Web page Classification Using Ant Colony Algorithm

Pan Ei San, Nilar Aye
University of Computer Studies, Yangon
paneisan1985@gmail.com

Abstract

In this paper we describe the new classification algorithm for web page classification that is ant colony optimization algorithm. The algorithm's aim is to solve for discrete problem and discreteness of text documents' features. In this paper, the system consists of two parts for classification: training processing and categorizing processing. In training process, the system removes the unnecessary part of the web page in preprocessing step. After preprocessing step, each text is represented by vector space model using TF-IDF formula. In the categorizing process, the testing web page is tested to classify appropriated class label by using ant colony algorithm. Ant colony algorithm works to find the optimal path or optimal class for text features by matching during iteration in the algorithm. Our proposed system is more robust and flexible than other traditional machine learning because it is based on swarm intelligence behaviors. The satisfactory accuracy of classification will get in this proposed system.

1. Introduction

Over the past decade we have witnessed an explosive growth on the Internet, with millions of web pages on every topic easily accessible through the Web. The Internet is a powerful medium for communication between computers and for accessing online documents all over the world but it is not a tool for locating or organizing the mass of information. Tools like search engines assist users in locating information on the Internet. They perform excellently in locating but provide limited ability in organizing the web pages. Internet users are now confronted with thousands of web pages returned by a search engine using simple keyword search. Searching through those web pages is in itself becoming impossible for users. Thus it has been of more interest in tools that can help make a relevant and quick selection of information that we are seeking. Web page classification can efficiently support diversified application, such as web mining, automatic web page categorization, information filtering, search engine and user profile mining. It describes the state of the art techniques and subsystems used to build automatic web page classification of the web pages. If all features of web pages are used in the representations, the number of dimensions of the vectors will usually be very high (hundreds of thousands). To reduce both time and space for computation, various methods are introduced to reduce the dimensionality. When a web page needed to be classified, the classifiers use the learned function to assign the web page to categories. Some classifiers compare the similarity of the representation of the web page to the representations of the categories. The category having the highest similarity is usually considered as the most appropriate category for the assigning the web page.

Ant Colony Optimization (ACO) is a relatively new computational intelligence paradigm inspired by the behavior of natural ants [3]. Ants often find the shortest path between a food source and the nest of the colony without using visual information. In order to exchange information about which path should be followed, ants communicate with each other by means of a chemical substance called pheromone. As ants move, a certain amount of pheromone is dropped on the ground, creating a pheromone trail. The more ants follow a given trail; the more attractive that trail becomes to be followed by other ants. This process involves a loop of positive feedback, in which the probability that an ant chooses a path is proportional to the number of ants that have

already passed by that path. Hence, individual ants, following very simple rules, interact to produce an intelligent behavior at the higher level of the ant colony. In other words, intelligence is an emergent phenomenon. In this article we present an overview of Ant-Miner, an ACO algorithm for discovering classification rules in data mining [6], as well as a review of several AntMiner variations and related ACO algorithms. All the algorithms reviewed in this article address the classification task of data mining. In this task each case (record) of the data being mined consists of two parts: a goal attribute, whose value is to be predicted, and a set of predictor attributes. The aim is to predict the value of the goal attribute for a case, given the values of the predictor attributes for that case.

2. Related Works

Rafael S.Parpinelli, Heitor S.Lopes and Alex A.Freitas[9] proposed an ant colony optimization (ACO) algorithm, for the classification task of data mining. In this task, the goal is to assign each case (object, record, or instance) to one class, out of a set of predefined classed, based on the values of some attributes (called predictor attributes) for the case. In the content of the classification task of data mining, discovered knowledge is often expressed in the form of IF-THEN rules, as follows: IF<conditions> THEN <class>.The rule antecedent (IF part) contains a set of conditions, usually connected by a logical conjunction operator (AND). They referred to each rule condition as a term, so that the rule antecedent is a logical conjunction of terms in the form IF term1 AND term2 AND.... Each term is a triple <attribute, operator, value>, such as <Gender=female>.The rule consequent (THEN part) specifies the class predicted for cased whose predictor attributes satisfy all the terms specified in the rule antecedent. From a data-mining viewpoint, this kind of knowledge representation has the advantage of being intuitively comprehensible for the user, as long as the number of discovered rules and the number of terms in rule antecedents are not large.

