December 2004

An Ant Colony Approach for Discovery of Users Preferred Navigation Paths

Haifeng Ling
Hefei University of Technology

Yezheng Liu
Hefei University of Technology

Shanlin Yang
Hefei University of Technology

An Ant Colony Approach for Discovery of Users Preferred Navigation Paths

Haifeng Ling
School of Management,
Hefei University of Technology, Hefei 230009, China
hfling@263.sina.com

Yezheng Liu
School of Management,
Hefei University of Technology, Hefei 230009, China
yezhengliu@sina.com

Shanlin Yang
School of Management,
Hefei University of Technology, Hefei 230009, China
slyang@mail.hf.ah.cn

Abstract

Web usage mining is the process of applying data mining technique to the discovery of user behavior patterns based on Web log data, for various applications. In the advance of e-commerce, the importance of Web usage mining grows larger than before. As an important research field of Web usage mining, mining users navigation patterns is the fundamental approach for optimizing the frame design of Web site. In this paper, we propose an ant colony approach for users navigation patterns. Firstly, a Web site model is built. Secondly, a navigation model is established, based on ant colony approach and Web logs. Finally, we design an algorithm regarding all users as entity for mining their preferred navigation paths.

Keywords: Web usage mining, Users navigation patterns, Ant colony optimization, Users navigation model

1. Introduction

With the rapid growth in Internet and WWW, the users browsing information is becoming enormous and pervasive. Mining the user access information, we can obtain the knowledge about user access manners, which can be used for the service providers and users. Therefore, how to automatically and effectively get the knowledge from the vast user access information, i.e. Web usage mining, became a new and important research field in the world. In the advance of e-commerce, the importance of Web usage mining grows larger than before. The development of E-commerce has made Web sites become an important environment for information service and the most popular medium between consumers and enterprises. At present, it is an urgent problem for E-commerce enterprises and researchers how to understand the users access behaviors accurately, which fertilizes to optimize Web site and improve commercial strategy. Web usage mining is a new and important research approach for this problem.

The overall process of Web usage mining is generally divided into two main tasks: data preprocessing and pattern discovery. Mining behavior patterns from Web log data needs the data preprocessing tasks that include data cleansing, user identification, session identification, and path completion. Cooley et al. presented a detailed description of data preprocessing methods for mining Web browsing patterns (Cooley et al. 1999). The pattern discovery tasks involve the discovery of association rules, sequential patterns, usage clusters, page clusters, user classifications or any other pattern discovery method. The usage patterns extracted from Web data can be applied to a wide range of applications such as Web personalization, site modification, business intelligence discovery, etc.

Interest navigation patterns mined from Web logs are useful knowledge in practice. Examples of applications of such knowledge include improving designs of Web sites, analyzing system performance as well as network communications, understanding user reaction and motivation, and building adaptive Web sites (Perkowitz et al. 2000; Sullivan 1997).

Essentially, an interest navigation pattern is a sequential pattern in a large set of pieces of Web logs, which is pursued intensively by users. Some research efforts try to employ techniques of sequential pattern mining (Agrawal et al. 1994; Agrawal et al. 1995), which is mostly based on association rule mining, for discovering Web navigation patterns from Web logs. Chen et al. introduced the concept of using the maximal forward references to break down user sessions into transactions for mining access patterns (Chen et al. 1998; Nanopoulos et al. 2000). Spiliopoulou proposed an algorithm for building and aggregating tree from web logs, then mining the Web access patterns by MINT mining language (Spiliopoulou 1999). Borges proposed a heuristic algorithm based on hypertext probabilistic grammar to mine the user navigation patterns (Borges 2000; Borges et al. 2000).

All above have the same character: determine the frequent navigation patterns based on association rule. Actually as an improved association rule, the above algorithms simply utilize the access frequency as support to mine regardless of confidence that is another factor of association rule revealing user interest. But the conventional concept of confidence is not suitable for the special frame of Web site. Hyperlink of current page determines the next page could be visited unless type the URL directly. In addition, many of the existing sites use some specific ways, such as color and graphical representations, to absorb users to access. Under these conditions, confidence cannot measure user interest precisely.

