Expert Systems with Applications 41 (2014) 1593–1600

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Identifying web sessions with simulated annealing

Tomás Arce [a,1], Pablo E. Román [b], Juan Velásquez [c], Víctor Parada [d,*]

[a] Departamento de Ingeniería Informática, Universidad de Santiago de Chile, Av. Ecuador 3659, Estación Central, Santiago, Chile
[b] Center of Mathematical Modelling (CMM) UMI CNRS 2807, Universidad de Chile, Av. Blanco Encalada 2120, Piso 7, Santiago, Chile
[c] Departamento de Ingeniería Industrial, Universidad de Chile, República 701, Santiago, Chile
[d] Departamento de Ingeniería Informática, Universidad de Santiago de Chile, Av. Ecuador 3659, Estación Central, Santiago, Chile

ARTICLE INFO

Keywords:
Web usage mining
Web session
Simulated annealing

ABSTRACT

Delivery of efficient service through a web site makes it compulsory in the redesigning stage to take into account the behavior of the users, which can be studied by means of a web log file that partially records information about user visits. The reconstruction of all of the sequences of pages that are visited by users who browse a web site is known as the web sessionization problem, and it has been formulated by means of an integer programming model; however, because a web log can accumulate a large amount of information, it is necessary to reconstruct the sessions over a period of weeks or months, thus the solution to this problem requires a long computational processing time. This paper presents a heuristic approach based on simulated annealing for the sessionization problem. Using this approach, it has been possible to reduce the processing time up to 166 times compared to the time that is required for the integer programming model. Furthermore, the metaheuristic solution finds new optimum values, which achieve increases on the order of 17% in the best cases.

1. Introduction

The Internet has become a flourishing source of data for different research fields related to social networks, such as sociology (Lin, Jheng, & Yu, 2012; Nohuddin et al., 2012), marketing (Fong, Zhou, Hui, Tang, & Hong, 2012; Wang, Ting, & Wu, 2013) and computer science (Devi, Devi, Rani, & Rao, 2012; Yin & Guo, 2013). Also, the massive usage of the Internet drives many lucrative businesses such as e-commerce. Thus, it is increasingly necessary for web sites to be designed with a structure that makes it easy for the user to obtain the service (Wang & Ren, 2009). To that end, it is necessary to study the behavior of the users which is partially kept in a privacy-compliant data file saved on the web server known as a web log. Legislation in several countries forbids the storage of personal information in order to safeguard personal privacy (Mayer & Mitchell, 2012). In order to follow national laws, internet companies must rely on the anonymity of web logs for extracting information about their user preferences and browsing behavior (Velásquez, 2013; Velásquez & Palade, 2008). The generated browsing sequences represent an input source for discovering the behavior patterns of users who visit a web site (Cooley, Mobasher, & Srivastava, 1999; Kosala & Blockeel, 2000; Tao, Hong, Lin, & Chiu, 2009).

Every time a user requests a web page or some resource contained on it such as images, videos, and sounds, a new record is created in the web log. The generated information allows the detection of the most visited pages, the common page access sequences, the users' preferred contents and even the generation of user profiles (Choi & Lee, 2009; Nasraoui, Soliman, Saka, Badia, & Germain, 2008). Such detection is called web usage mining oriented toward extracting knowledge from web logs (Román, L'Huillier, & Velásquez, 2010), which requires the extraction of individual sequences of a user's web interaction while maintaining anonymity (Mayer & Mitchell, 2012). The following data are typically recorded in the file: the IP address from which the inquiry is made, the time and date of the inquiry, the requested resource, a code indicating the result of the operation, the number of bytes transferred in the request, a string that contains the address of the site from which the present request was originated and the browser and operating system that was used. The generated browsing sequences represent an input source for discovering the behavior patterns of users who visit a web site (Cooley et al., 1999; Kosala & Blockeel, 2000; Tao et al., 2009).

With expert assistance the information stored in the web log can be used to support decision-making that can help in restructuring a web site to improve access to the preferred contents; the information can also be used to detect market segments based on the buying behavior of the users and to implement resource suggestion systems for the users. The sequence of individual interactions is called a session and the reconstruction of all of the page

* Corresponding author. Tel.: +56 2 7180900.
E-mail addresses: tomas.arcec@usach.cl (T. Arce), proman@dii.uchile.cl (P.E. Román), jvelasqu@dii.uchile.cl (J. Velásquez), victor.parada@usach.cl (V. Parada).
[1] Tel.: +56 2 7180900.

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.08.056

sequences visited by the users during their browsing through a web site is known as the web sessionization problem (WSP) (Berendt, Mobasher, Spiliopoulou, & Wiltshire, 2001; Chitraa & Selvdoos, 2010; Huynh & Miller, 2009; Poženel, Mahnic, & Kukar, 2010; Velásquez & Palade, 2008). A correct reconstruction of the sessions must take into account that the validity of the generated patterns depends largely on the credibility of the sessions obtained (Berendt et al., 2001). In spite of that, the reconstruction of sessions is not a trivial process because of factors that hinder reconstruction, such as proxy servers or the use of cache memory in the client's web browser.

