INTRODUCTION

Data mining and knowledge discovery is an interdisciplinary field merging ideas and concepts from databases, statistics, machine learning, lattice theory etc. The phenomenal and explosive growth of data in different spheres of human endeavour and the impending need to extract useful information from large amounts of data has been the driving force for the proliferation of data mining efforts in different directions. Formally, **Data Mining** is defined as the non-trivial extraction of implicit, previously unknown and potentially useful information from data.

Data mining is part of a bigger framework referred to as Knowledge Discovery in Databases (KDD) that covers complex processes, from data preparation to knowledge modeling. Figure 1.1 (Han and Kamber, 2001) depicts the architecture of a typical data mining system. Within this process, data mining techniques and algorithms are the actual tools that analysts have at their disposal to find unknown, interesting patterns and correlate on the occurrences of patterns in data. Data cleaning and integration techniques are applied to the raw data obtained from data warehouses and other repositories. The database server fetches the relevant data, based on the user’s request and the domain knowledge is used to guide the search. The data mining engine is a set of functional modules for tasks such as characterization, association, classification, cluster analysis and deviation analysis. The pattern evaluation module employs interestingness measures and interacts with the data mining modules, so as to focus the search towards
interesting patterns. The graphical user interface which communicates between the users and the data mining system, helps focus the search and performs exploratory data mining based on intermediate results.

![Architecture of a Typical Data Mining System](image)

Figure 1.1. Architecture of a Typical Data Mining System

Typical data mining tasks are classification, clustering and association rule mining. A host of algorithms have been developed to perform these tasks. The techniques merge many fields of science from statistics through neural and fuzzy computing. Many of the techniques have their roots in fields like Lattice theory, Formal Concept Analysis and Representation Theory.
The earliest applications of data mining were predominantly focused on customer purchasing behaviour. Market Basket Analysis is one of the most common applications in this category. However, at present data mining is being applied in several areas of traditional and social sciences as well. Among the traditional sciences astronomy, physics, biology and medicine provide rich sources of applications to data miners. An important and widely researched application is the World Wide Web. The web provides the ability to access one of the largest data repositories and there is immense potential for the application of data mining techniques in this area. Among other applications are stock market databases, biological databases and the human genome project which are excellent targets for the application of data mining techniques.

Two main forms of data mining identified in KDD applications are verification driven data mining and discovery driven data mining (Simoudis, 1996). In verification driven data mining, the user postulates a hypothesis and the system tries to validate it. The common verification driven operations include query and reporting, multidimensional analysis, and statistical analysis. Discovery driven mining, on the other hand automatically extracts new information. Typical discovery driven tasks are:

- **Discovery of Association Rules**

  Association Rule Mining (ARM) finds interesting associations or correlation relationships among a large set of data items. With massive amounts of data being continuously collected and stored, many industries benefit from deriving interesting associations from items in their database. The discovery of interesting association relationships among huge amounts of business transaction records helps in many decision making processes, such as catalog design, cross marketing, formulating marketing strategies and customer profiling.
Sequential Pattern Discovery

Sequence discovery aims at extracting sets of events that commonly occur over a period of time (Agrawal and Srikant, 1995). An example of a sequential pattern could be that “70% of the people who purchase a computer also purchase a web camera within a month”. Such sequence patterns are useful in retail sales, monitoring network failures and medical diagnosis.

Classification and Regression

Classification aims at assigning a new data item to one of several predefined categorical classes. Since the field being predicted is pre-labeled, classification is also known as supervised induction (Weiss and Kulikowski, 1991; Michie et al. 1994). While there are several classification methods including neural network and genetic algorithms, decision trees are particularly suited to classification, since they can be constructed relatively quickly, are simple and easy to understand. While classification predicts a categorical value, regression is applied if the field comes from a real-valued domain. Common applications of classification include fraud detection in credit card usage and insurance risk analysis.

Clustering

Clustering is used to partition a database into subsets or clusters, such that elements of a cluster share a set of common interesting properties that distinguish them from other clusters (Jain and Dubes, 1988; Cheeseman et al. 1988). Unlike classification which has predefined labels, clustering must in essence automatically come up with the labels. For this reason, clustering is also called unsupervised induction. Applications of clustering include demographic or market segmentation for identifying common traits in groups of people, discovery of new types of stars in datasets of stellar objects and so on.
• Similarity Search

Similarity search is performed on a database of objects to find objects that are within a user defined distance from the queried object, or to find all pairs within some distance of each other (Agrawal et al. 1993b; Faloutsos et al. 1994). This kind of search is especially applicable to temporal and spatial databases. Example applications include discovering stocks with similar price movements, identifying companies with similar sales patterns etc.

