Finding All Maximal Paths In Web User Sessions

Murat Ali Bayir*
Microsoft Bing Ads R&D Center
Bellevue, WA, USA, 98004

Ismail Hakki Toroslu
Middle East Technical University
Ankara, 06530, Turkey

ABSTRACT

This paper introduces a new method for the session construction problem, which is the first main step of the web usage mining process. The proposed method is capable of extracting all possible maximal navigation sequences of web users. Through experiments, it is shown that when our new technique is used, it outperforms previous approaches in web usage mining applications such as next-page prediction.

Keywords

Web Mining, Linked Data, Graph Theory

1. INTRODUCTION

The purpose of Web Usage Mining (WUM) [4] is to find interesting knowledge about navigation behaviors of web users. The first step of WUM includes the session construction from user logs which directly affects the quality of patterns discovered in WUM process. Previous approaches [5] for session reconstruction have two problems. They are either using time information without link data or add artificial backward movements (noise) to complete paths in web topology. These problems can be handled by using cookies and adding client specific information in server requests. However, for various reasons, such as security and changes in the internal structure of web site, some site owners may not want to use proactive approaches at all. Instead of that, these site owners prefer to process only their raw server logs.

Our previous method Smart-SRA [2] solved most of the problems mentioned above. However, it still can not capture particular user behaviors due to its greedy nature. Consider the following example on web topology given in Figure 1. A web user follows the path $[P_1, P_{13}, P_{49}]$ and goes back to P_{13}. Then, the same user visits P_{34} from links on P_{13}. In this case, there will be two maximal sessions sequences that represents user's paths in the graph $\{[P_1, P_{13}, P_{49}], [P_1, P_{13}, P_{34}]\}$.

*Corresponding author, email: mbayir@microsoft.com

Copyright is held by the author/owner(s).
WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889351.

However, none of the previous heuristics [5] including Smart-SRA [2] are capable of extracting both.

To overcome the problems of Smart-SRA and previous approaches, we propose a new technique, called as Complete Session Reconstruction Algorithm (C-SRA). C-SRA is very powerful algorithm which produces complete set of maximal paths that can be obtained from given page request sequence and web topology. It produces sessions that better represent complex navigation behavior where user goes and backs between web pages and jumps to different nodes in the topology of given web site.

The next section describes the details of C-SRA. Next, the pattern discovery on top of C-SRA sessions is explained. Finally, we present our preliminary experimental results.

2. C-SRA ALGORITHM

C-SRA is a two phased method that produces sessions as a set of all possible maximal sequences. A Maximal sequence is a path in the graph which is not subsequence[1] of any other sequence generated from the same session.

In the first phase of C-SRA, user log sequences from server logs including (IP, URL, Time) tuples are partitioned into smaller candidate sessions by using time constraints. The second phase of C-SRA constructs all maximal navigation sequences from the candidate sessions generated at the first phase. The input and outputs of second phase of C-SRA is given below:

Input: A possibly cyclic directed graph $G = (V, E)$ such that $V = \{v_1, v_2, \ldots, v_n\}$ is vertex set and $E \subset V \times V$ is a set of edges, and an ordered sequence of vertices $S = [vs_1, vs_2, \ldots, vs_k]$ where each $vs_i \in V$ (without any repetition for our problem, since the second request of the same page is always provided by the browser cache for limited time interval).

Output: Set of maximal sequences O, where each $O_j = [vs_{j1}, vs_{j2}, \ldots, vs_{jm}] \in O$ is a maximal navigation sequence that corresponds to a path in G. That is, for every pair of consecutive vertices in a sequence O_j, such as vs_{jp} and $vs_{j(p+1)}$, there exists an edge $\langle vs_{j(p)}, vs_{j(p+1)} \rangle \in E$. In addition, in order to satisfy the maximality property, there is no sequence $O_q \in O$ such that O_j is a sub-string of O_q.

The main part of the C-SRA consumes the site graph and vertex sequence (G, S) and produces the output O mentioned above. The details of each phase are given below:

Phase 1 constructs candidate session set from user page request sequence by using time thresholds (applying both

[1]subsequence relation here is same as substring relation

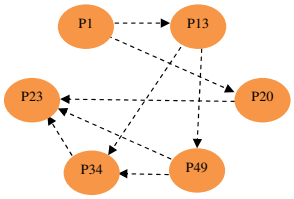

Figure 1: Topology