Nicholas Holden and Alex Freitas[5] utilized 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 their experiments. It also investigates the benefits and dangers of several linguistics-based text preprocessing techniques to reduce the large numbers of attributes associated with web content mining. Ant- miner starts by initializing the training set to the set of all training cases (web pages, in this project), and initializing the discovered rule list to an empty list. Then it performs an outer Repeat-Until loop. Each iterations of this loop discovers one classification rule. This first step of this loop is to initialize all trails with the same amount of pheromone, which means that all terms have the same probability of being chosen (by the current ant) to incrementally construct the current classification rule. In this paper, Ant-Miner produces accuracies that are at worst comparable to the more established C5.0 algorithm; and (b) Ant-Miner discovers knowledge in a much more compact form than C5.0, facilitating the interpretation of the knowledge by the user.

Mehdi Hosseinzadeh Aghdam, Nasser Ghasem-Aghaee and Mohammad Ehsan Basiri[7] proposed a major of text categorization is the high dimensionality of the feature space; therefore, feature selection is the most important step in text categorization. The authors presented a novel feature selection algorithm that is based on ant colony optimization. Ant colony optimization algorithm is inspired by observation on real ants in their search for the shortest paths food source. This proposed algorithm was easily implemented and because of use of a simple classifier in that, its computational complexity is very low. The performance of the proposed algorithm is compared to the performance of information gain and CHI algorithms on the task of feature selection in Reuters-21578 dataset.

Esra Sarac and Selma Ayse Ozel [3] proposed a new ant colony optimization based feature selection for web page classification. In this paper, the aim of the system was to reduce the number of features to be used to improve runtime and accuracy of the classification of web pages. In this paper, they used an ant colony

algorithm to select the best features, and then they applied the well -known C4.5, naïve bayes and k nearest neighbor classifiers to assign class labels to web pages. They used the WebKB and conference datasets.

Allen Chan and Alex A. Freitas [1] proposed a new ant colony algorithm for the multi-label classification task. The new algorithm, called MuLAM (Multi-Label Ant-Miner) is a major extension of Ant-Miner algorithm, the first ant colony algorithm for discovering classification rules. MuLAM obtained predictive accuracies considerably better than the predictive accuracies obtained by the simple majority classifier and by C5.0.

3. Background Theory

3.1. Web Page Classification

Web pages are different from text, and they contain a lot of additional information, such as URLs, links, HTML tags such as script, which are not supported by text documents. Because of this property of web pages, web classification is different from traditional text classification. A major problem of the web page classification is the high dimensionality of the feature space. We need to select "good" subsets of features from the original feature space to reduce the dimensionality and to improve the efficiency and run time performance of the classification process. The classification process is vital to the efficiency of the overall system, as only relevant pages will then be considered by the extraction process, thus drastically reducing processing time and increasing accuracy. The underlining technique used for our classifier is based on the ACO algorithm, due to the independence noticed in the data corpus analyzed [4].

3.1.1. Web Page Representation

The first step in web page classification is to transform a web page, which typically composes of strings of characters, hyperlinks, images and HTML tags, into a feature vector. This procedure is used to remove less important information and to extract salient features from the web pages. The subject-based classification prefers features representing contents of subjects of web pages and these features may not represent genres of the web pages. There are presented different web page representations for the two basic classification approaches [10].

I. *Representations for Functional classification*

It is based on an analysis of the unique business functions and activities of an organization, but is independent of the organization's administrative structure. This makes functional classification more flexible and stable as business units and divisions are likely to change over time. It also promotes effective information sharing in the organization, with the 'ownership' of files shared across different business units. Functional classification is used not only for titling and retrieval purposes, but it can also help define access and security restrictions and determine retention periods for records. This can be achieved by aligning classification tools such as Business Classification Schemes (BCS) and functional thesauri to other tools, such as a security classification scheme and a Retention and Disposal Schedule.

II. Representations for Subject Based Classification

Most work for subject-based classifications believes the text source (e.g. words, phases, and sentences) represents the content of a web page. In order to retrieve important textual features, web pages are first preprocessed to discard the less important data.

In our proposed system, it is used the subject based classification representation because the system classifies the different category such as health, sport, entertainment that based on the text sources of the content of web pages.

