ABSTRACT

Different approaches have been presented to classify web documents. The low precision of classification and low flexibility of models presented to classify web documents can however be mentioned as major problems among them. In this paper, we intend to model users’ behavior by the ant colony algorithm to classify WebPages. This method is based on the idea that if users sequentially browse some WebPages, then these pages are the same and will be classified in one group. This of-course will apply if the user does not pause on a page until the threshold time and continuously performs their survey process. If a site has several users with different tastes and interests to visit the WebPages accordingly, then the site webpages will be classified into several groups according to users’ tastes. By allocating an ant to each user in the proposed method & the amount of the pheromone existing on the users’ survey graphs manes, it will be specified that which documents are similar and lie in one group of users’ interests. Users will also be classified in cointerest groups using this same ants survey graph. The results of this study can be used in proposing and classification systems based on the documents users’ behavior or documents clustering. Simulations results have shown that the proposed method enjoys higher precision compared with previous methods.

Key words: Web mining, web application mining, ant colony algorithm, classification
Introduction

The size of available information on the web is very large and increasing year by year. According to a study conducted in 2000, over 2.1 billion unique WebPages were exposed to the public, which is increasing at the rate of 7.3 million pages per day. This is a very good piece of news. The web is ubiquitous and contains almost any type of imaginable information to use on the web. Yet, the very large volume of information is matchless and it will take a whole life time to screen information to find a subject matter one is seeking for[1]. As the size of information considerably increases and the web expands and extends, the need for methods and techniques by which to provide the possibility of data access and information extraction to users is felt more than ever. Web mining is one of research fields in which by applying data mining techniques, web services & documents are exposed to information automatic extraction and discovery. Web mining is in effect the discovery process of information and useful, unknown knowledge from web data. Based on the type of the data mined, web mining methods are classified into 3 groups: web content mining, web structure mining and web application mining[2].

Web content mining deals with useful, unknown information discovery and extraction from the web documents content. In web structure mining, it is tried to model the web links structure. That is, from each super link (inside each document), which other sources and documents can be linked to, and from where (which super links) the present document can be accessed. Web application mining concentrates on methods which can model users’ behavior at a specified time based on web application mining[3].

In WebPages classification. There is one or some predefined classes, and one classification model will allocate WebPages to one or more predefined classes. WebPages classification is in fact responsible for allocating one class or one set of predefined classes to a webpage [4]. One of the major problems of web content classification is the high dimensions of features space. We should select proper subsets of features from the main features space to reduce the features space dimensions and optimize classification efficiency and performance. Yet, the high space of features in web documents classification based on content is one of the basic problems in this type of classifications [5], [6], [7].

In [8], an ant colong algorithm has been used to classify WebPages. The aim is in fact to discover a good set of classification rules to classify WebPages under the basic subject of WebPages themselves. The classification algorithm applied, the ant miner algorithm, is the 1st algorithm of ants community used to discover classification rules[9]. In this process, ants behavior and collective intelligence are somehow modeled, and an algorithm is
proposed to search the environment and classify the data existing in it. In this method, text preprocessing and words etymology are initially performed by the lexical network. Then, by identifying the ants community model, data will be recognized and classified. And it has been shown that the ant miner is not, not only comparable with other algorithms, but also tends to find more simpler rules. The disadvantage of this method is that, due to the very many features and complication of relation between words, classification precision has dropped considerably. One of the other problems of WebPages classification based on content is the application of key words. Key words suffer from the problems of synonyms (which have different forms but the same meanings) and homonyms (having the same form yet different meanings). Besides, when a subject is new or has not yet been known to the user, the user may not be able to exactly find any related key word[10]. The second group of problems is due to the fact that the allocation of good key words to documents requires a tremendously laborious effort and special expertise: Only those who are well acquainted with this field and have well studied the document can determine the key words with a proper classification. Due to the above mentioned problems, the user behavior-web interaction has been used [11].

In [12] , the work idea is based on metadata extraction, not from documents content, but from the document application model. Since users do not randomly move among documents, and supposing that users are relatively aware of the content of the document they intend to use in their next step and that they select the next document based on their own information requirements, and if some users sequentially request for some WebPages, such pages have responded to the same information requirements, they are thus similar to each other.