The techniques that solve the WSP from a web log must deal mainly with the nonexistence of a clear identification of the users. Matching the IP address as a single criterion is not sufficient, because when multiple users have access to a site through a proxy server, they are registered in the web log with the proxy server's address. To partially remove the identification error, the techniques use the IP address and the web browser that was used by the web site visitor as identification criteria (Pirolli, Pitkow, & Rao, 1996). Other techniques directly track user operations by using cookies with client-side scripts obtaining accurate sessions, but they are not recommended since they violate user privacy laws (Mayer & Mitchell, 2012). Therefore, the web usage mining on web logs is a safe way to analyze user preferences, as long as we take into account that in general a high quality sessionization is not possible by means of machine learning techniques that are known to be subject to data error (Román et al., 2010).

Current improvement of the web usage mining processing relies on accurately solving the WSP, for which several heuristics have been proposed. The most widely used technique to tackle the WSP is the time heuristic, which considers that the sessions have a time limit (Catledge & Pitkow, 1995; Cooley, Mobasher, & Srivastava, 1997; Huynh & Miller, 2009). This technique groups the records according to the IP address and the web browser that was employed by the user; it arranges the records of each resultant group in a temporal order and then obtains the sessions, ensuring that the first and last records of a given session do not exceed a time limit. The major drawback of this algorithm consists of the null consideration of the hyperlink topology of the web site. A revised version of this heuristic considers the site's link structure together with a temporal criterion (Pirolli et al., 1996), in addition to the criteria that the consecutive records of the same session must refer to pages between which there is a link. Nowadays, with modern web browser navigation it becomes hard to track the information due to the multiple ways that pages are cached, loaded, and navigated. Multitab navigation, back button browsing and history jumps are commonly not reflected on web logs (Román et al., 2010), so hyperlink topology restriction is relaxed from being a strong restriction on session properties.

A recent approach addresses the WSP as an optimization problem by defining an integer programming model (Román, Dell, & Velásquez, 2010). All of the possible reconstructions of sessions from a given web log constitute the feasible solution space of the problem. The choice of a specific reconstruction implies a search for the reconstruction that has a maximum value in terms of a specific objective function. Web hyperlink topology and time sequencing are incorporated as restrictions in the model. Relaxed browsing behavior could even be easily included (Dell, Román, & Velásquez, 2009). However, because a web log in a single day can accumulate an enormous number of records and the supporting decision making requires the reconstruction of sessions over a period of weeks or months, the number of variables of the integer programming models is huge for the currently available capacity for solving integer programming problems. Solving even small instances of the problem requires several hours of computing time. The performance is even worse when considering more common browsing behaviors such as parallel tabs and back buttons. The ideal situation would consider the existence of an algorithm that can process the information in real time by requiring a short computing time.

A novel algorithm for solving the WSP was presented in Bayir, Toroloslu, Demirbas, and Cosar (2012). The algorithm is based on graph modeling of the sessions that are constructed considering maximal path length, hyperlink topology and back button browsing. They found improved accuracy in recovering sessions relative to previous models. However, their major drawback comes from the theoretical justification of the model affecting its reproducibility.

Instead of generating the optimum solution by means of an algorithm that solves the integer programming problem, we obtain a good quality solution using a small amount of computing time, by means of Simulated Annealing (SA) which is a metaheuristic that emulates the physicochemical process that takes place in the cooling of pure substances, systematically generating a new solution from the current solution and allowing, at some instant, the choice of a *poor solution*, with a certain probability that decreases with time (Kirkpatrick, Gelatt, & Vecchi, 1983; Talbi, 2009). This method has been shown to be very efficient for solving problems belonging to the *NP-Hard* class, such as the design of electronic circuits, the reconstruction of images, the generation of roads, set partition problems and planning problems (Suman & Kumar, 2005). This paper presents an approach that is based on simulated annealing for the WSP for the purpose of reconstructing each session of the users that visit a web site.

The second section presents the WSP representation under an SA, identifying the elements needed for the experimental phase. The third section presents the main results, and in the last section, the main conclusions are given.

2. Methods and materials

2.1. The web sessionization problem

The model founded on integer programming (Dell, Roman, & Velasquez, 2008) is based on ensuring that a session is constituted by a set of records that share the same IP address and the same web browser. For consecutive records within a session, the constraints ensure the maintenance of the link structure of the site and that the session does not go beyond a certain time limit. Thus, if a record is found directly after another record within the same session, then the following are true:

- Both share the same IP address and web browser.
- There is a link between the referenced page in both records.
- The time difference between the recorded application for both records does not exceed a certain *mtp* value or time window.

Let

- *r* and *r*' be records of a web log;
- *o*be the order of a record in a given session, $o = 1, 2, \ldots, O$, where is the maximum size that a session can have;
- *s* be the identifier of a session;
- C_o be the value of the coefficient of the objective function when there is a record assigned to position *o*;
- x_{ros} be the binary decision variable, which has a value of 1 when record *r* is assigned to position *o* in a given session *s* and a value of 0 in any other case;
- *Bpage*(*r*) be the set of records that can be directly before record *r* in the same session, according to the criteria of having the same IP address, corresponding to the same web browser, having a link between consecutive records and maintaining the time window between consecutive records;