in that experiments. It also investigated the benefits and dangers of several linguistics-based text pre-processing techniques to reduce the large numbers of attributes associated with web content mining.

METHODOLOGY

To implement the above goals, following methodology is followed:

1. Two datasets are harvested from the BBC standard repository. One dataset has 5 classes, other has 3 classes.

2. For improving the total execution time, the total number of web pages are restricted to 500 and 300 respectively in both the datasets.

3. Each dataset is taken as an input to a Graphical User Interface (GUI), implemented in JAVA.

4. The dataset is classified into relation, attributes, classes and instances.

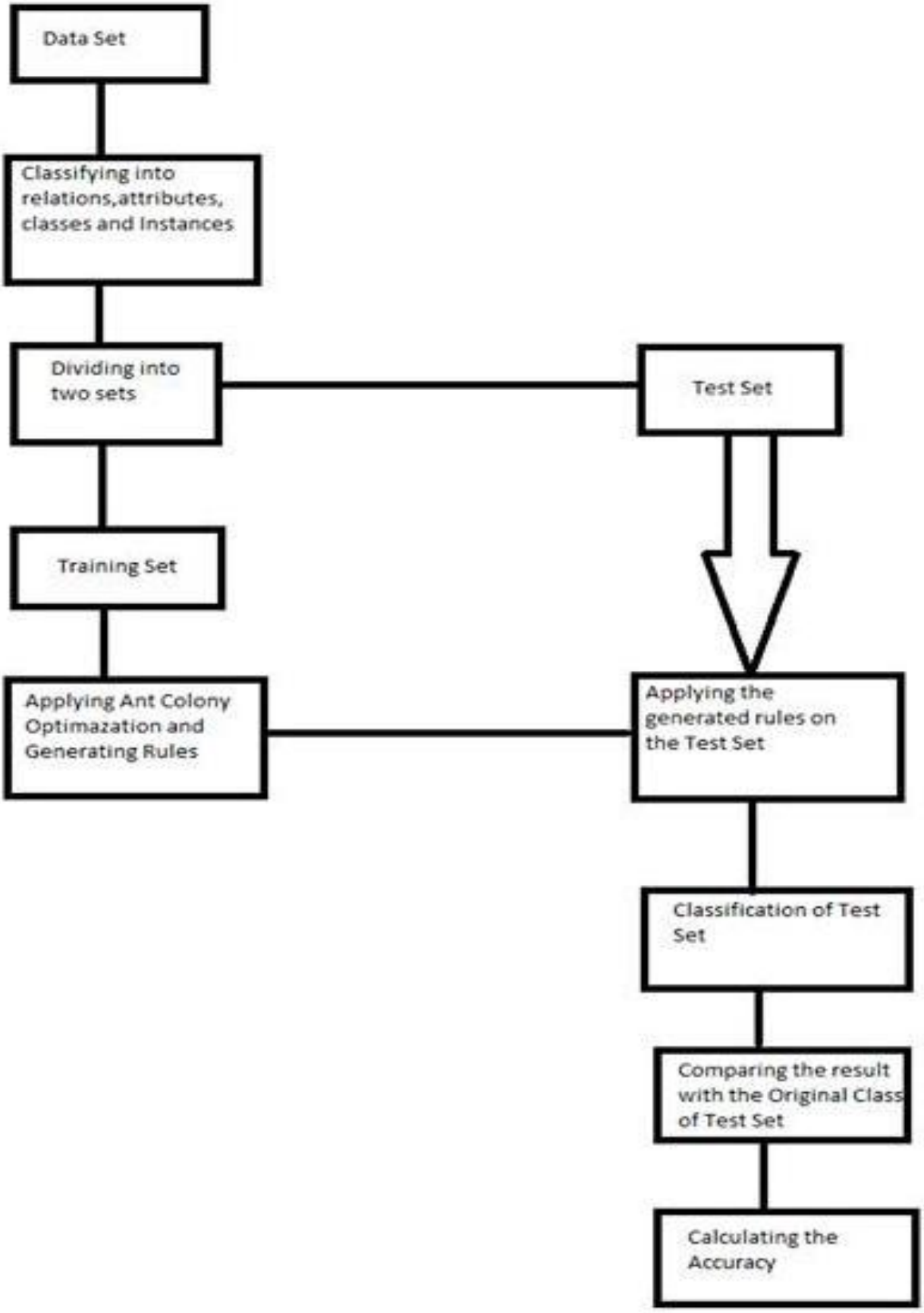
5. Dataset is divided into two sets, namely, training set and test set. For this, fivefold cross-validation is implemented, i.e. four-fifth of the dataset is used as training set while one-fifth of the dataset is used as test set.

6. Ant Miner Algorithm is implemented on the training set, and certain rules are generated based on the attribute-value pairs of the training set.

