Engineering Applications of Artificial Intelligence 25 (2012) 889–897

Contents lists available at SciVerse ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Predicting web user behavior using learning-based ant colony optimization

Pablo Loyola *, Pablo E. Román, Juan D. Velásquez

Department of Industrial Engineering, Universidad de Chile, República 701, P.O. Box 8370439, Santiago, Chile

ARTICLE INFO

Article history:
Received 15 July 2011
Received in revised form
29 September 2011
Accepted 19 October 2011
Available online 6 December 2011

Keywords:
Ant colony optimization
Web usage mining
Multi-agent simulation
Text preferences

ABSTRACT

An ant colony optimization-based algorithm to predict web usage patterns is presented. Our methodology incorporates multiple data sources, such as web content and structure, as well as web usage. The model is based on a continuous learning strategy based on previous usage in which artificial ants try to fit their sessions with real usage through the modification of a text preference vector. Subsequently, trained ants are released onto a new web graph and the new artificial sessions are compared with real sessions, previously captured via web log processing. The main results of this work are related to an effective prediction of the aggregated patterns of real usage, reaching approximately 80%. In the second place, this approach allows the obtaining of a quantitative representation of the keywords that influence the navigational sessions.

1. Introduction

Since the beginning of the World Wide Web, one of the aspects that has caught the attention of the emerging researchers was the way in which the users interact with the structure and content of web sites. For that purpose, multiple analyses and models were generated to understand web user behavior in order to display relevant content and maximize traffic. The emergence of e-commerce represented a major change in terms of the valuation of the web user, not only as a consumer of content, but also as a client that has to be seduced into making a purchase. This, together with the rise of the Web 2.0 paradigm, promoted the unification of knowledge and techniques in what is now commonly known as Web usage mining (WUM), a field specializing in the study of web user behavior. One of its main objectives is to achieve web user *personalization*, which means the capability of generating adaptability within the web site, both through link structure transformation and real time suggestions, in relation to the user's preferences and generated paths.

In general, WUM uses traditional behavioral models, operations research and data mining methods to deal with web usage data. However, some modifications are necessary according to their respective application domain. Two families of techniques have been used to analyze the sequential patterns: deterministic and stochastic. Each one has been used depending on the approaches that have been adopted.

Soft computing methodologies have gained a considerable amount of relevance in relation with WUM, due to their flexible implementation and results in the field of recommendation-based systems and adaptive web sites (Lin and Tseng, 2010). Within these fields, special attention has been concentrated on bio-inspired metaheuristics, which are commonly ruled by the concept of *swarm intelligence*, the ability of a group of agents to perform complex tasks through a collaborative process. Instead of trying to mimic human intelligence, the inspiration is taken from the observation of social insects such as ants or bees (Christensen et al., 2007).

Ant colony optimization (ACO) is one of the tools used for these purposes. This metaheuristic is inspired by the way ants optimize their trails for food foraging based on releasing chemical substances into the environment called *pheromones*. This simple idea is applied to the web user trails of visited web pages, also called sessions (Liu, 2007). Artificial ants are trained through a web session clustering method modifying an intrinsic text preference vector which represents the importance given by the users to the set of most important keywords. Furthermore, trained ants are used to predict future browsing behavior.

This paper is organized as follows. Section 2 provides an overview of related work. In Section 3 the proposed model is presented. Then, in Section 4 an application of our work on a real web site is described. Finally, conclusions and future work are presented in Section 5.

2. Related work

ACO is inspired by the behavior of some insect species which is related to the presence of a social component in their

* Corresponding author. Tel./fax: +56 2 978 4834.
E-mail addresses: ployolah@dii.uchile.cl (P. Loyola), proman@ing.uchile.cl (P.E. Román), jvelasqu@dii.uchile.cl (J.D. Velásquez).

0952-1976/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2011.10.008

organizational structure, allowing them to perform complex tasks through a cooperative and coordinated process. The key component behind this is their capability of generating an indirect communication by modifying the environment. This mechanism is called *stigmergy*. Specifically in ants, it is based on the use of chemical *pheromones* which can be deposited on the ground and detected by the colony in order to execute labors such as food foraging, cooperative transport and corpse grouping. The methodology is based on the progressive construction of pheromone trails from the nest to the food source, searching for the minimum path. This is achieved through an auto-catalytic process in which pheromone levels are reinforced in inverse proportion to the time needed to walk through its respective path. As ants choose paths with higher pheromone levels, suboptimal solutions evaporate and the algorithm returns the shortest path. Biological studies have shown that colony-level behavior can be explained using models based on stigmergic communication, in terms of simulating the levels of self-organization (Dorigo and Stutzle, 2004).