Since the focus of the research work is on association rules and incremental mining, section 1.1 elaborates on various classifications and applications of ARM. The relevance of incremental mining in the present scenario is also discussed. Open issues and research avenues in the field of ARM is presented in section 1.2. The main contributions of the thesis is outlined in section 1.3.

1.1 ASSOCIATION RULE MINING

With massive amounts of data being continuously collected and stored, it becomes imperative for business organizations to derive interesting patterns and trends from this data. This discovery of interesting relationships from huge amounts of business transaction records helps in many decision making processes, such as catalog design, cross-marketing and sales promotional offers. A typical application of association rule mining is Market Basket Analysis (Agrawal, 1993). This process analyzes customer buying habits by finding associations between the different items that customers place in their “shopping baskets” as shown in Figure.1.2.
The discovery of such associations can help retailers develop marketing strategies by gaining insight about items that are frequently purchased together by customers. A typical supermarket example could be analyzing if, “customers who buy soft drinks also tend to buy chips on the same trip to the supermarket”. Such information can lead retailers to do selective marketing strategies to promote sales of particular products and also plan the store layout. Such a buying pattern, may induce the retailer to place soft drinks and chips within close proximity, or place them further apart with the in between space being allocated to infrequently purchased products, so that they get noticed when the customer crosses them to pick up the chips.

In a broader sense, the association rules identify relationships among a set of items in a database. These relationships are not based on the inherent properties of the data themselves, but rather based on co-occurrence of the data items. The following example further illustrates the use and objective of inferring different associations among items in a grocery store.
A grocery store has weekly specials for which advertising supplements are created for the local newspaper. When an item, such as peanut butter, has been designated to go on discount sale, the management determines the other items that are frequently purchased along with peanut butter. They find that bread is purchased 30% of the time and jelly is purchased with it 40% of the time. Based on this association, special displays of jelly and bread are placed near the peanut butter, which is on sale. They also decide not to offer any discount on jelly and bread. These actions are aimed at increasing overall sales volume by taking advantage of the frequency with which these items are purchased together.

There are two associations of interest in the above example. The first one states that, when peanut butter is purchased, bread is purchased 30% of the time. The second one states that 40% of the time, whenever peanut butter is purchased so is jelly. The discovered association rules can be used by the management to increase the effectiveness and reduce the cost associated with advertising, marketing, inventory and shelf placement. While this is a typical example of the use of association rules in Market Basket Analysis, it is now extensively used in applications, such as failure predictions in telecommunication networks and in establishing trends in the stock market.

A formal definition of the association rule (Agrawal, 1993) is given below:

**Definition 1.1:** Let \( I = \{I_1, I_2, \ldots, I_m\} \) be a set of \( m \) distinct attributes, also called literals. Let \( D \) be a database, where each record (tuple \( T \)) is a unique identifier and contains a set of items such that \( T \subseteq I \). An association rule is an implication of the form \( X \Rightarrow Y \), where \( X, Y \subseteq I \) are sets of items called itemsets, and \( X \cap Y = \emptyset \). Here, \( X \) is called the antecedent and \( Y \) the consequent.

Not all rules are interesting. Two measures which determine the interestingness of rules are the "support" and confidence which are defined as follows:
**Definition 1.2:** The support ($S$) of an association rule is the ratio (in percentage) of the records that contain $X \cup Y$ to the total number of records in the database.

Therefore, when we say that the support of a rule is 5%, it means that 5% of the total records contain $X \cup Y$. Support is the statistical significance of an association rule. Grocery store managers probably would not be concerned about, how peanut butter and bread are related if less than 5% of the total transactions have this combination. A high support would render the rule interesting in this case. However, this is not always the case, if we were using association rules to predict the failure of telecommunications switching nodes, based on the set of events that occur prior to failure, then even if the events were not very frequent, they would still be considered important. The user-defined threshold of support is therefore largely determined by the nature of the application.

**Definition 1.3:** For a given number of records, confidence is the ratio (in percentage) of the number of records that contain $X \cup Y$ to the number of records that contain $X$.

