

# Chapter 1

## Introduction

Knowledge discovery and data mining (KDD) has been defined as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from data" [20]. Clearly this definition serves more as a guideline than an operational definition. "Novel," "useful," and "understandable," are not quantitative criteria that can be implemented computationally. 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 (model). Designing a good model and evaluating what patterns are generated from a particular model is difficult. The model should ultimately lead to useful and understandable patterns.

In this dissertation, we examine closely the problem of mining sequential patterns. Sequential pattern mining finds interesting patterns in sequences of sets. Conventionally the sets are called *itemsets*. Since it has been proposed in [2], mining sequential patterns in large databases has become an important data mining task and has broad applications, such as social science research, policy analysis, business analysis, career analysis, web mining, and security.

Sequential pattern mining is commonly defined as finding the complete set of frequent subsequences in a set of sequences. However, conventional sequential pattern mining methods based on this *support model* may meet inherent difficulties in mining databases with long sequences and noises. They may generate a huge number of short and trivial patterns but fail to find the underlying patterns (interesting patterns approximately shared by many sequences) in the data.

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

- We propose a novel model for sequential pattern mining, *multiple alignment sequential pattern mining*, in databases with long sequences. 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 develop an efficient and effective algorithm, **ApproxMAP** (for APPROXimate Multiple Alignment Pattern mining), to mine consensus sequential patterns from large databases. **ApproxMAP** finds the underlying consensus patterns directly through multiple alignment.
- We design a general evaluation method for assessing the quality of any sequential pattern mining algorithm empirically. We then employ the method to conduct a detailed study of **ApproxMAP** as well as a through comparison study with the conventional model. We complete the evaluation with a real case study that illustrates the effectiveness of **ApproxMAP**.

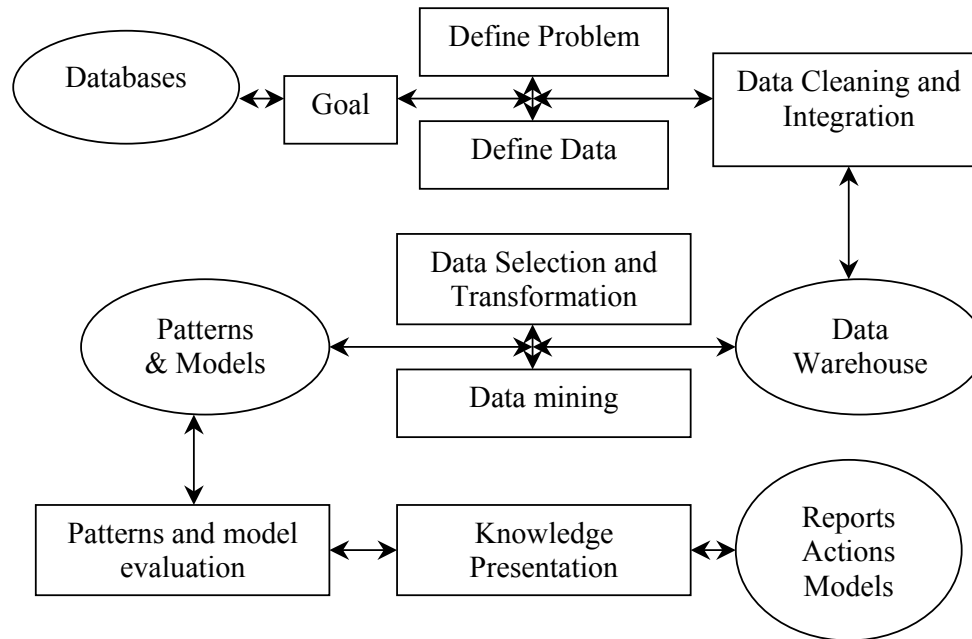
## 1.1 Knowledge Discovery and Data Mining (KDD)

Our society is accumulating massive amounts of data, much of which resides in large database management systems (DBMS). With the explosion of the Internet the rate of accumulation is increasing exponentially. Methods to explore such data would stimulate research in many fields. Knowledge discovery and data mining (KDD) is the area of computer science that tries to generate an integrated approach to extracting valuable information from such data by combining ideas drawn from databases, machine learning, artificial intelligence, knowledge-based systems, information retrieval, statistics, pattern recognition, visualization, and parallel and distributed computing [18, 20, 31, 56]. It has been defined as "The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [19]. The goal is to discover and present knowledge in a form, which is easily comprehensible to humans [19].

