Research on Information Retrieval System Based on Ant Clustering Algorithm

Peiyu Liu, Zhenfang Zhu, Lina Zhao
School of Information Science and Engineering, Shandong Normal University, Shandong Ji'Nan 250014 , China
liupy@sdnu.edu.cn, zhuzhfyt @163.com, pujingna2000 @126.com

***Abstract*—Internet is more and more widely used, which provide a valuable information resource for users. How retrieve the information users prefers rapidly and accurately become the focus nowadays. With introducing ant-based clustering and sorting, it makes a more precise and rapid clustering result. Consequently, it increases the speed and efficiency of information retrieval.**

***Index Terms*—ant-based clustering, clustering result, information retrieval**

I. INTRODCUTION

Information technology has infiltrated our society in every corner of life， and affects every person's life and work and study. We have already entered the times of "information explosion" with ever-growing popularity of the internet. The field of information technology is a vibrant and strong vitality of the area. Information retrieval has increasingly become an important research direction. INTERNET information retrieval can help users find the content of their own interest, and information retrieval is on the basis of information filtering, which has been the key technology for information filtering system laid the foundation for the study.

There are many information retrieval technologies in recently year. Traditional text retrieval: full text scanning, the inverted file, signature files, as well as clustering. The new information retrieval methods which introduce semantic information of the: natural language processing (NLP), Latent Semantic Indexing (LSI), the Neural Network. Text classification is a higher value and pivotal technology in information retrieval technologies, which is a beneficial tool to organize and manage documents.

Clustering Analysis: the similar objects aggregate clusters, the objects have the greatest similarity as far as possible within the clusters, and objects in different clusters have the greatest dissimilitude. Clustering is a process that a group of physical or abstract objects dispart a similar target group consisting of a number of clusters. The method in this paper based on cluster hypothesis: closely linked documents have relation with the same documents for requirements .Document retrieval process can be accelerated by clustering similar. However, the disadvantage is that clustering method clustering spending too much time and reducing efficiency.

Ant Algorithm has been proposed by the Italian scholar Dori_go, Maniezzo and others, in earlier stage of the 1990s of **20** century [1]. Ant Colony Algorithm is that footpath way carries out to imitate reaches one kinds of bionic algorithm on Nature ant .In order to explains the principle of ant colony algorithm , firstly brief introducing the process that the ant searches for the food : During the period of the ant group finds food, they always can find the optimum route between den and food. This is because that they are able to release out one kinds of pheromone on the route when seeking route ,While they meet a crossing not having go by choose a route randomly moving forward. Now and then release out pheromone relevant to length of route meanwhile. the longer route , the lower concentrations of hormones released, probability only is capable to do the higher thickness route choosing hormone while ant in the afterwards meets this crossing once again relatively more, such has formed a positive feedback. concentrations of hormones of optimum route is more and more big but concentrations of hormone at other route was reduce as time goes by, ultimately whole ant group is able to find out optimum route. The ant can also adapt to change of environment, when the ant group moves forward that obstruction appears abruptly on the route, the ant is able to find optimum route quickly again . In process of whole searching route, the effect through hormone although single ants choice ability is limited, the information being exchanging route between whole ant group, ultimately find out optimum route .According to this principle of work, It is one kind of late-model optimization method, it is not dependent on concrete problem mathematics optimizing an ability describing, having an global optimization ability. Ant colony algorithm has been successfully applied to the circuit design, digital data analysis, Text Mining and network data packet routing, and other fields. this paper will be discussed in the text Ant clustering algorithm classification and the study of information retrieval and application.

II. TEXT PRESENT BASED ON VECTOR SPACEMODEL

Vector space model basic thought that an article expressed as vector ,namely（W1，W2，W3…，Wn），

Conesponding author: Peiyu Liu.

doi:10.4304/jsw.4.9.1032-1036

Wi are mansion i feature item weight value , the feature item of an article expresses character and word etc .Therefore after segment to the text, an article is expressed for the set of the word. In the early stage if an article embraces the word that the vector dimension of article is 1, otherwise is 0 .this method has no way to come to express the feature item accurately in an article. Therefore, because commonly used method is the text of that text-based feature vector space model. An article expressed as a vector, the dimension of the vector feature of the set is the number of the vector should be on the characteristics of each component in this article in the weights. Hypothetical documentation set D={d_i},|D|=S, the set of characteristics T={t_j},|T|=M. t_i feature of the definition in the document t_j in the weight Wij:

$$w_{ij} = \frac{tf_{ij}}{df_{ij}}, 1 \le i \le S, 1 \le j \le M \tag{1}$$

Among them, tfi said in the document di appeared in the number of feature vector ti . dfi is the number of feature item t_j in documents D. Thereby, establishing
document vector space model, the t_1, t_2, t_3,… t_M to coordinate axis, the document d_i is presented M-dimensional vector abstract objects dispart a similar target group consisting of a number of clusters. The method in （w_{i1},w_{i2},…w_{iM}）.