Marco Dorigo proposed the first ACO model to solve the TSP[1] in 1992 (Dorigo and Gambardella, 1997). This work stated the principles of the methodology in terms of the use of four components: graph representation, problem restrictions, pheromone trails and solution construction. Most of the calculation is produced by using two variables that manage trails generation. The first is a support measure η_{ij}, also called heuristic information, which contains problem information that is not directly accessible to the ants, but has been introduced externally. This measure can store either costs or utility associated with a decision. The second, a pheromone function, $\tau_{ij(t)}$ is defined as

$$\tau_{ij}(t+1) = (1-\rho)\tau_{ij}(t) + \Delta\tau_{ij}, \tag{1}$$

where $\Delta\tau_{ij} = Q\eta_{ij}$, with Q a constant and ρ an attenuation factor, which is associated with the pheromone evaporation. Finally, the trail preference $p_{ij}(t)$

$$p_{ij}(t) = \frac{[\tau_{ij}(t)]^{\alpha} * [\eta_{ij}]^{\beta}}{\sum_{k=1}^{n} [\tau_{ik}(t)]^{\alpha} * [\eta_{ik}]^{\beta}}, \tag{2}$$

where α and β are the weights for pheromones and support, respectively.

This allows the development of the metaheuristic and its further applications in fields from vehicle routing problems to image processing and machine learning (Dorigo and Stutzle, 2004).

The main advantages of ACO are related with its inherent parallelism (which can be useful when working together with an effective multi-threading computing strategy), its flexible implementation (which allows its use in dynamic problems), and the positive feedback (which accounts for rapid discovery of good solutions). As drawbacks of these techniques, it can be mentioned that although convergence was mathematically demonstrated (Dorigo and Stutzle, 2004), convergence time is difficult to predict and a deficient parameter setting can lead to a local optimal solution instead of a global one (Umarani and Selvi, 2010).

The implementation of ACO in web intelligence tasks, together with other bio-inspired heuristics, has gained a considerable amount of attention due to the necessity of finding new ways to explain and simulate web user behavior.

Ling (Lin and Tseng, 2010) proposes an ACO-based method for discovering navigational patterns, by using a support measure based on real transitions between pages. The resulting user navigational preferences, represented by pheromone levels, are directly proportional to the real frequency of usage. Clustering is another area in which ACO has been used widely (Bharne et al., 2011). Abraham and Ramos (2004) propose an approach based on the clustering of web user sessions. The model takes inspiration from the ability of certain kinds of ants to group corpses and generate *ant cemeteries*.

White et al. (2010) use an ACO implementation to progressively find the best correlation between web content and a set of online advertisements in order to maximize the probability of profit. To achieve this, a vector space model is used together with the notion of web user preference vector.

3. A methodology for simulating web user behavior using ACO

The key aspect of the proposed model is to enable artificial ants to learn from real user behavior in terms of the settings of the intrinsic parameters of an ACO implementation, which have to be adapted to a web environment.

3.1. Web user sessions clustering

The learning algorithms to be used in this paper are based on a continuous comparison between the artificial sessions generated by the ants and the web user behavior represented through real web user sessions. A first approach could be to compare each artificial session with the complete set of real sessions available. This method could take a considerable amount of time due to the large number of sessions generated in the study time intervals. Thus, it is proposed to perform a clustering process of the web user sessions in order to reduce the number of subsequent comparisons.

3.1.1. Similarity measure

To generate a valid similarity measure two main subjects must be considered. In the first place there are the sets of web pages belonging to each session, and in the second place there is the order in which those web pages were visited. Most approaches to this problem have used a similarity measure based on the number of operations needed to transform one session into another, for instance, using the Levenshtein distance (Velásquez and Palade, 2008) or the sequence alignment method (Hay et al., 2004). But those methods do not incorporate the notion of sequential path between the elements, which is fundamental in terms of the existence of a link element which makes the user trail possible. We propose to use a degree of similarity (Spiliopoulou et al., 2003) between sessions based on the calculation of longest common subsequence (LCS) (Gusfield, 1999).

Given r a real session and c an artificial session, the similarity measure is defined by

$$sim(r,c) = \frac{LCS(r,c)}{\max\{\|r\|,\|c\|\}}, \tag{3}$$

where $\|r\|$ represents the length of a given session r and $LCS(r,s)$ is the longest common subsequence between sessions. This method allows to compare two sessions regardless if they have different lengths.

3.1.2. Clustering method

Web user sessions are basically a categorical representation of the paths followed by users and do not have an intrinsic value. Rather, it is their components, web pages, which give them an inner characteristic. Thus, the notion of a mean value for representing them does not exist (Liu, 2007; Murtagh, 1983; Zhao and Karypis, 2002). We propose to use a hierarchical method for clustering consisting of an agglomerative process that at every step group sessions according to the similarity measure. This method uses only the similarity (or distance) between elements which fit into the categorical nature of web user sessions (Liu, 2007). It also allows visualizing at every step the partial groups being generated. This is

[1] Traveling Salesman Problem.