3.2. Ant Colony Optimization

Ant Colony Optimization [8] is typically used to solve minimum cost problems. We may usually have N nodes and A undirected arcs. There are two working modes for the ants: either forwards of backwards. The ant's memory

allows them to retrace the path it has followed whiles searching for the destination node before moving backward on their memorized path, they eliminate any loops from it. While moving backwards, the ants leave pheromones on the arcs they traversed. At the beginning of the search process, a constant amount of pheromone is assigned to all arcs. When located at a node i an ant k uses the pheromone trail to compute the probability of choosing j as the next node:

$$p_{ij}^{k} = \begin{cases} \dfrac{\tau_{ij}^{\alpha}}{\sum_{l \in N_i^k} \tau_{il}^{\alpha}} & if\ j = N_i^k \\ 0 & if\ j \neq N_i^k \end{cases} \qquad (1)$$

When the arc (i,j) is traverse, the pheromone value changes as follows: By using this rule, the probability increases that forthcoming ants will use this arc. After each ant k had moved to the next node, the pheromones evaporate by the following equation to all the arcs: Steps for solving a problem by ACO [6].

a) Present the problem in the form of sets of components and transitions, or by a set of weighted graphs, on which ants can build solutions.
b) Define the meaning of the pheromone trails
c) Define the heuristic preference for the ant while constructing a solution
d) If possible implement an efficient local search algorithm for the problem to be solved.
e) Choose a specific ACO algorithm and apply to problem being solved.
f) Tune the parameter of the ACO algorithm.

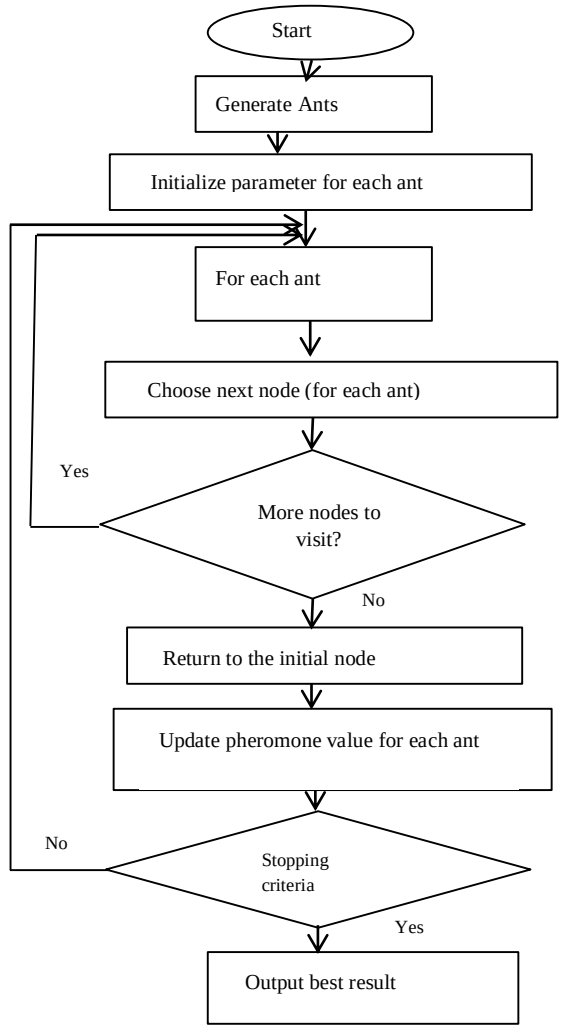

Figure1. Overview of Ant Colony Optimization

4. Proposed System

In our proposed system, there are two parts of the system. They are training process and categorizing process. The training process includes the preprocessing step, calculating the weight of each term for training database by using vector space model. The categorizing process includes the same as training process such as preprocessing and weighting for term and then uses the new classifier (Ant Colony algorithm) to classify the web page and assign the appropriate class or category. The detailed

explanations step by step are described in follows:

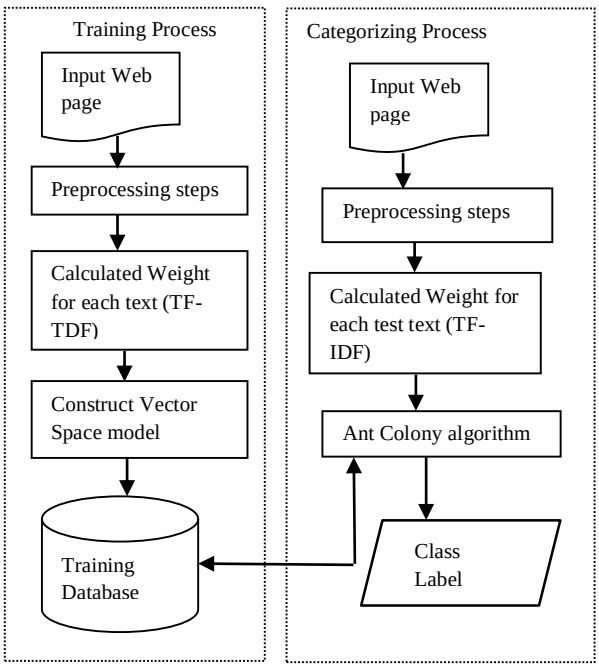

Figure2. Proposed System for web page classification

4.1. Preprocessing Steps

The preprocessing consists of the following steps:

(i) Removing HTML tags: HTML tags indicate the formats of web pages. For instance, the content within <title> and </title> pair is the title of a web page; the content enclosed by <table> and </table> pair is a table. These HTML tags may indicate the importance of their enclosed content and they can thus help weight their enclosed content. The tags themselves are removed after weighting their enclosed content.