The information in the Internet is organized half-formatted Web pages, documents in the form of HTML format. The expression between HTML-formatted data and text mode data is different in which there are two factors influenced the weight of feature items[6]. One is the frequency of the feature item appearing the HTML document , and the other is the location of the feature in the document. Because the HTML file contains many markers, different markers contain different information. Therefore, in order to express information on the Internet more accurately, markers are defined:

N={TITLE,H1,H2,H3,H4,H5,H6,B,U,I,Meta}

Weight Set: $W = \{W_k | k \in N\}$

$$w_{ij} = \frac{\sum_{k \in N} W_k \times tf_{ij}^{\ k}}{\sum_{k \in N} W_k \times df_j^{\ k}} \tag{2}$$

Among them, w_k is marker k's weight, and $W_{TITLE} > W_{H1} > W_{H2} > ... > W_{Meta}$; $tf_i^{\ k}$ is the frequency of feature i in the marker k.

In segmentation, the improved scanning method is used. The process as follow: firstly ,some no practical meaning words are eliminated according to the inactive Chinese word list; secondly, some obvious feature words are recognized and splited among the character string that is about to analyze; third, the obvious features of the word as breakpoints, the original string divided into smaller sub-string and then mechanical cut, which can reduce rate of error matching.feature selection using information gain; select some words having the highest information gain as feature, thereby reducing the total number of features，defined as follows:

$$\begin{aligned} IG(t) = & -\sum_{i=1}^{m} p(c_i) \log p(c_i) + \\ & p(t) \sum_{i=1}^{m} p(c_i | t) \log p(c_i | t) + \\ & p(\bar{t}) \sum_{i=1}^{m} p(c_i | \bar{t}) \log p(c_i | \bar{t}) \end{aligned} \tag{3}$$

Among them: *IG(t)* for the whole of the term reflects the amount of information provided by the classification; $p(\bar{t})$ is not a word t that the probability; $p(c_i|t)$ described *t* is the word appears in the text under the category of c_i probability, $p(c_i|\bar{t})$ is not the word *t* that the circumstances of the text c_i the probability category.

III. CLUSTER ANALYSIS BASED ON THE PRINCIPLE OF ANT HEAP

At present, the mainly clustering methods are K-means, fuzzy clustering, the clustering based on genetic algorithms, wavelet transform clustering algorithm and the result of the effective combination of improved method. Following introduces principle of various clustering algorithm respectively:

（1）K-means algorithm : In 1967, MacQueen have brought forward K-means cluster-algorithm firstly (K-means algorithm). So far, many clustering working all chooses that classics algorithm. Owing algorithmic core thought is to find out K clustering centre c_1 , c_2, ,c_k , will do every number point x_i and clustering centre c_v that the square distance is minimized (square distance to be called deviation D)

Advantage of K-means algorithm: it can carry out high efficiency classification on large-scale data collection, whose computational complexity is be O (tKmn) , among them, t is iteration numbers , K is a clustering number , m is a feature attribute number , n is objects number of classification, commonly K , m , t <<n , When clustering to large-scale data, the K-means algorithm is much quicker than hierarchical clustering algorithm . Disadvantage: be able to end when gaining a local optimum value generally; be suitable for only to numerical value clustering; clustering result that apply to the convexity data collection(class fascicles is convexity)

(2) Fuzzy Clustering: In 1969, Ruspini have applied fuzzy set theory to cluster analysis firstly , have brought forward the fuzzy clustering algorithm (fuzzy c-means , for short FCM). The FCM algorithm is one of that the image segmentation using most method , its success to owe to mainly for belonging of the image pixel introduced into fuzziness. Comparing with fragile

