Chapter 14

Sequence Pattern Mining for Web Logs

Pradeep Kumar
Indian Institute of Management Lucknow, India

Raju S. Bapi
University of Hyderabad, India

P. Radha Krishna
Infosys Labs, Infosys Limited, India

ABSTRACT

Interestingness measures play an important role in finding frequently occurring patterns, regardless of the kind of patterns being mined. In this work, we propose variation to the AprioriALL Algorithm, which is commonly used for the sequence pattern mining. The proposed variation adds up the measure of interest during every step of candidate generation to reduce the number of candidates thus resulting in reduced time and space cost. The proposed algorithm derives the patterns which are qualified and more of interest to the user. The algorithm, by using the interest, measure limits the size the candidates set whenever it is produced by giving the user more importance to get the desired patterns.

INTRODUCTION

Finding frequent sequence pattern from large transactional databases is one of the successful data mining endeavors introduced by Agarwal and Srikant (1995). It obtains frequent sequential patterns of items satisfying the condition that the number of their occurrences, called support, in the item sequence, called transaction database, is greater than or equal to a given threshold, called minimum support. The obtained frequent patterns could be applied to analysis and decision making in applications like time-series stock trend, web page traversal, customer purchasing behavior, content signature of network applications, etc.

The task of sequence pattern mining is to discover the frequently occurring subsequences from the large sequence database. Regardless of how frequent these sequences occur it is also required to exploit the relationships among the sequences.

One of the challenging problems with sequence generating systems is the large number
of sequences being generated. These generated rules may be of no practical value or interest to the analyst. To overcome the problem researchers have started using to measure the usefulness or interestingness of rules. Whenever a interestingness measure is applied, there is clear tradeoff between accuracy and the coverage of knowledge.

Interestingness decreases with coverage for a fixed number of correct responses (remember accuracy equals the number of correct responses divided by the coverage).

In this chapter our focus is to mine sequential patterns from sequence database. For this work we choose web usage mining domain is used to demonstrate our approach. The current approach is highly applicable in any domain where data exhibits sequentiality in nature.

In this chapter we introduce a general framework of mining sequential patterns using interest measure. The sequential patterns obtained due to the modified algorithm are compared to the original sequence pattern mining algorithm, AprioriALL (Agarwal & Agarwal, 1995).

Our research is motivated by following two observations:

- Limited customization, the user has no option to choose the type of pattern catering to his need depending on his interest.
- The patterns derived are not interesting as Support is not a good interestingness measure for either association rules or sequential patterns.

Now we formally define our research problem addressed in this work. The problem of Sequential Pattern Mining in general to web mining can be stated as “Given a set of user sessions,, with each session consisting of a list of elements and each element consisting of a set of items and given user specified minimum interest value, min_support, the problem is to generate all candidates which satisfy the minimum interest value and to find all the sequences whose occurrence frequency in the set of sequences is no less than min_support “

Mining sequential patterns has become an important data mining task with broad applications in business analysis, career analysis, policy analysis, and security. Many papers on sequential pattern mining focus on specific algorithms and evaluating their efficiency (Ayers et al, 2002, Pei et al 2001, Srikant & Agarwal, 1996).

In this work, we focus on the problem of mining sequential patterns. Sequential pattern mining finds interesting patterns in sequence of sets. Mining sequential patterns has become an important data mining task with broad application areas. For example, supermarkets often collect customer purchase records in sequence databases in which a sequential pattern would indicate a customer’s buying habit.

Currently after many years of research in the Market basket analysis through Sequence Pattern Mining problem (Agarwal & Agarwal 1995, Pei et al 2001, Srikant & Agarwal, 1996) the trend is shifting to the other areas of application of sequence pattern mining. One such area is web mining. Lot of research has been done to make the process of finding useful information and (interesting) knowledge from web data more efficient.

The current work is motivated by the candidate set and the test approach used in the basic AprioriAll algorithm (Agarwal & Agarwal, 1995). Similar to AprioriAll algorithm traversal of sequences takes place using the breadth first search technique. All the combinations of candidate set and frequent itemset takes place at the K-Level. As we are concentrated on web user traversals i.e, the user can visit back and froth a sites, the proposed algorithm considers the combinations of back and forth nature ((1,2) and (2,1)).Web data exhibit sequentiality in nature. The interrelationship among the web-pages visit with in a session can be used to predict the navigational behavior of the user.