Thus, if we say that a rule $X \Rightarrow Y$ has a confidence of 85%, it means that 85% of the records containing $X$ also contain $Y$. The confidence of a rule indicates the degree of correlation between $X$ and $Y$ in the dataset. Confidence is a measure of the rule's strength. Often a large confidence is required for association rules.

Most of the applications in ARM involve finding all rules that meet the user specified threshold support and confidence. Association rules can be classified in various ways based on the several criteria.
1.1.1 Classification of Association Rules

- **Boolean and Quantitative Association Rules**

If a rule concerns associations between the presence or absence of items, it is a Boolean association rule. Given below is an example of a Boolean association rule obtained from market basket analysis.

\[
\text{Computer} \Rightarrow \text{Web Camera} \quad (\text{Support} = 2\%, \ \text{Confidence} = 60\%) \quad - \text{Rule 1}
\]

A support of 2% for the above rule means that 2% of all the transactions under analysis, show that computer and web camera are purchased together. A confidence of 60% means that, 60% of the customers who purchased a computer also purchased a web camera.

If a rule describes associations between quantitative items or attributes, then it is a quantitative association rule. In these rules, quantitative values for items or attributes are partitioned into intervals. Given below is a quantitative association rule, where X is a variable representing a customer.

\[
\text{Age}(X, \text{"30 .. 39"}) \land \text{income}(X, \text{"42k .. 50k"}) \Rightarrow \text{buys } (X, \text{Plasma TV}) \quad - \text{Rule 2}
\]

Here the quantitative attributes, age and income have been described.

- **Single Dimension and Multidimensional Association Rules**

If the items or attributes in an association rule reference only one dimension, then it is a single dimensional association rule. Rule (1) could be rewritten as

\[
\text{buys}(X, \text{"Computer"}) \Rightarrow \text{buys}(X, \text{"Web Camera"}) \quad - \text{Rule 3}
\]

The above rule is a single dimensional association rule since it refers to only one dimension, “buys”. If a rule references two or more dimensions, then it is a
multidimensional association rule. Rule (2) is considered a multidimensional rule since it involves three dimensions, “age”, “income” and “buys”.

- **Single Level and Multilevel Association Rules**

  Association rules can also be classified based on the levels of abstraction. For example, consider an association rule mining system which includes the following rules:

  \[
  \text{Age} (X, \text{“30 \ldots 39”}) \Rightarrow \text{buys}(X, \text{“Laptop Computer”}) \quad - \text{Rule 4}
  \]

  \[
  \text{Age} (X, \text{“30 \ldots 39”}) \Rightarrow \text{buys} (X, \text{“Computer”}) \quad - \text{Rule 5}
  \]

  In the above two rules, the items bought are referenced at different levels of abstraction. Computer is a higher level abstraction of “laptop computer”. The rule set mined, consists of multilevel association rules. If the rules within a given set do not reference items or attributes at different levels of abstraction, then the set contains single level association rules. For many applications, it is difficult to find strong associations among data items at low or primitive levels of abstraction due to sparsity of data in multidimensional space. Strong associations discovered at high concept levels may represent common sense knowledge. However, what may represent common sense to one user may seem novel to another. Therefore, data mining systems should provide capabilities to mine association rules at multiple levels of abstraction and traverse easily among different abstraction spaces. Concept hierarchies can be used to mine multiple level or multilevel association rules (Han et al.2001). Concept hierarchies provide useful background knowledge for expressing discovered patterns in concise, high level terms and facilitate mining of knowledge at multiple levels of abstraction.
1.1.2 Mining on Different Kinds of Data

Data mining techniques have focused on mining relational databases and data warehouses formed by the transformation and integration of structured data. Vast amounts of data, in various complex forms have been growing explosively owing to the rapid progress of data collection tools, advanced database system techniques and World Wide Web technologies. Therefore, an increasingly important task of data mining is to mine complex types of data including complex objects, spatial data, multimedia data, time-series data, text data and the World Wide Web.

- Multidimensional analysis and data mining can be performed in object relational and object oriented databases by:
  
  (i) Class based generalization of complex objects, including set-valued, list-valued and other sophisticated types of data, class/subclass hierarchies and class composition hierarchies.
  
  (ii) Constructing object data cubes.
  
  (iii) Performing generalization based mining.

- Spatial data mining is the discovery of interesting patterns from large geospatial databases. Spatial data cubes that contain spatial dimensions and measures can be constructed. Spatial OLAP can be implemented to facilitate multi dimensional spatial data analysis. Spatial data mining includes spatial data description, classification, association, clustering, spatial trends and outlier analysis.