(crisp) or tough division method, FCM is able to reserve more initial image information. But, a shortcoming of FCM is that take no account of any space information in image context, so it is very sensitive to noise and artificial image. People have carried out a great quantity studying to FCM algorithm. In2006, people suggests that a new fuzzy clustering algorithm NFWFCA based on feature weighting. Both tradition fuzzy K-means algorithm, K-modes algorithm and K-prototype algorithm assume that every dimension of sample vector is identical to clustering contribution. But in actual application, every dimension of sample vector comes from different sensors and existence measures difference such as accuracy and reliability, every sample vector of sample vector is not identical to clustering effect . Take fuzzy K-prototype algorithm as basis, algorithm NFWFCA adopt the ReliefF algorithm to ascertain every feature weight, computational method of numerical feature weighting is:

$$\lambda^r = \lambda^r - \frac{diff\_hit^r}{R} + \frac{diff\_miss^r}{R} \tag{4}$$

attribute feature weighting reckoning is:

$$\lambda^c = \lambda^c - \frac{diff\_hit^c}{R} + \frac{diff\_miss^c}{R} \tag{5}$$

The experiment results shows that this algorithmic clustering result should be more accurate and high-effect comparing with tradition fuzzy K-means algorithm, K-modes algorithm and K-prototype algorithm. At the same time, an algorithm can analyze the contribution degree of every dimension feature to clustering, and effective go along feature extracts and selective preference, that this is of certain significance to studied and their application to cluster algorithm.

(3)The clustering based on genetic algorithms: genetic algorithm is a complicated optimization problem draw having that biosphere natural selection and evolution mechanism develop one kind finding the solution with self-adapting and self-organization algorithm. That its main merit is to simple, popularity , robust and is suitable to parallel processing, is a solved model having nothing to do with problem. Genetic algorithm is one kind of the broad global optimization method. It is operation being in progress to much individual made up of population, use the genetic operator to exchange information of each other individual, individual of in group can be evolved, and approach the optimum solution step by step. K-means algorithm can overcome disadvantage of sensitivity since genetic algorithm global optimization a function , people begins to use genetic algorithm to do clustering.

(4) Wavelet clustering: wavelet is one kind of signal processing technology, it becomes the signal decomposition different frequency wave band , the n times making use of wavelet transforming, the wavelet model can be used for n dimension signal. During the process of the wavelet transforming, the data reserves relative among object distance with the arrangement of ideas varying in different resolution ratio , this make data clustering becomes especially easy to differentiate , seeks high density area , to ascertain clustering in new space. Sheikholeslami,chatteriee and zhang put forward to wavelet clustering , which is one kind of multi-resolution clustering algorithm , firstly impose on many dimensions grid structure by depending on the data space assembling data , and then adopt wavelet transforming to transformed the original feature space and find density area in transformed space.

With the rise of ant colony algorithm research, it was found that in certain areas, using ant colony model clustering closer to the actual problem of clustering. Clustering method based on the ant colony algorithm can be divided into four principles: (1) adopting principles of ants looking for food, data clustering is realized by pheromone[4], (2) using self-aggregation behavior of ants,(3) data clustering based on the principle of ant heap,(4) according to ant mounds classification model, date clustering is realized by chemical identification system of ants.

A. Clustering principle based on ant-foraging

The ant- foraging process can be included search for food and carry food. But every ant may release pheromone on the route in the process of motion , moreover can perception plain intensity of pheromone, the stronger the pheromone intensity ,the higher the probability that ant can choose that route , the more the ant going by ,while pheromone may be volatilization as time goes by , the behavior of entire ant colony comes into being a positive feedback effect according to the intensity of pheromone .

Basic thought that the ant choose what route to realize clustering analysis according to the intensity of pheromone. Look upon the data as the ant having the different attribute, data clustering process is the process of ant seek food source also. Assume that the data clustering objects is: $X = \{X_{i1}, X_{i2}, ...X_{in}), i = 1,2..N\}$, initialization of algorithm firstly , pheromone of each route are set to 0 , namely $T_{ij}(0) = 0$, the radius of a cluster is r , statistics error is ε and so on , weighted Euclidean distance between X_j and X_i is d_{ij}, pheromone T_{ij} on every route:

$$T_{ij}(t) = \begin{cases} 1, & d_{ij} \Leftarrow r \\ 0, & d_{ij} > r \end{cases} \tag{6}$$

Probability of object X_i merge into the X_j :

$$p_{ij}(t) = \frac{T_{ij}^{\alpha}(t)\eta_{ij}^{\beta}(t)}{\sum_{\varepsilon s} T_{sj}^{\alpha}(t)\eta_{sj}^{\beta}(t)} \tag{7}$$

$p_{ij}(t)$ greater than threshold value p_0 , X_i is merged into X_j .

