論文の内容の要旨

論文題目　Fast Algorithms for Sequential Pattern Mining
（シーケンスパタンマイニングの高速アルゴリズムに関する研究）

氏 名　楊 征路

Sequential pattern mining, which extracts frequent subsequences from a sequence database, has attracted a great deal of interest during the recent surge in data mining research because it is the basis of many applications, such as customer behavior analysis, stock trend prediction, and DNA sequence analysis. The sequential mining problem was first introduced in ¥cite{Agrawal: SequenceMining}; two sequential patterns examples are: ``80% of the people who buy a television also buy a video camera within a day", and ``Every time Microsoft stock drops by 5%, then IBM stock will also drop by at least 4% within three days". The above patterns can be used to determine the efficient use of shelf space for customer convenience, or to properly plan the next step during an economic crisis. Sequential pattern mining is also very important for analyzing biological data, in which a very small alphabet (i.e., 4 for DNA sequences and 20 for protein sequences) and long patterns with a typical length of few hundreds or even thousands, frequently appear.

Efficient sequential pattern mining methodologies have been studied extensively in many related problems, including the general sequential pattern mining, constraint-based sequential pattern mining, incremental sequential pattern mining, frequent episode mining, approximate sequential pattern mining, partial periodic pattern mining, temporal pattern mining in data stream, maximal and closed sequential pattern mining.

Although there are so many problems related to sequential pattern mining explored, we realize that the general sequential pattern mining algorithm development is the most basic one because all the others can benefit from the strategies it employs, i.e., Apriori heuristic and projection-based pattern growth. Hence we aim to develop an efficient general sequential pattern mining algorithm in this thesis.

Much work has been carried out on mining frequent patterns, however, their performance is still far from satisfactory because of two main challenges: large search spaces and the ineffectiveness in handling dense data sets. To offer a solution

to the above challenges, we have proposed a series of novel algorithms, called the LAst Position INduction (LAPIN) sequential pattern mining, which is based on the simple idea that the last position of an item, α, is the key to judging whether or not a frequent k-length sequential pattern can be extended to be a frequent (k+1)-length pattern by appending the item α to it. LAPIN can largely reduce the search space during the mining process, and is very effective in mining dense data sets. Our performance study demonstrates that LAPIN outperforms PrefixSpan and SPADE by up to an order of magnitude on long pattern dense data sets.

However, we found that the improvement is at the price of much memory consuming when building the list of item's last position because LAPIN uses a bitmap strategy. We aim to obtain an efficient and balanced pattern mining algorithm with low memory consuming and thus, we proposed an improved algorithm which makes good use of not only the position of item but also the intermediate value (support value) of k-length pattern when fining (k+1)-length pattern. The experiments demonstrated that our improved algorithm performs the best in limited resource environments.

Ayres et al. claimed that SPAM is very efficient for long pattern mining and it can outperform PrefixSpan by up to an order of magnitude. Our experiments show that, although SPAM can handle long patterns in dense data sets, it is limited in the length of long patterns it can handle, and its high speed comes at a price of large space consumption. We proposed a new algorithm named LAPIN_SPAM, which combines the idea of LAPIN and SPAM. The experiments demonstrated that LAPIN_SPAM significantly outperforms the original SPAM, and is the best under unlimited resource assumption.

The WWW provides a simple yet effective media for users to search, browse, and retrieve information in the Web. Web log mining is a promising tool to study user behaviors, which could further benefit web-site designers with better organization and services. Although there are many existing systems that can be used to analyze the traversal path of web-site visitors, their performance is still far from satisfactory. In this thesis, we propose our effective Web log mining system based on our efficient sequential mining algorithm, LAPIN_WEB, an extension of previous LAPIN algorithm to extract user access patterns from traversal path in Web logs. Our experimental results and performance studies demonstrate that LAPIN WEB is very efficient and outperforms well-known PrefixSpan by up to an order of magnitude on real Web log datasets. Moreover, we also implement a visualization tool to help interpret mining results as well as predict users' future requests.

Recently, the skyline query has attracted considerable attention because it is the basis of many applications, e.g., multi-criteria decision making, user-preference queries and microeconomic analysis. Given an N-dimensional dataset D, a point p is said to dominate another point q if p is better than q in at least one dimension and equal to or better than q in the remaining dimensions. Skyline mining aims to find those non-dominated points, in a d-dimensional spatial dataset. This problem can be seen as a special class of pareto preference queries.

Efficient skyline querying methodologies have been studied extensively. However, all the papers concerned only the pure dominant relationship among a dataset, i.e., a point p is whether dominated by others or not, and got those non-dominated ones as results. However, in the real world, users are more interested in the detail of the dominant relationship in a dataset, i.e., a point p dominates how many other points and whom they are. This problem can be seen as a general dominant relationship analysis to the skyline query and has not been studied.

In this thesis, we find the interrelated connection between sequential pattern mining and the general dominant relationship. Based on this discovery, we propose efficient algorithms to answer the general dominant relationship queries by using efficient sequential pattern mining algorithms and several other strategies. Extensive experiments illustrate the effectiveness and efficiency of our methods.