A key characteristic particular to KDD is that it uses "observational data, as opposed to experimental data" [33]. That is, the objective of the data collection is something other than KDD. Usually, it is operational data collected in the process of operation, such as payroll data. Operational data is sometimes called administrative data when it is collected for administration purposes in government agencies. This means that often times the data is a huge convenience sample. Thus, in KDD attention to the large size of the data is required and care must be given when making generalization. This is an important difference between KDD and statistics. In statistics, such analysis is called secondary data analysis [33].

In KDD, useful information extracted are often in the form of (1) previously unknown relationships between observations and/or variables or (2) summarization or compression of the huge data in novel ways allowing humans to "see" the large bodies of data. The relationships and summaries must be new, understandable, and potentially useful [18, 33]. "These relationships and summaries are often referred to as models or patterns" [33]. Roughly, models are global summaries of a data set while patterns are local descriptions of a subset of the data [33]. As done in this dissertation, sometimes a group of local patterns can be a

Figure 1.1: The complete KDD process



global model that summarizes a data set.

This dissertation deals with summarizations and compressions. The summarizations and compressions must be done "in such a way that the result is more comprehensible, without any notion of generalization" [33]. The key to successful summarization is how you view the data and the problem. It requires flexibility and creativity in finding the proper definitions for both. As with much in life, one innovative perspective can give you a simple answer to a complicated problem. The difficulty is in realizing the right point of view for the given question and data. As such, it is essential to interpret the summaries in the context of the defined problem and understanding of the data.

An important aspect of KDD is that it is an ongoing iterative process. Following are the steps in the iterative KDD process along with an example from the real world [33]:

1. **Data Acquisition:** The raw data is collected usually as a by product of operation of the business.
  - As various welfare programs are administered, many raw data have been collected on who has received what welfare programs in the past 5 years. One might be a database on all the TANF <sup>1</sup> welfare checks issued and pulled.
2. **Choose a Goal:** Choose a question to ask the database.

<sup>1</sup>TANF (Temporary Assistance of Needy Families) is the monthly cash assistance welfare program since Welfare Reform.

- There are data on various welfare program participation. From the database, one could be interested in the following questions: What are the common patterns of participation in these welfare programs? What are the main variations?
3. **Define the Problem:** Define the problem statement so that the data can give the answer.
    - What are the underlying patterns in the sequences of sets? (See chapter 3 for details)
  4. **Define the Data:** Define/view the data to answer the problem statement?
    - Organize the welfare data into sequences of sets of welfare programs received during a particular month per person. See Table 1.1 in section 1.2.1 for an example.
  5. **Data Cleaning and Integration:** Remove noise and errors in the data and combine multiple data sources to build the data warehouse as defined in the "Define the Data" step.
    - Cleaning: Clean all bad data (such as those with invalid data points)
    - Integration: Merge the appropriate database on different welfare programs by recipients.
  6. **Data Selection and Transformation:** Select the appropriate data and transform them as necessary for analysis.
    - Selection: Separate out adults and children for separate analysis.
    - Transform: Define participation for each program and transform the data accordingly. For example, if someone received at least one TANF welfare check then code as T for that month.
    - Transform: Build the transformed welfare program participation data into actual sequences.
  7. **Data Mining:** The essential step where intelligent methods are applied to extract the information.
    - Apply ApproxMAP to the sequences of monthly welfare programs received. (See chapter 4 for details)
  8. **Patterns and Model Evaluation:** Identify interesting and useful patterns.

- View the consensus sequences generated by **ApproxMAP** for any interesting and previously unknown patterns. Also, view the aligned clusters of interest in more detail and use pattern search methods to confirm your findings.

9. **Knowledge Presentation:** Present the patterns and model of interest as appropriate to the audience.

- Write up a report on previously unknown and useful welfare patterns of participation in welfare programs for policy makers.

Figure 1.1 shows the diagram of the complete KDD process [18, 31]. Although the process is depicted in a linear fashion keep in mind that the process is iterative. The earlier steps are frequently revisited and revised as needed while future steps have to be taken into account when completing earlier steps. For example, when defining the problem, along with the databases and the goal one should consider what established methods of data mining might work best in the application. This dissertation presents a novel data mining technique along with the appropriate problem and data definitions.

