

Chapter 9

Conclusions

We conclude the dissertation with a summary of our research and a discussion of areas for future work.

9.1 Summary

In any particular data mining problem, the first and most important task is to define patterns operationally. The algorithms are only as good as the definition of the patterns (model). In this dissertation, we propose a novel model for sequential pattern mining, *multiple alignment sequential pattern mining*. Its goal is to organize and summarize sequence of sets to uncover the underlying consensus patterns. We demonstrate that multiple alignment is an effective model to find such patterns that are approximately shared by many sequences in the data.

We develop an efficient and effective algorithm, ApproxMAP (for APPROXimate Multiple Alignment Pattern mining), for multiple alignment sequential pattern mining. ApproxMAP uses clustering as a preprocessing step to group similar sequences, and then mines the underlying consensus patterns in each cluster directly through multiple alignment. A novel structure, weighted sequences, is proposed to summarize and compress the alignment information. The consensus sequences are then extracted from the weighted sequences via strength cutoffs. The strength cutoff is a powerful and expressive mechanism for the user to specify the level of detail to include in the consensus patterns.

To the better of our knowledge, this is the first study on mining consensus patterns from sequence databases. It distinguishes itself from the previous studies in the following two aspects. First, it proposes the theme of approximate sequential pattern mining, which reduces number of patterns substantially and provides much more accurate and informative insights into the sequential data. Second, it generalizes the multiple alignment techniques to handle sequences of itemsets. Mining sequences of itemsets extends the application domain substantially. The method is applicable to many interesting problems, such as social science

research, policy analysis, business analysis, career analysis, web mining, and security.

Our extensive evaluation demonstrates that **ApproxMAP** will effectively extract useful information by organizing the large database into clusters as well as give good descriptors (weighted sequences and consensus sequences) for the clusters using multiple alignment. We demonstrate that 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, scalable with respect to the size of the database, and in comparison to the conventional support model **ApproxMAP** can better recover the underlying patterns with little confounding information under most circumstances.

In addition, our case study on social welfare service patterns illustrates that approximate sequential pattern mining can find general, useful, concise, and understandable knowledge and thus is an interesting and promising direction.

9.2 Future Work

This dissertation is our first step towards the study of effective sequential pattern mining. Following the approximate frequent pattern mining model, many interesting research problems need to be solved.

First, more recent advances in multiple alignment come from Gibbs sampling algorithms, which use hidden Markov models [57]. These methods are better for local multiple alignment. Local multiple alignment is to find substrings of high similarity. Formally, given a set of strings, local multiple alignment first selects a substring from each string and then finds the best global alignment for these substrings. Since DNA sequences are very long, finding local similarity has many benefits. One possible future direction would be to expand **ApproxMAP** to do local alignment and investigate the benefits of local multiple alignment for sequences of sets in KDD applications.

Second, in the optimization of sample based iterative clustering the hash table implementation needs to be explored further. The optimization is made in order to speed up the running time for large datasets. But the current hash table implementation has a large memory requirement for large datasets. In our experiments, we ran out of memory for datasets $N_{seq} > 70000$ given 2GB of memory. We already know that there are other more efficient implementations of hash tables. But ultimately, to make our method practically scalable, we need to explore an implementation that stores only the possible proximity values (limited by the given memory size), and recalculate the other distances when needed. This will make the application work with any memory size and still give a significant reduction in time (Figure 6.6(d)).

The most interesting future direction is to expand the distance metric to be more comprehensive. First, it could be expanded to handle sequences of multisets or sets with quan-

titative information. Many of the data mining applications have sets that have more than one of the same item (multiset). For example, people buy many packs of diapers at once. If ApproxMAP could be expanded to handle multisets, it can find quantitative sequential patterns.

Second, user specified taxonomies could be used to customize the replacement cost. For example, two toys should be considered more similar to each other than a toy and a piece of furniture. Under the current model, {doll}, {crib}, and {ball} are all equally distant. If a user specified a taxonomy tree putting doll and ball under the same ancestor and crib in a separate branch, the distance metric could be expanded to a weighted distance metric which can incorporate this information.

Last, a practical improvement to ApproxMAP would be to automatically detect the best strength threshold, θ , for each cluster of sequences. An interesting approach could be analyzing the distribution of the item weights dynamically. Initial investigation seems to suggest that the item weights may follow the Zipf distribution. Closer examination of the distribution might give hints for automatically detecting statistically significant cutoff values customized for each cluster. When presenting an initial overview of the data, such approach could be quite practical.