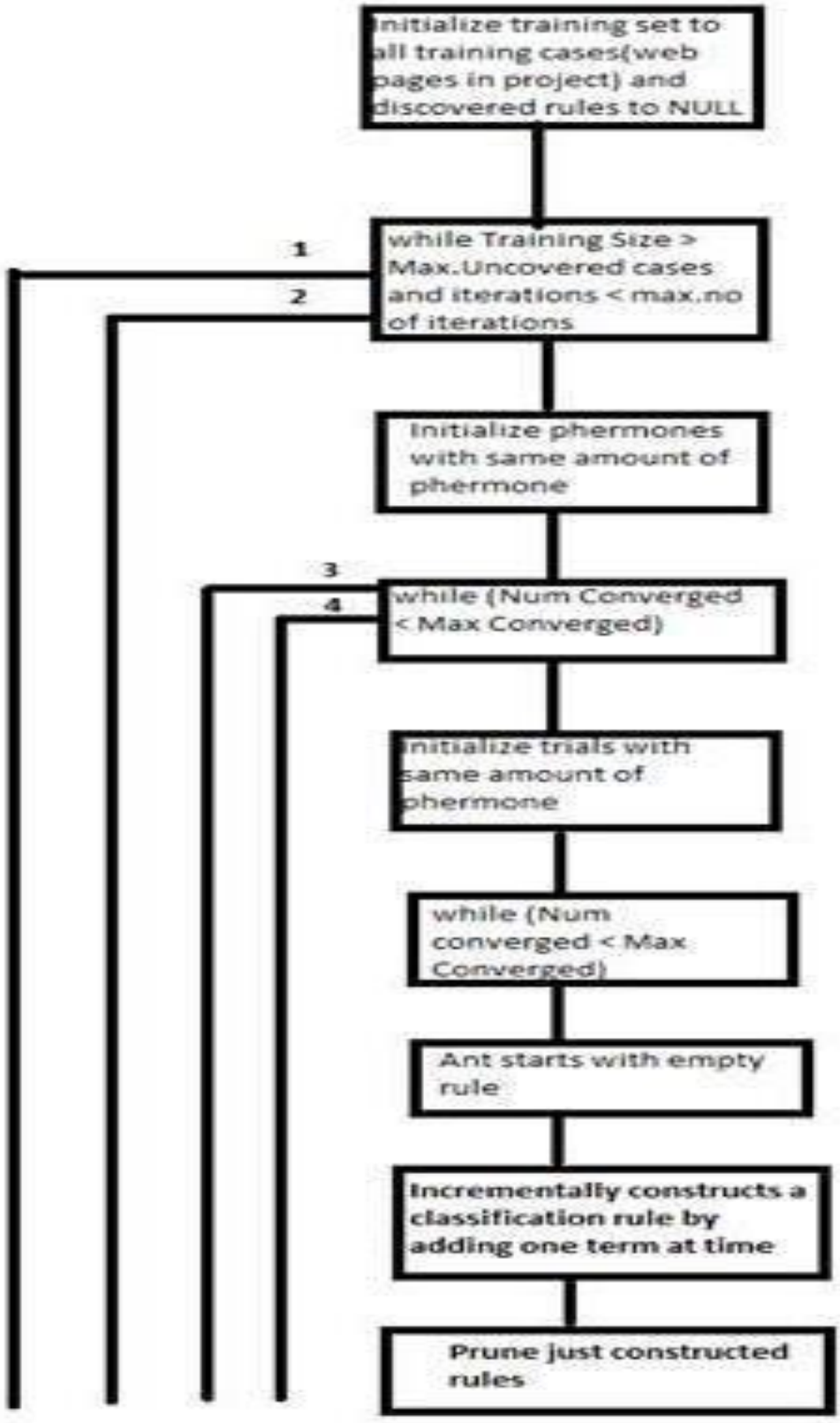
7. The rules generated are applied to the test set, and the test set is classified into its classes.