Data analysts have different objectives for utilizing KDD. Some of them are exploratory data analysis, descriptive modeling, predictive modeling, discovering patterns and rules, and retrieval of similar patterns when given a pattern of interest [33]. The objective of this dissertation is to assist the data analyst in exploratory data analysis by descriptive modeling. The models are built through loss compression and data reduction. "The goal of descriptive modeling is to describe all the data" [33] through the model. These models built in KDD are empirical by nature. Such models are "simply a description of the observed data" [33] and should not be viewed outside the context of the data. "The fundamental objective [of the model] is to produce insight and understanding about the structure of the data, and see its important features" [33].

## 1.2 Sequential Pattern Mining

Classical exploratory data analysis methods used in statistics and many of the earlier KDD methods tend to focus only on basic data types, such as interval or categorical data, as the unit of analysis. However, some information cannot be interpreted unless the data is treated as a unit, leading to complex data types. For example, the research in DNA sequences involves interpreting huge databases of amino acid sequences. The interpretation could not be obtained if the DNA sequences were analyzed as multiple categorical variables. The interpretation requires a view of the data at the sequence-level. DNA sequence research has developed many methods to interpret long sequences of alphabets [25, 30, 57].

In fact, sequence data is very common and "constitute a large portion of data stored in computers" [5]. For example, data collected over time is best analyzed as sequence data.

**Table 1.1: Different examples of monthly welfare services given to clients in sequential form**

clientID	Sequential Data
A	{AFDC (A), Medicaid (M), Food Stamp (FS)} {A, M, FS } {M, FS} {M, FS} {M, FS} {FS}
B	{Report (R)} {Investigation (I), Foster Care (FC)} {FC, Transportation (Tr)} {FC} {FC, Tr}
C	{Medicaid (M)} {AFDC (A), M} {A, M} {A,M} {M, Foster Care (FC)} {FC} {FC}

**Table 1.2: A segment of the frequency of combinations of the first four events table**

Pattern ID	Sequential Pattern	Frequency	Percentage
1	Medicaid Only	95,434	55.2%
2	Medicaid Only $\Rightarrow$ AFDC	13,544	7.8%
3	Medicaid Only $\Rightarrow$ AFDC $\Rightarrow$ Foster Care	115	0.1%

Analysis of these sequences would reveal information about patterns and variation of variables over time. Furthermore, if the unit of analysis is a sequence of complex data, one can investigate the patterns of multiple variables over time. When appropriate, viewing the database as sequences of sets can reveal useful information that could not be extracted in any other way.

Analyzing suitable databases from such a perspective can assist many social scientists and data analysts in their work. For example, all states have administrative data about who has received various welfare programs and services. Some even have the various databases linked for reporting purposes [26]. Table 1.1 shows examples of monthly welfare services given to clients in sequential form. Using the linked data it is possible to analyze the pattern of participation in these programs. This can be quite useful for policy analysis: What are some commonly occurring patterns? What is the variability of such patterns? How do these patterns change over time? How does certain policy changes effect these patterns?

### 1.2.1 Sequence Analysis in Social Welfare Data

However, the methods to deal with such data (sequences of sets) are very limited. In addition, existing methods that analyze basic interval or categorical data yield poor results on these data due to exploding dimensionality [32].

Conventional methods used in policy analysis cannot answer the broad policy questions such as finding common patterns of practice. Thus, analysts have been forced to substitute their questions. Until recently, simple demographic information was the predominant method used (66% of those receiving Food Stamp also received AFDC<sup>2</sup> benefits in June). Survival analysis is gaining more popularity but still only allows for analyzing the rate of occurrence of some particular events (50% of participants on AFDC leave within 6 months). In [26], they studied specific transitions of interest (15% of children who entered AFDC in Jan 95, were in foster care before entering AFDC). These methods are very limited in their ability to describe the whole body of data.

<sup>2</sup>AFDC (Aid to Families with Dependent Children) was the cash assistance welfare program that was the precursor TANF.