By using the user’s subjective model, a virtual relation will be established between documents. This relation will not necessarily comply with documents observable relation (like documents relation based on key words). In this method, by using the analysis of the data obtained from the user’s behavior, with no use of experts, precious information has been acquired. In this method, relations between documents are modified using a method such as Heb’s law. In Boulon’s method, the practice is that for the existing documents, a relation matrix is considered in which any entry aij shows the relation between documents i and j. As the user moves from document i to document j, the link between these 2 documents will be more secure, showing the amount of similarity between these 2 documents.

In [13], this method has been used to determine the amount of similarity between documents via information on how users use the web: a learning automaton is allocated to each webpage, whose function is to learn the
amount of similarity between this page and other WebPages, and it has been shown that this method is more efficient than the method in [12] in WebPages similarity identification.

In [14], using the ant colony algorithm, the determination of documents similarity is dealt with. In the proposed algorithm, an ant is allocated to each user. As the user moves on documents, his/her corresponding ant will be moved on the documents graph. The proposed algorithm will estimate the amount of similarity between each document and other documents based on the residual pheromone from the user’s movement on the documents relations graph. The results of simulations show that the proposed algorithm identifies documents similarity better than the methods stated in [12] and [13] do.

Using the user’s behavior modeling employing pages survey information by the user in this paper, we present a new model based on the user's survey graphs, cointerest groups and the user's target page.

The benefit of the method proposed to preserve classification quality and classes variability for documents is continuous. In the meantime in this method, users lie in cointerest classifications using the appropriate survey behaviors study, the number of documents and users included in the model is variable and there is no limitation to this number. On the other hand, documents enjoying more similarity to each other will lie in one class, either.

The paper construct is as follows: we will proceed to describle the algorithm as used in the paper, and at the end, the results of simulation and conclusion will follow.

Finding the documents which are similar to each other: In users’ modeling in this paper, it is assumed that users’ survey on documents resembles the ants in search space.

Different documents are interrelated provided the user has visited these pages sequentially. Yet, if the mane between the documents in the user’s survey graph has more pheromone, there will thus be a stronger relation between those documents, which is per se a problem of finding the shortest course (the highest similarity), and besides, it is an NP-hard combination optimization problem, extensively studied. An ant colony algorithm can be used to solve the problem of users’ survey in the “ant system” method in which any (artificial) ant is considered as a user surveying on a document which has been randomly selected, and any artificial ant has a memory which saves information on its course(path) up to now( a relative relation). This is only a start point. When the ant lies on the target page(which will be described in the next section), it will cover(survey) a complete course from the survey graph by the documents probable selection to move to the next document
unless it reaches its intended target document. When ant k is on document i, it will select the unvisited document j using the probability obtained from Eq.(1)[14]

\[ P_k(i, j) = \frac{[\tau(i, j)]^\alpha [\eta j(i, j)]^\beta}{\sum_{N_k} [\tau(i, j)]^\alpha [\eta j(i, j)]^\beta} \]  

In Eq.(10, (i,j)-1, a (i,j) where d(i,j)= distance between documents i and j and shows the explored information available to the ants. N_k = “possible(viable)” neighborhood of ant k. In another word, it shows all the documents not yet visited by ant k. (i,j) is a tail value of the pheromone between documents i and j. A (pheromone) and b (explored information) are parameters which determine the relative effect of the explored information and pheromone. If α=0, the ants will perform a random, voracious search using the heuristics of the nearest neighborhood. If β=0, then the ants will each reach its target document, and the pheromone tail will be expressed based on the target updating law defined in Eq.2.

\[ \tau(i, j) = p^* \tau(i, j) + \sum_{k=1}^{m} \Delta \tau_k(i, j) \]  

where p refers to a parameter of pheromone evaporation which depletes the pheromone tail and m is the number of ants(users). A specified amount of pheromone, \( \Delta \tau_k(i,j) \), left by each ant k on the tail, has been determined using Eq.3.

\[ \Delta \tau_k(i, j) = \begin{cases} \frac{\rho \tau(i, j)}{L_k} & \text{if arc } i \rightarrow j \text{ is used by ant } k \\ 0 & \text{Otherwise} \end{cases} \]  

where L_k= travel course(length) of ant k. This implies that the shorter the ant’s travel, the more pheromone on the arcs will remain in the course and thus, this course arcs will most probably be selected in the next repeat.