Therefore, it is necessary to propose a new concept that can replace the conventional concept of confidence to reflect user interest accurately under Web environment. In this paper, we recast the recently proposed ant colony optimization algorithm to suit the need for mining interest navigation patterns efficiently and accurately from Web logs. The main contributions are as follows. First, the new concept of pheromone reflecting user interest is proposed. Second, considering the structure of Web site, an ant colony navigation algorithm is developed for mining user preferred navigation patterns. It can get user access preferred path by the page-page transition probability statistics of users behaviors.

2. Ant colony optimization

Colonies of social insects can exhibit an amazing variety of complex behaviors. The study of ant colonies behavior gives rise to a completely novel field of research, now known as ant colony optimization (ACO).

The ACO technique has emerged recently as a novel heuristic in the class of naturally derived problem-solving strategies. Other categories include neural networks, simulated annealing, and evolutionary algorithms. Marco Dorigo first introduced the Ant System (AS), the earliest version of the ACO methods, in his dissertation in 1992 (Dorigo 1992). The algorithm, called Ant System, was applied to the traveling salesman problem (TSP). The AS optimization algorithm is basically a multi-agent system in which low level interactions between single agents (i.e., artificial ants) result in the complex behavior of the colony as a whole.

The ACO algorithms mimic the techniques employed by real ants to rapidly establish the shortest route between food source and their nest. It is well known that the medium used to communicate information among individual ants regarding paths is pheromone trails. Ants start searching the area surrounding their nest in a random manner. When an isolated ant comes across some food source in its random sojourn, it deposits a quantity of pheromone on the ground. Other randomly moving ants in the neighborhood can detect this marked pheromone trail. Further, they follow this trail with a very high degree of probability and simultaneously enhance the trail by depositing their own pheromone. More and more ants follow the pheromone rich trail and the probability of the trail being followed by other ants is further enhanced by increased trail deposition. It is this auto catalytic process characterized by positive feedback mechanism that helps the ants to establish the shortest route.

Given a set of *n* cities, set various distances apart, the TSP involves finding the shortest closed path that visits every city once. D_{ij} denotes the length of the path between cities i and j. Moreover, an instance of the TSP is given by a graph (N, E), where N denotes the set of cities and E represents the set of edges between cities. The ACO algorithm for this problem can be simply represented as follows (Dorigo et al. 1997).

```
Step1: Set parameters, initialize pheromone trails
Step2: While (termination condition not met) do
                Construct Solutions
                Apply Local Search
                Local_Pheromone_Update
         Endwhile
         Global_ Pheromone_Update
```

In fact, currently there are many successful applications of ant algorithms to adaptive routing in communications networks and combinatorial optimization problems.

3. The proposed approach

In this paper, we propose an ACO algorithm for users navigation patterns. This algorithm has the following three important strategies: (a) establishing a simple Web site model, (b) utilizing Web logs to build users navigation model, (c) adopting ant colony algorithm to discover users preferred navigation paths. If user visit a Web site, the viewing frequency and viewing time of a page reveals users interest. The more frequent visit and the longer viewing time, the more interested and preferred visiting, vice versa. For the simplification of this problem, we don’t consider the factor of viewing time in the algorithm proposed.

3.1 A Web site model

Definition 3.1 (web site model). A web site model is given by a graph (N,E), where N denotes the set of Web pages and E represents the set of hyperlinks between Web pages.

