

ABSTRACT

In any particular data mining problem, the most important task is to define patterns operationally. The model (definition of patterns) should ultimately lead to useful understandable patterns. Designing a good model and evaluating what patterns are generated from a particular model is difficult.

In this dissertation, we examine closely the problem of mining sequential patterns. Sequential pattern mining finds interesting patterns in sequences of sets. Since it has been proposed in [2], mining sequential patterns in large databases has become an important data mining task with broad applications.

Sequential pattern mining is commonly defined as finding the complete set of frequent subsequences. However, such conventional model may meet inherent difficulties in mining databases with long sequences and noises. They may generate a huge number of short trivial patterns but fail to find the underlying interesting patterns.

Motivated by these observations, we examine an entirely different model for analyzing sequential data. In this dissertation,

- We propose a novel model, *multiple alignment sequential pattern mining*. By lining up similar sequences and detecting the general trend, the multiple alignment model effectively finds consensus patterns that are approximately shared by many sequences.
- We demonstrate that **ApproxMAP** (for APPROXimate Multiple Alignment Pattern mining) is an efficient algorithm to find such consensus patterns.
- We design a general evaluation method for sequential pattern mining based on how well the underlying (embedded) patterns are recovered and how much confounding information is included in the results.
- We conduct an extensive performance study. The trend is clear and consistent. Together the consensus patterns form a succinct but comprehensive and accurate model of the sequential data. Furthermore, **ApproxMAP** is robust to its input parameters, robust to noise and outliers in the data, and scalable with respect to the size of the database. In comparison to the conventional model, **ApproxMAP** can better recover the underlying patterns with little confounding information under most circumstances.
- In addition, we report a successful case study on social welfare service patterns which illustrates that multiple alignment sequential pattern mining can find general, useful, concise and understandable knowledge and thus is an interesting and promising direction.