The algorithm will be repeated in each of these steps unless the algorithm stop criteria are fulfilled.

**Calculation of the distance between documents**

In the users’ survey modeling problem, the distance between documents does not have a physical concept. There is however a vectorial concept between documents. Such a distance should thus be specified using specific parameters. The major approach to calculating the distance between documents is to use methods to study
documents content, the vocabulary (lexicon) existing in them based on the importance and also the number of words frequency in the document, and then, extract features using features selection, convert into features vectors, and in the end, calculate the distance between these vectors.

**Target page**

It is known that users’ movement between WebPages to perform different operations such as reading a specific page, finding specific subject in the intended field, software downloading,... This variable behavior of users is performed based on instantaneous decisions. That is, the user searches for a specific objective on a website. Documents are thus surveyed (mined) in a targeted mode. Yet, a problem which has not so far been dealt with in any relevant research is the problem of the user’s target document which refers to the documents which the user gets to start performing a specific behavior such as pausing on the document, document printing, file downloading or document reserving, and the user may afterwards, proceed to survey other pages and follow another target document. In this paper, the user’s pause on the page is used as a method to specify the target document. That is we consider that if the user pauses on a page say for example 20 s, he/she has thus achieved his/her target document. So, if the user starts a new survey, it will be indicative of search for a new target document.

Each user thus visits one document or one set of documents. In order to show this survey, we will act in the form of Eq. 4:

\[
u_i = \{p_1, p_2, ..., p_n\}
\]

\[
u_n = \{p_1, p_2, ..., p_n\}
\]  

Eq. 4 shows users’ movement between documents.

Eq.5 is used to normalize ants’ (users’) movement and the importance level of the relations between web documents:

\[
\eta_{ij} = \frac{c_{ij} \ast n}{\sum_{k=1}^{n} c_{ik}} 
\]  

(5)
where \( c_{ij} \) is the amount of the pheromone existing on the link between document \( i \) and document \( j \), \( n \) is the number of users’ survey frequency using this course, and \( \sum_{k=1}^{n} C_{ik} \) shows the total amount of the pheromone existing in the courses taken out of document \( i \) to other documents. This equation expresses the importance level of the link in the graph relative to other links existing in this graph. Here, a condition should also be defined for this Eq.

The condition in question is for the nodes in the graph, that is the graph end nodes, for this Eq. We will thus express a new condition as follows:

\[
\eta_{ij} = \frac{c_{ij} \cdot n}{\sum_{k=1}^{n} C_{ik}}, \forall c_{ik} > 0
\]  

(6)

This condition states that if the user surveys a document from the documents graph and this document does not have a link to another document, the link Eq. will thus be considered by the user only in the form of Eq. 6.

It should of course be mentioned that a hypothesis should be considered for the random survey behavior separation, and the hypothesis comes from the fact that if we assume that a user searches for special information among documents, when the user obtains his/her own target document, he/she will start to read the information or download the intended file existing in that document and/or reply to relevant questions or items.

Eq. 7 can thus be considered for documents:

\[
p_{\text{target}} = \sum_{i=1}^{t \geq n} x_i, \forall x \in X, \forall p_{\text{target}} \in P, \forall t \geq n
\]  

(7)

where \( P_{\text{target}} \) shows the user’s target document, \( x_i \) shows the observation time interval for user \( i \) who has referred to the site for web documents survey and \( t \) is the time duration of the user’s pausing on that document to read it or perform his/her other intended operations.