Thus, some have tried enumeration [26, 58] - frequency counts of participants by program participation. Table 1.2 gives a small section of a table included in the technical report to Department of Health and Human Services (HHS) [26]. It reports the frequency of combinations of the first four events. For example the third line states that 0.1% of the children (115 sequences) experienced "Medicaid only" followed by "AFDC" followed by "foster care". Client C in Table 1.1 would be encoded as having such a pattern.

There are numerous problems with enumeration. First, program participation patterns do not line up into a single sequence easily. Most people receive more than one service in a particular month. For example, most clients receiving AFDC also receive Medicaid. As a work around, analysts carefully define a small set of alphabets to represent programs of most interest and its combinations. In [26], only three programs, AFDC, Medicaid, and foster care, were looked at. In order to build single sequences, they defined the following five events as most interesting.

- Medicaid Only
- AFDC: probably receiving Medicaid as well but no distinction was made as to whether they did or not
- Foster care: could be receiving Medicaid in conjunction with foster care, but no distinction was made
- No services
- Some other services

Second, even with this reduction of the question, many times the combinatoric nature of simple enumeration does not give you much useful information. In the above example, looking at only the first four events, the number of possible patterns would be  $5^4 = 625$ . Not surprisingly, there are only a few simple patterns such as, "AFDC  $\Rightarrow$  Some other service", that are frequent. The more complex patterns of interest do not get compressed enough to be comprehensible. The problem is that almost all the frequent patterns are already known and the rest of the information in the result is not easily understandable by people. There were a total of 179 patterns reported in the analysis [26].

A much better method developed in sociology, *optimal matching*, has not yet gained much popularity in social sciences [1]. Optimal matching is a direct application of pattern matching algorithms developed in DNA sequencing to social science data [25, 30]. It simply applies the hierarchical edit distance to pairs of simple sequences of categorical data and runs the similarity matrix through standard clustering algorithms. The researcher looks at the clusters and tries to manually organize and interpret the clusters. This is possible because up to now researchers have only used it for fairly small data sets that were collected and coded manually. Optimal matching could be applied to the data discussed in the previous paragraph. It should give more useful results than simple enumeration. Then the analysis would not be limited to the most important programs or the first  $n$  elements.

There are two problems with optimal matching. First, you are limited to strings. Thus, one could not handle multiple services received in one month very well. In real datasets used in social science, sequences of sets are much more common than sequences of letters. Second, once the clustering and alignment is done there is no mechanism to summarize the cluster information automatically. Finding consensus strings from DNA sequences have not been applied to social science data yet. Thus, the applicability needs to be investigated. Without some automatic processing to produce cluster descriptors social scientists would be limited to very small data sets.

### 1.2.2 Conventional Sequential Pattern Mining

Traditionally in data mining, sequential patterns have been defined as a subsequence that appears frequently in a sequence database [2]. Such definition leads to a *support* model based sequential pattern mining methods. Currently as far as we know, the only methods available for analyzing sequences of sets is based on such a support model.

Although the support model based sequential pattern mining has been extensively studied and many methods have been proposed [4, 34, 48, 55, 66], there are some inherent obstacles within the conventional model. These methods meet inherent difficulties in mining databases with long sequences and noise. These limitations are discussed in detail in chapter 7.

More importantly, finding frequent subsequences will not answer the policy questions about service patterns in social welfare. In order to find common service patterns and the main variations, the mining algorithm must find the general underlying trend in the sequence data through summarization. However, the conventional methods have no mechanism for summarizing the data. In fact, often times many more patterns are output from the method than the number of sequences input into the mining process. These methods tends to generate a huge number of short and trivial patterns but fail to find interesting patterns approximately shared by many sequences.

## 1.3 Multiple Alignment Sequential Pattern Mining

In this dissertation, we present a new approach to sequential analysis, *multiple alignment sequential pattern mining*, that can detect common patterns and their variations in sequences of sets. Multiple alignment sequential pattern mining partitions the database into similar sequences, and then summarizes the underlying pattern in each partition through multiple alignment.

The exact solution to multiple alignment sequential pattern mining is too expensive. Here, we design an effective and efficient approximate solution, **ApproxMAP**, to mine consensus patterns from large sequence databases. Our goal is to assist the data analyst in exploratory data analysis through organization, compression, and summarization. **ApproxMAP** has three

steps. First, similar sequences are grouped together using  $k$ -nearest neighbor clustering. Then we organize and compress sequences within each cluster into *weighted sequences* using multiple alignment. In the last step, the weighted sequences for each cluster is summarized into the longest consensus pattern best fitting each cluster via user specified *strength* cutoff. We use color to visualize item strength. *Item strength*, color in the consensus sequence, indicates how many sequences in the cluster contain the item in that position.