(ii) Removing stop words: stop words are frequent words that carry little information, such as prepositions, pronouns and conjunctions. They are removed by comparing the input text with a "stop list" of words.

(iii) Removing rare words: low frequency words are also that rare words do not contribute significantly to the content of a text. This is to be done by removing words whose number of occurrences in the text are less than a predefined threshold.

(iv) Performing word stemmed: this is done by grouping words that have the same stem or root, such as computer, compute, and computing. The Porter stemmer is well-known algorithm for performing this task. After the preprocessing we select features to represent each web page.

4.2. Vector Space Model for texts

Term Frequency (TF): Term frequency known as TF measures the number of times a term (word) occurs in a document.

Inverse Document Frequency (IDF): The main purpose of doing a searching is to find out relevant documents matching the query. In the first step all terms are considered equally important. In fact certain terms that occur too frequently have little power in determining the relevance. It is to weigh down the effects of too frequently occurring terms. Also the terms that occur less in the document can be more relevant. The system weighs up the effects of less frequently occurring terms.

$$IDF(term) = 1 + \log_e\left(\frac{total\ number\ of\ Documents}{Number\ of\ Documets\ with\ term}\right) \quad (2)$$

*TF*IDF*: For each term in the test document multiply its normalized term frequency with its IDF on each document.

Vector Space Model (Cosine Similarity): From each document the system derives a vector. The set of documents in a collection then is viewed as a set of vectors in a vector space. Each term will have its own axis. Using the formula given below the system finds out the similarity between any two documents.

$$\text{Cosine Similarity} = S(i,j) = \cos(i,j) = \frac{\sum w_i . w_j}{\sum w_i^2 . \sum w_j^2} \quad (3)$$

where, w is the weight of each term.

4.3. Ant Colony Algorithm for Classification

A feature term of test documents is regarded as a node in the algorithm. All ants are divided into several clusters. Ants which have same category information in the same cluster traverse all of the nodes. The numbers of the ants in one type

colony determine ants crawling iterations of this type. A class path $I_k$which can describe the optimum of this class will be generated after some a type of ants K have completed all the nodes. The classification result will come out by comparing pheromone concentrations b in their own roads I_k after all of ant's iteration. The classification k described by the road $I_k$that has the max pheromone concentration is the class of this text. There need the three steps

a) *Determination the next node of the road*

The next node of the road is determined by both similarity and transition probability of current node. The similarity of a node can be calculated using formula (3) above, the transition probability can be calculated using formula (4) as follows.

$$P_{ij} = \frac{\tau_j}{\sum_\tau \tau_i} \quad (4)$$

Where, P_{ij} is transition probability and τ_j is pheromone value of j and τ_i is pheromone values of node i.

b) *Calculating the pheromone to be updated after getting some a node j*

$$\tau_j = \rho\tau_j + \triangle \tau_j \quad (5)$$

The symbol $\triangle \tau_j$ is equal to w_{ij} which is the weight of term j belonging to category k above and ρ is constant value.

c) *Finding the optimal covering collection in all of them (optimal path) of categories*

Every category covering collection contains all nodes in the optimal path of this category. The optimal covering collection is the path whose similarity with text category is the closed. The similarity of every path and text category can be calculated by pheromone concentration how many pheromones in unit distance. The formula (6) is the calculating of pheromone.

$$b_k = \frac{\sum_i^n \tau_i}{n} \quad (6)$$

Where, b_k is the pheromone concentration for category k and τ is the pheromone value and n is the total number of node.