**Calculation of the courses (manes) pheromone evaporation level**

Based on any survey the user performs, the user leaves an amount of pheromone on the documents link courses. \( Q \) is used to show the amount of this pheromone. In order to calculate the pheromone evaporation level on each mane, the algorithm should be implemented in a parallel manner for time elapse modeling in the algorithm, that
is, at moment t=0, no mane has so far been updated. Yet, at moment t=1, manes have been updated, whose number equals that of the movement of the ants in question. For parallel modeling therefore, Eq. 8 is used:

$$\tau_{ij}(t+1) = (1-p)\tau_{ij}(t) + \Delta \tau_{ij}$$

wherein it is mentioned that at moment t+1 the level of the pheromone existing in mane i-j equals that of the pheromone evaporated from the start time to moment t plus \(\Delta \tau_{ij}\), indicating pheromone level variations in the previous times. Eq. 9 is used for \(\Delta \tau_{ij}\) calculation:

$$\Delta \tau_{ij} = Q \eta_{ij} T_{ij}$$

Eq. 1 indicates the pheromone level variations on the mane in question in unit time. In this Eq., Q shows the intended pheromone level and Tij indicates the number of users’ movements from i to j.

**The updating parallel algorithm of the mane surveyed by users**

The algorithm code sample presented in Eq. 1 will follow:

```plaintext
Step 1: Set parameters, and initialize pheromone trails
Set t=0, tij(*) = ∗,
\(n=\forall .(*)\) Time Is For Diagnosis target Page By
Hold User Time On Page;
MatNavigate Mi=[], Matrix For All Navigation For All Users
NavigateMi=(); Array For navigation To Find Each Target Page For user
M=Number User in site navigation pages(=Number of ANTS)
Step 1: For (k=1 to M)
Update nj
Update Tij(t+1) on every path
Set t=t+1
NavigateMi=[Navigate Pages i,j];
If M stop on page t=n
MatNavigate Mi=MatNavigateMi=NavigateMi;
NavigateMi=new[];
End for
Step 2: Initialize a, b
Calculate:
```

Fig. 1 Algorithm code sample presented
As seen in the code sample presented, the target document identified time for the user has been considered to be 20 s.

Eq. 10 states that 2 documents have more similarities if the pheromone amount existing in the mane course between them is more than that existing on other manes linking this document.

\[
\text{sim}(x, y) = x_{\text{pheromon} \rightarrow y} \cap y_{\text{pheromon} \rightarrow x} \\
\forall \text{sim}(x, y) > x_{\text{pheromon} \rightarrow \text{All other pages}} \cap \text{All pages} \\
\text{equation} (10)
\]

The Eq. in algorithm 11 shows that if the pheromone amount is the same between 2 documents or other documents of the links (manes) of other documents, the amount of the mane shared with this mane between different users will be studied, and the mane under study will lie in the cointerest group of the users who have the most manes sharing together and sharing with this mane.

\[
\text{IF Pheromon}(x, y) \\
x_{\text{pheromon} \rightarrow \text{All other pages}} \cap \text{All other pages} x, y-> \\
\text{equation} (11)
\]

\[
\text{IF} \\
\text{Pages - oneclass}(x, y)_{i,j} = i_{\text{PathGraph}} \cap j_{\text{PathGraph}} \\
\forall i_{\text{PathGraph}} \cap j_{\text{PathGraph}} > i_{\text{PathGraph}} \cap \text{Allusers}
\]

**Users classification in cointerest groups**

Eq. 12 is used to classify the existing users in different classifications based on interest.
(1) \[ \text{Group}^{\text{favorites}}(i, j) = \text{PathGraph}_i \cap \text{PathGraph}_j \]

\[ \forall i, j \in \text{PathGraph}_i \cap \text{PathGraph}_j \quad \text{PathGraph}_i \cap \text{UserPathGraph} \]

2) \[ \text{Group}^{\text{graphpath}}(i, j) = \text{PathGraph}_i \cup \text{PathGraph}_j \]

\[ \forall i, j \in \text{PathGraph}_i \cup \text{PathGraph}_j \quad \text{UserPathGraph} \cap \text{PathGraph}_i \cap \text{PathGraph}_j \]

3) \[ \text{NewGroup}(i, j) = \text{PathGraph}_i \cap \text{PathGraph}_j > \text{PathGraph}_i \cap \text{PathGraph}_j \cap \text{Groups}^{\text{favorites}} \]

The 1st part of Eq. 12 states that 2 users will lie in an interest class when they have the highest sharing between their surveys manes.