The choice of α, β will severely affect running efficiency of algorithm and result clustering, which can adopt different methods to avoid sinking into algorithm local convergence according to different condition. But in actual computation, under the condition that circulation number of times presets leading to algorithmic finding the solution is inefficient, convergence is bad, the optimum solution is difficult to find.

B. Clustering based on the principle of ant heap

The phenomenon of ant heap clustering is the dead are moved by ergates to form the dead ants heap.That is, a small ant heap become bigger gradually by attracting more ergates moving dead ants , of which the positive feedback will lead to the ant heap more and more bigger gradually. According to this mechanism, data objects in space will affect the distribution of clustering results.The basic idea as follows: the assumption is that all the data objects are randomly distributed in the two-dimensional network Grid space which pro rata flex with data set[5]. The distance or similitude of the two objects o_i ,o_j in the same grid is d (o_i, o_j) which is their Euclidean distance. If o_i and o_j is the same type, d (o_i, o_j) is equal to 0, on the contrary d (o_i, o_j) is equal to 1, and thus binary similarity matrix is gained.

Supposed a number of ants in the two-dimensional grid continuously move and the implementation of picked up or down repeatedly targeted operation, at a moment ants find o_i in the location of r, the local density can be calculated by the following formula :

Among them:

$s = \{X_i \mid d_{sj} <= r, s = 1,2...j+1, N\}$ .If

$$f(o_i) = \max\left\{0, \frac{1}{s^2}\sum_{o_j \in Neigh_{s\times s}(r)}\left[1 - \frac{d(o_i,o_j)}{\alpha(1+\frac{(v-1)}{v_{\max}})}\right]\right\} \quad (8)$$

$f(o_i)$ is the density of similarity, $s \times s$ is the neighborhood area of unit r, $o_j \in Neigh_{(s\times s)}(r)$, α is the different degrees factor, $s \times s$ is the neighborhood area of field r. Ant speed v evenly distributed in [1， v_{max}]. During each cycle, ants picking up or dropting an object are guided by the following principles: If an ant does not move any object, it will randomly pick up an object from its neighboring cell.

Picked up probability is:

$$p_p(o_i) = \left(\frac{k_1}{k_1 + f(o_i)}\right)^2 \quad (9)$$

If an ant is moving object, it will randomly choose a adjacent empty unit and put it down .Or if the moving object is similar with the neighboring object, it will dropt it.

Dropting probability:

$$p_d(o_i) = \begin{cases} 2f(o_i) & f(o_i) < k_2 \\ 1 & f(o_i) \geq k_2 \end{cases} \quad (10)$$

k_1 and k_2 are threshold constants.

The algorithm is based on a grid and density clustering[5]. Firstly, High-dimensional data space must mapped to a low-dimensional grid space in order to dispose it easily, which ensure that the mapping distance within cluster is less than the distance between clusters, and the fine grid of clustering will affect the outcome of the fineness. Secondly, the action that ant pick up or down the object is affected by its similar local density f (o_i). The similar local density is more, the picked up probability pp (o_i) is smaller. The dropting probability is more, data objects are inclined to remain in the cluster and vic versa. At the same time: which record previous operating target, and put the current operating object at the one of the most similar to it ,then which can avoid repeatly and randomly searching.

C. Description Algorithm

Step1initialization parameters: CycleNumber, AntNumber, r, k1, k2, MaxCounterNumber;

Step2 randomly cast data object to a plane, and give a timely coordinate (x, y) to the data objects;

Step 3 initial target ants, ants' initialization state;

Step 4 CycleNumber + +;

Step 5 a group of ants began clustering;

Step 6 calculate f (oi) and radius is r;

Step 7 if the ants unloading, calculate Pp;

Step8 comparing Pp to a random probability Pr, if Pp> Pr, ants pick up objects;

Step 9 if ants load, calculate Pd;

Step10 using Px to calculate Pd, if Pd> Pr, ants put down their target , assign coordinates of this ant to target object, the status of ants is unloading, other data is randomly assign ants, ants start moving target and a new random coordinates assign to the ant;

Step11 if a group of ants completed, jump the next step, otherwise, select an ant without loading then go to (5) step;

Step12 if CycleNumber <MaxCycleNumber, goto step(4).Otherwise output clustering results and algorithm end.