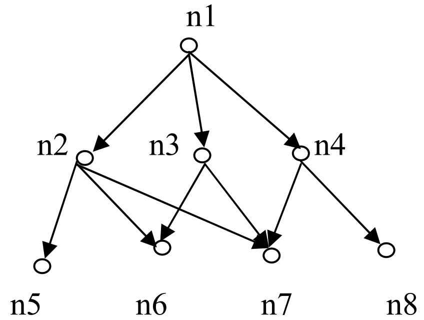

Fig. 3.1 A simple Web site model

3.2 Users navigation model

Given a set of users access as follows.
S={(n1,n4,n8), (n1,n4), (n1,n4,n7), (n1,n3,n7), (n1,n3,n7), (n1,n3,n6), (n1,n2,n6), (n1,n2,n5,n3,n7), (n1,n2,n5), (n4,n7)}

According to Fig. 3.1 and users navigation behaviors, we propose the users navigation model as shown in fig. 3.2

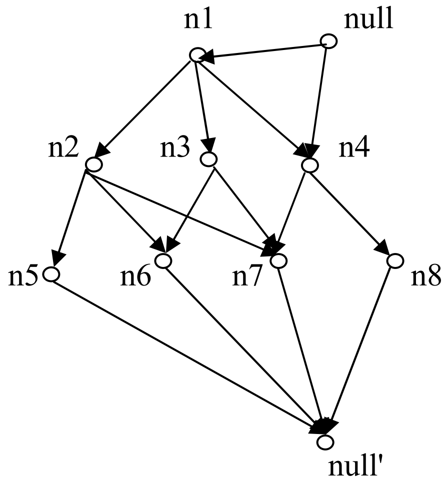

Fig. 3.2 Users navigation model

Definition 3.2 (Users Navigation Path). If user visits a Web site, the sequence of his related Web requests is called user navigation path. N is the URL set of a Web site. Null represents that users visit page by inputting URL directly, using bookmark or hyperlink on other Web sites. Null' represents the end of user access or browse other Web sites.

Definition 3.3 (Support). Support η_{ij} indicates the user access frequency for the path from node i to j. Support η_{ij} is defined as follows.

$$\eta_{ij} = C_{ij} / ((\sum_{k=1}^{n} C_{ik}) / n) \qquad (1)$$

where C_{ij} is the access times for the path from node i to j, and there are n different selections from node i to all the next nodes.

Usually, the paths browsed more frequently are regarded as the preferred navigation patterns. It is inaccurate. For example, the navigation frequency of a path relates to its position. In general, the nearer to home page the paths, the more frequently it is browsed, vice versa. The page having lots of links may be visited more frequently, but it does not show that users have more interest in it. Because maybe users have to visit this page before access a target page. In addition, many of the existing sites use some specific ways, such as color and graphical representations, to absorb users to access. The navigation frequency of a path cannot measure users true preference precisely. Therefore, we present a new definition of user access frequency and utilize it as support to mine.

Definition 3.4 (Pheromone function). The pheromone function $\tau_{ij}(t)$ indicates the user access interest for the path from node i to j. The pheromone function $\tau_{ij}(t)$ is defined as follows.

$$\tau_{ij}(t+1) = (1-\rho) \cdot \tau_{ij}(t) + \Delta\tau_{ij} \qquad (2)$$

where $0<\rho<1$, ρ represents the pheromone attenuation ratio

$$\Delta\tau_{ij} = Q \cdot \eta_{ij} \qquad (3)$$

where Q denotes a constant relating to the quantity of pheromone

According to ant algorithm for TSP, the shorter the trail, the richer the pheromone deposited on the trail. The algorithm for TSP utilizes reciprocal of the length of the trail as support. Therefore, we present the definition of the increment of pheromone $\Delta\tau_{ij}$, in which the user access frequency η_{ij} determines the increment of pheromone $\Delta\tau_{ij}$.

Definition 3.5 (Preference function). Preference function $p_{ij}(t)$ indicates the users access preference for the path from node i to j. Preference function $p_{ij}(t)$ is defined as follows.

$$p_{ij}(t) = \frac{[\tau_{ij}(t)]^{\alpha} \cdot [\eta_{ij}]^{\beta}}{\sum_{k=1}^{n} [\tau_{ik}(t)]^{\alpha} \cdot [\eta_{ik}]^{\beta}} \qquad (4)$$

Where α, β indicate the relative importance of pheromone and support respectively.