4.3.1 Classification Algorithm

Algorithm: Ant Colony Classifier

Input: Terms t of testing document D

τ0← Initialize Pheromone value ();
Cover point I← {};
*i ←0/*ant index*/*
j←0/ category index*/*

For j=1 to category m do
Create ants for sub category;
Release ant randomly;
do
i←i+1
p_{ij} ←Choose the next node ();
S← Calculate the similarity value ();
x← Calculate S×p_{ij} ;
If (x< standard value)
Then crawling of the ant is ended;
Else
Update pheromone (j);
I+={j};
End if
While (un- iterated ant in the sub category)
j←j+1;
I_k= {I_1, I_2, … , I_n};
b_k=Calculate pheromone concentration (I_k);
End for
Output: max b_k to be the best found solution or category for testing document;

In the classification process, Ant Colony algorithm has included three steps. The first step defines the ants and divides all ants into m categories. Feature terms of testing documents are hashed randomly. The second step is iteration process for m categories. In this step, every ant in ant colony of class k (a_k) traversals all nodes sequence. Pheromone value of every node is initialized into τ0 equally. Select a starting node randomly and start to crawl after releasing the pheromoneτ1.The set of cover point I is initialized into empty set φ, I= { }. The algorithm does until there is un-iterated ant in the population. The next node j is selected based on the max value of x which is product of similarity

with current node i and transition probability. For the formula of $x=S\times p_{ij}$, the value of S and p_{ij} can be calculated by formula (3) and formula (4). And then if x is less than the standard value then this ant is ended the crawling. Otherwise, accessing node j and updating the pheromone of j by formula (5). Node j is added into I and updating the covering collection I (maybe get rid of redundant node element. Getting the covering collection of class k to this document, $I_k= \{I_1, I_2, I_n\}$. The algorithm is calculating the pheromone concentration using formula (6) and loop next category. The step3 is the output process that the category (a) of the covering collection (I) accorded by Max(b_k) is the text's category.

5. Experimental Result

5.1 Data set

The test dataset from two sources: development and evaluation data sets from CleanEval competition (671 HTML documents) and The dataset gathered from several web sites (NY Times, Yahoo, BBC) (736 HTML documents). Our system involves classifying web pages in order to a number of domains as specified by the user. Also, our system belongs to the category of multi-class classification, with multiple classes or categories such as entertainment, medicine, sports, education and business. Every testing document is classified by iterative computation in training process used the classification algorithm above. Classification results are evaluated by precision and recall rate which are accepted internationally.

5.2 Performance Measure

In this system, the classification accuracy is measured by precision, recall and F1-measure. These equations are as followed.

$$\text{Precision} = \frac{categories\ found\ and\ correct}{total\ categories\ found} \quad (7)$$

$$\text{Recall} == \frac{categories\ found\ and\ correct}{total\ categories\ correct} \quad (8)$$

$$\text{F1-measure} = 2.\ \frac{Precision * recall}{Precision + recall} \quad (9)$$

6. Conclusion

We have described an approach for the classification of web pages that uses the different web pages. The results obtained are quite encouraging. This approach could be used by search engines for effective categorization of web pages. We have currently used our approach to categories the web pages into very broad categories. The same algorithm could also be used to classify the pages into more specific categories by changing the feature set.

References

[1] A.Chan and A.A.Freitas, "A New Ant Colony Algorithm for Multi-Label Classification with applications in Bioinformatics", GECCO,2006, pp27-34.

[2] B. Liu, "Introduction of Pre-processing for text classification", CS583, UIC

[3] E.Sarac and S.Ayse Ozel, "An Ant Colony Optimization Based feature Selection for web page classification" , The Scientific World Journal, Volume 2014, 16 pages

[4] G.l Fiol-Roig, M. Miro-Julia, E. Herraiz, "Data Mining Techniques for Web Page Classification", Advances in Intelligent and Soft Computing Volume 89, 2011, pp 61-68

[5] N.Holden and A. A. Freitas, "Web Page Classification with an Ant Colony Algorithm _ Ant-Miner", Computing Laboratory, University of Kent Canterbury, CT2 7NF, UK.

[6] M.Dorigo and T.Stutzle, "Ant Colony Optimization", ACM, USA,2004

[7] M.Hosseinzadeh Aghdam, N. Ghasem-Aghaee and M. Ehsan Basiri,"Application of Ant Colony Optimization for Feature Selection in Text Categorization", IEEE,2008, PP 2872-2878.

[8] P. Kaur and R. Kaur, " A survey of Optimization Algorithms for Web Page Classification", IJCST, Vol. 5, April-June 2014

[9] R. S.Parpinelli, H. S.Lopes and A. A.Freitas,"Data Mining With an Ant Colony Optimization Algorithm", IEEE , VOL. 6, NO. 4, AUGUST 2002.

[10] X.Qi and D. Davison ,"Web Page Classification Feature and Algorithms", Department of Computer Science & Engineering Lehigh University, June 2007.