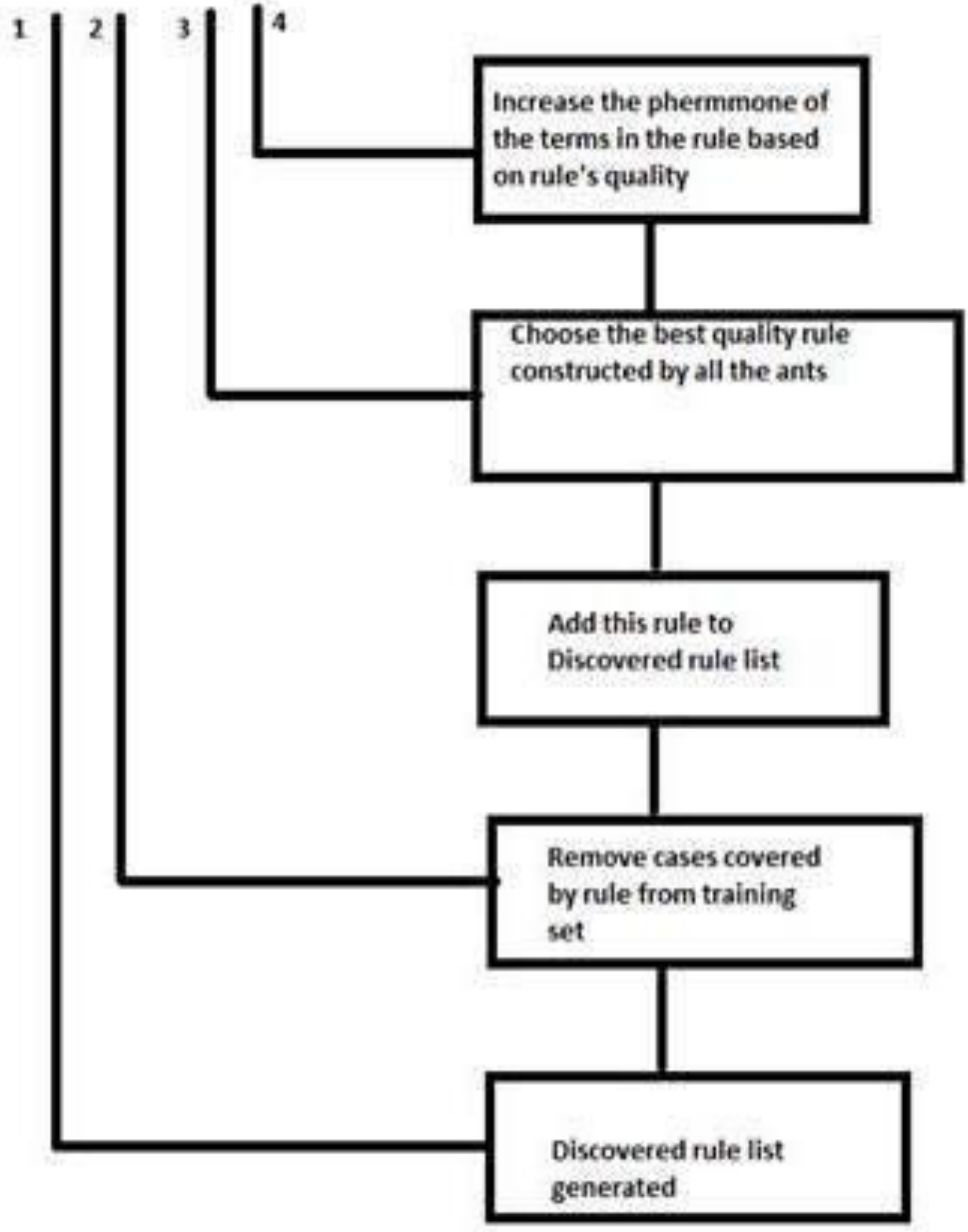
8. Proposed classes are then compared to the actual classes and hence accuracy is calculated.

METHODOLOGY BLOCK DIAGRAM



ANT MINER BLOCK DIAGRAM




1
2
3
4
Increase the phermmone of the terms in the rule based on rule's quality
Choose the best quality rule constructed by all the ants
Add this rule to Discovered rule list
Remove cases covered by rule from training set
Discovered rule list generated

HARDWARE AND SOFTWARE USED

- Programming Language: **JAVA**
- IDE : **Eclipse Luna**
- Libraries: **JDK 1.8**
- Datasets: **Standard BBC dataset**

RESULTS AND DISCUSSION

Attempting to classify information on the web is challenging task. The Ant Miner algorithm has proved again to be a powerful classification tool. Not only this but it presents knowledge in a much more compact and comprehensible way. Another alteration that would be useful for the Ant Miner algorithm for using in the web mining field is the ability to tolerate larger amounts of attributes. This is the largest problem facing the system described which is why so much effort has been put into trying to solve it, with only limited success.

Classification rules have been generated from the training set and those rules are applied on the test set. This classification of test set is then compared with the actual classes of those pages as mentioned in the dataset. Thus, Classification Accuracy is determined.

FUTURE SCOPE OF THE WORK

Currently, the great majority of problems attacked by ACO are static and well defined combinatorial optimization problems, that is, problems for which all the necessary information is available and does not change during problem solution. For this kind of problems ACO algorithms must compete with very well established algorithms, often specialized for the given problem. Also, very often the role played by local search is extremely important to obtain good results. Although rather successful on these problems, we believe that ACO algorithms will really evidentiate their strength when they will be systematically applied to “ill-structured” problems for which it is not clear how to apply local search, or to highly dynamic domains with only local information available.

A first step in this direction has already been done with the application to telecommunications networks routing, but more research is necessary.

REFERENCES

1. I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools with Java Implementations*,
Morgan Kaufmann Publications, 2000.
2. U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth. From data mining to knowledge discovery:
an overview. In: U.M. Fayyad et al (Eds.) *Advances in Knowledge Discovery and Data*
Mining, 1-34. AAAI/MIT, 1996.
3. R.S. Parpinelli, H.S. Lopes and A.A. Freitas. Data Mining with an Ant Colony Optimization
Algorithm. *IEEE Trans. on Evolutionary Computation, special issue on Ant Colony algorithms*, 6(4), pp. 321-332, Aug. 2002.

4. R.S. Parpinelli, H.S. Lopes and A.A. Freitas. An Ant Colony Algorithm for Classification

Rule Discovery. In: H.A. Abbass, R.A. Sarker, C.S. Newton. (Eds.) *Data Mining: a Heuristic*
Approach, pp. 191-208. London: Idea Group Publishing, 2002.