3.3 Algorithm for mining preferred navigation paths

Provided that the number of the elements in the user access set is m, then the concrete algorithm is shown as follows.

Step1: Set parameters, and initialize pheromone trails
    Set t=0, $\tau_{ij}(0)=0$
Step2: For (k=1 to m)
        Update η_{ij}
        Update $\tau_{ij}(t+1)$ on every path
        Set t=t+1
    Endfor
Step3: Initialize α, β
    Calculate $p_{ij}(t) = [\tau_{ij}(t)]^{\alpha} \cdot [\eta_{ij}]^{\beta} / \sum_{k=1}^{n} [\tau_{ik}(t)]^{\alpha} \cdot [\eta_{ik}]^{\beta}$

According to the page-page transition probabilities, we can get users access preferred paths.

4. Conclusion

At present, it is an urgent problem for the applications of Internet E-business how to develop users preferred patterns accurately, which fertilizes to optimize Web site and improve commercial strategy. In this paper, we propose the new concept of pheromone, which can reflects user navigation interest. Considering the structure of Web site, an ant colony navigation algorithm is developed for mining user preferred navigation patterns. It can get user access preferred paths by the page-page transition probability statistics of users behaviors. This algorithm is suitable for applications in E-business, such as to optimize Web site or to design personalized service. We plan to perform extensible testing and investigate possible improvements.

Ackowledgements

This paper is supported by the Key Research Project of Anhui Province (No.03021057) and the Key Science and Technique Research Project of Education Ministry (No.02127).

References

Agrawal, R., and Srikant, R. "Fast algorithms for mining association rules," *Proceedings of the 20th International Conference on VLDB*, 1994, pp. 487-499.

Agrawal, R., and Srikant, R. "Mining sequential patterns," *Proceedings of the 11th International Conference on Data Engineering*, 1995.

Borges, J. "A Data Mining Model to Capture Visitor Web Navigation Patterns," *Ph.D. thesis*, Department of Computer Science, University College London, July 2000.

Borges, J., and Levene, M. "A heuristic to capture longer visitor web navigation patterns," *Proceedings of the first International Conference on Electronics Commerce and Wen Technologies*, (EC-Web 2000), London-Greenwich, UK, September 2000, pp. 155-164.

Chen, M.S., Park, J.S., and Yu, P.S. "Efficient data mining for path traversal patterns," *IEEE Trans. Knowledge Data Engn* (10:2), 1998, pp. 209-221.

Cooley, R., Mobasher, B., and Srivastava, J. "Data preparation for mining world wide web browsing patterns," *J. Knowledge Information Syst* (1:1), 1999.

Dorigo, M. "Optimization, learning and natural algorithms," *ph.D. Thesis* , Dipartimento di Elettronica, Politecnico di Milano, Italy, 1992.

Dorigo, M., and Gambardella, L. M. "Ant colonies for the travelling salesman problem," *BioSystems* (43), 1997, pp73-81.

Nanopoulos, A., and Manolopoulos, Y. "Finding generalized path patterns for web log data mining," *"Proceedings of the East-European Conference on Advances in Databases and Information Systems (ADBIS'00)"*, 2000, pp. 215-228.

Perkowitz, M., and Etzioni, O. "Towards adaptive web sites: conceptual framework and case study," *Artificial Intell* (118:2), 2000.

Spiliopoulou, M. "The laborious way from data mining to web mining," *Int. J. Comput. Syst., Sci. Engng,* 1999, pp. 113-126.

Sullivan, T. "Reading reader reaction: a proposal for inferential analysis of web server log files," *Proceedings of the third Conference on Human Factors and The Web*, Denver, Colorado, 1997.