The 2nd part of Eq. 12 states that the cointerest class of the users’ behavior equals the sum of shared survey behaviors (manes) between the users under study, and of-course, if the number of survey behaviors (manes) has a sharing with other users which is smaller than that of the users under study.

The 3rd part of Eq. 12 expresses the time of forming the users interest new group which is indicative of the fact that if sharing between 2 users in the number of shared manes is higher that the sharing of these users with other users, a new group of users' interest should thus be established.

**Number of users cointerest classes**

The proposed model acts variably and the number of classes will depend on the number of users' different behaviors. The number of documents can have no limitation , and using the algorithm proposed here the problem of cold start will also be solved employing the method proposed in this paper.

**Assessment**

In this paper, the similarity of 2 documents is the reverse of the distance between these documents. The distance between 2 documents \( I_j \) (entry \( d_{ij} \), matrix \( D \)) or Euclidian distance has been used based on the vectors of their content , ie, \( (\text{content} \ j, \text{content} \ i) \) by Euclidian distance employing the inverse relation of document similarity and features vectors.

If there is the amount of pheromone existing in the course of the linking mane for each document, the similarity between the 2 documents will be shown by \( (d'_{ij}) \). This similarity can be displayed using Eq.
In order to show the results of the proposed algorithm, the 2 matrices (D, D') correlation of Eq. 14 can be used. Since the similarity level between 2 documents is inversely related to their content vector distance amount, at the time when the algorithm achieves its best state of performance, those 2 vectors correlation will be negative. In this relation, D = actual similarity between the documents and D' = the similarity obtained using the users' survey behavior modeling algorithm. In this state, correlation will be negative. The smaller the correlation value, the better function (performance) of the algorithm it will show.

In order to show the results of the proposed algorithm, the 2 matrices (D, D') correlation of Eq. 14 can be used. Since the similarity level between 2 documents is inversely related to their content vector distance amount, at the time when the algorithm achieves its best state of performance, those 2 vectors correlation will be negative. In this relation, D = actual similarity between the documents and D' = the similarity obtained using the users’ survey behavior modeling algorithm. In this state, correlation will be negative. The smaller the correlation value, the better function (performance) of the algorithm it will show.

\[ C_i = \begin{bmatrix} c_{w_1} & c_{w_2} & \cdots & c_{w_M} \end{bmatrix} \]

\[ d_{ij} = \sqrt{\sum_{k=1}^{M} (c_{w_k} - c_{w_k}')^2} \]

\[ d_{ij}' = \tau(i, j) \]

(13)
Simulation results

For simulation, the model proposed in [16] is used to generate the set of users' documents and movements. In this model, users' interests profile and their movement strategy and motivation have been modeled using statistical distributions. By changing these statistical distributions parameters, one can generate users with different interests, motivations and strategies. We have designed a statistical software and installed it on a website to extract users' survey information. After 15 days, information was collected and analyzed. Using words frequency and weighting methods, these pages were analyzed, reviewed and labeled for classification in different classes. 10000 pages of documents were surveyed in this study. In this paper, the results were compared with the method in [14].

Fig.2 has compared the 2 methods performance level by correlation.

Fig.3 displays the proposed method and the 2 ant colony methods applied in [14] in terms of correlation comparison. The problem of the classes' and documents' numbers' being static and the cold start in [14] has been resolved in the proposed method. The Fig. results show the better performance of the algorithm proposed in this paper.

Fig.3: Similarity matrix correlation
Conclusion

This paper presents a new algorithm enabling to find similar documents and classify them in similar classes using the users' behavior survey criterion employing the ant algorithm. In the proposed method, problems such as a certain number of documents, document labeling and users' interest document which are used in the proposing systems have been dealt with, and the results show the better performance of the proposed method compared with the previous ones. In this method accordingly, the problem of cold start, which is one of the problems in classifying and proposing systems, has been well resolved.

References