IV. INFORMATION RETRIEVAL

The input require is expressed a t-dimensional vector that compare with each class,then search the most similar categories. For carrying out clustering analysis on the set of document, firstly be going to carry out the characteristic analysis on all documents , to gain the set of characteristic , building all documents under system in a unified coordinate and computing each other the degree

of similarity. In principles of pattern recognition, In general, similarity has method such as Euclidean distance and angle cosine similarity degree etc, degree of similarity relatively poor because using Euro-style distance to be document similarity measurement dimension is able to lead too poor for document clustering. Mainly adopt cosine similarity degree to accomplish the information retrieval job in paper. In this searching algorithm, the angle between the vectors used cosine function: rapid ants can form a large rough cluster and slow ants can accurately congregate data object in a small scope. Lastly, the ants can also keep the "short-term memory" in the process of clustering,

$$sim(C,D)=\cos\theta=\frac{C.D}{\|C\|.\|D\|}=\frac{\sum_{i=1}^{n}U_iW_i}{\sqrt{\sum_{i=1}^{n}U_i^2\sum_{i=1}^{n}W_i^2}} \qquad (11)$$

Setting similarity threshold R, if $\mathrm{sim}(C, D) \geq R$, D corresponding to the content meets the requirements, the results should be submitted to the document database, otherwise, and are not submitted.

V. EXPERIMENTAL

Experimental data are downloaded from the Internet ,which are1565 items, including political (267), economy (200), military(581),Sports(248) and computer (269), a total of five categories.

During the experiment, the information retrieval algorithm based on ant colony clustering (ACA) raised in this paper is compared with frequently used the information retrieval based on the incremental learning information (Incremental Learning, IL). The test results are as follows:

TABLEI.
ANALYSIS OF THE RESULTS

Categories Technology	IL	ACA
political	0.692	0.829
economy	0.701	0.795
military	0.636	0.773
Sports	0.812	0.887
computer	0.742	0.744

From the experiment result, we can find that ACA information retrieval technology is more practical and gain good results.

VI. CONCLUSION

The clustering algorithm based on the ant heap is not pre-specified the number of clusters, the arbitrary shape of cluster can be constructed. This algorithm is the same with pile, which can be picked up or down again like a single object and form a new cluster again. Therefore, the results of clustering algorithms have more accuracy and accelerate the cluster speed, thereby increase the speed and efficiency of information retrieval.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support from National Nature Science Foundation of China (60873247), Nature Science Foundation of Shandong Province (Y2006G20). The authors also would like to express appreciation to the anonymous reviewers for their very helpful comments on improving the paper.

REFERENCES

[1] Dorigo M,Maniezzo V,Colorni A.Ant System: Optimization by a Colony of Cooperation gents. *IEEE Trans On System.Man,and Cybernetics*, 1996; 26(1):29-41.

[2] E Bonabeau, M Dorigo,G Theraulaz. Swarm Intelligence-From Natural to Artificial System. *New York, NY: Oxford University Press,*1999.

[3]]Handl,B Meyer.Improved ant-based clustering and sorting in a document retrieval interface. *In: Proceeding of the Seventh International Conference on Parallel Problem Solving from Natuer, Springer-Verlag,* Berlin, Germany, 2002; 2439: 913-923.

[4] Yang Bin, Sun Jing Hao, Huang Road. A new method for evolutionary studying clustering . *Computer Engineering and Application*, 2003; 39 (15): 60-62.

[5] M Parag Kanade, O Lawrence Hall. Fuzzy Ants as a Clustering Concept.Dept of Computer Science Engineering . *In: 22nd international conference of the North American fuzzy information processing society*, 227-232.

[6] However,Zhang Jiao, Chun-Huang Liu, Yinxiao feng. The research of Ant Algorithm on data mining .*Computer Engineering and Application*, 2004; 40 (28):197-193.

[7] Zhang Jianhua,Jiang He,Zhang ,Xianchao. Survey of Ant Colony Clustering Algorithms. *Computer Engineering and Application*, 2006;16;171-174

Prof. Peiyu Liu: (Male, born in 1960,China)graduated from East China Normal University and obtained his master degree in computer applications in the East China Normal University in 1986. He has worked in Shandong Normal University since he was graduated from school and now is a full-time professor in Shandong Normal University . He was appointed as Doctoral tutor in 2005. Now, he is the president of department of communication engineering and the vice-president of Network College.

At present, his Main research directions are computer network and information security, network system planning, network information resources development and software development technology.

Zhenfang Zhu: (Male, born in 1981) Postgraduate. Main research directions are information security and genetic algorithm.

Lina Zhao: (Female, born in 1982) Postgraduate. Main research directions are information filtering and genetic algorithm.