## 1.4 Evaluation

It is important to understand the approximating behavior of **ApproxMAP**. The accuracy of the approximation can be evaluated in terms of how well it finds the real underlying patterns and whether or not it generates any spurious patterns. However, it is difficult to calculate analytically what patterns will be generated because of the complexity of the algorithm.

As an alternative, in this dissertation we have developed a general evaluation method that can objectively evaluate the quality of the results produced by any sequential pattern mining method. It uses the well known synthetic data generator built by [2]. The evaluation is based on how well the base patterns are recovered and how much confounding information (in the form of spurious patterns, redundant patterns, or extraneous items) is in the results. The base patterns, which are output along with the database by the data generator in [2], are the patterns used to generate the data.

We conduct an extensive and systematic performance study of **ApproxMAP** using the evaluation method. The trend is clear and consistent. The results show that **ApproxMAP** (1) is robust to the input parameters, (2) returns a succinct but accurate summary of the base patterns with few redundant or spurious patterns, (3) is robust to both noise and outliers in the data, and (4) is effective and scalable in mining large sequence databases with long patterns.

We further employ the evaluation method to conduct a comparative study of the conventional support model and the multiple alignment model. To the best of our knowledge, no research has examined in detail what patterns are actually generated from the popular support model for sequential pattern mining. The results clearly demonstrate that too much confounding information is generated. With so much redundant and spurious patterns there was no way to pick out the primary underlying patterns in the results. In addition, our theoretically analysis of the expected support of random sequences under the null hypothesis demonstrate that support alone cannot distinguish between statistically significant patterns and random occurrences. Furthermore, the support model is vulnerable to noise in the data because it is based on exact match. We demonstrate that in comparison, the sequence alignment model is able to better recover the underlying patterns with little confounding information under most circumstances including those where the support model fails.

We complete the evaluation by reporting on a successful case study of **ApproxMAP** in a real application: mining the North Carolina State welfare services database. We illustrate some interesting patterns mined by **ApproxMAP**. Moreover, the mining result of **ApproxMAP** triggered some important investigations by our clients. The application results show that approximate sequential pattern mining can find interesting patterns that are highly promising in many applications.

## 1.5 Thesis Statement

The author of this dissertation asserts that multiple alignment is an effective model to uncover the underlying trend in sequences of sets. We will show that **ApproxMAP**, a novel method to apply multiple alignment techniques to sequences of sets, 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. Furthermore, we will show that **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, we will demonstrate the usefulness of **ApproxMAP** using real world data.

## 1.6 Contributions

This dissertation makes the following contributions:

- defines a new model for sequential analysis based on multiple alignment, *Multiple Alignment Sequential Pattern Mining*,
- describes a novel solution **ApproxMAP** (for APPROXimate Multiple Alignment Pattern mining) that (1) introduces a new metric for itemsets, (2) a new representation of alignment information for sequences of sets (weighted sequences), and (2) the effective use of strength cutoffs to control the level of detail included in the consensus patterns,
- designs a general evaluation method to assess the quality of results from any sequential pattern mining algorithm,
- employs the evaluation method to conduct an extensive set of empirical evaluations of **ApproxMAP** on synthetic data,
- employs the evaluation method to compare the effectiveness of **ApproxMAP** to the conventional methods based on the support model,

- derives the expected support of random sequences under the null hypothesis of no patterns in the database to better understand the behaviour of the support model,
- and demonstrates the usefulness of **ApproxMAP** using real world data.

## 1.7 Synopsis

The rest of the dissertation is organized as follows. Chapter 2 reviews related works. Chapter 3 defines the model, *Multiple Alignment Sequential Pattern Mining*. Chapter 4 details the method, **ApproxMAP**. Chapter 5 introduces the evaluation method and chapter 6 reports the evaluation results. Chapter 7 reviews the conventional support model, details its limitations, and reports the results of the comparison study. Chapter 8 presents a case study on real data. Finally, in chapter 9, we conclude with a summary of our research and discuss areas for future work.