- Multimedia data mining is the discovery of interesting patterns from multimedia databases that store and manage large collection of multimedia objects, including audio, image, sequence and hypertext data containing text, text markups and linkages. Issues in multimedia data mining include content
based retrieval and similarity search, generalization and multidimensional analysis, classification and prediction analysis, and mining associations in multimedia data.

- A time series database consists of a sequence of values or events changing with time, such as stock market data, business transaction sequences, dynamic production processes, medical treatments, web page access sequences and so on. Research issues in time series and sequence data mining cover trend analysis, similarity search in time series analysis, and mining sequential and periodic patterns in time related data.

- Text mining has become increasingly important, since a substantial portion of the available information is stored in text or document databases that consist of a large collection of documents, such as news articles, technical papers, books, digital libraries, email messages and web pages. Text mining goes a step beyond keyword based and similarity based information retrieval, to discover knowledge from semi structured text data using methods such as keyword based association and document classification.

- The WWW serves as a huge, widely distributed global information service center for news channels, advertising agencies, financial management, education, government, e-commerce and many other services. Web mining includes mining web linkage structure, web content and web access patterns.

In developing techniques for web usage, first the success of web log file analysis depends on what and how much valid and reliable knowledge is discovered from large, raw log data. Secondly, with the available URL, time, IP addresses and web page content information, a multidimensional OLAP analysis can be performed to find the top N users, top N accessed web pages, most frequently accessed time periods, and so on, which help discover potential customers, users, markets and other information on customer preferences.
Third, data mining can be performed on web log records to find association patterns, sequential patterns and trends in web access. Study of web log files can help in improving system design and performance by web caching, web page Prefetching and web page swapping. Adaptive web sites could be designed that improvise their structure by learning from user access patterns.

The web is a highly dynamic source of information, growing at a rapid pace. Its information is also constantly updated. It is also said that 99% of the web information is useless to 99% of web users (Han et al. 2001). Although, this may not seem obvious, it is true that a particular person is generally interested in only a tiny portion of the web, while the rest contains information that is uninteresting to the user and may swamp desired search results. How can the portion of the web that is truly relevant to one’s interest be determined? How can one find high quality web pages on a specific topic? How can previous results of web queries be utilized to improve the search? All these challenges have promoted research in several directions laying emphasis on concept hierarchies and incremental mining strategies during the last decade. Research in this direction also heralds the semantic web of the future. In the next section, the need for incremental mining techniques in the current scenario is examined.

1.1.3 Incremental Mining in Dynamic Databases

Incrementality is a major challenge in data mining. Incremental mining focuses on the maintenance of discovered patterns over time as data is continuously being added to the database. Agrawal et al. (1994) originally proposed the term “Incremental mining”. It is increasingly being applied in web applications, time series databases and other dynamic databases. To stay competitive in a fast paced environment like the World Wide Web, companies must be able to extract knowledge from their web access logs, web transaction
logs and web user profiles. However, the vast amount of web data makes batch processing and recomputation in mining of log data unreliable, therefore online and incremental mining techniques become indispensable in the quest for cutting edge knowledge.

In dynamic databases, users may periodically or occasionally insert or remove data from the database. The update to the database may cause not only the generation of new rules but also invalidation of some existing rules. Thus the efficient maintenance of the association rules is significant, paving the way for many incremental strategies. Such strategies focus on reducing the cost of processing the updated database by making use of previously obtained results.

In inductive databases, it makes sense to exploit the effort already done by the DBMS with previous queries. Applying incremental techniques, the mining engine can "reuse" some of the information contained in them, in order to reduce the computational effort (Meo et al. 2006).

Several incremental strategies based on context dependent constraints have been proposed. Context dependent constraints occur in many important application domains such as stock market analysis (Liu et al. 2002), meteorological forecasts (Feng et al. 2001) and in Market Basket Analysis (Grahne et al. 2001).

Analysis of stock market data can be explained with an example. Let us assume that the database to be analyzed contains the following attributes.

Date - the date of interest;
Stock - the name of the stock;
Price - a categorical attribute assuming values in {increased, constant, decreased}

In such a context, the user may be interested in knowing, whether there exists any negative correlation between groups of stock items. For instance, it may want to
associate stocks for which \textit{price-increased} with ones for which \textit{price-decreased}. From the mining results one will be able to discover that \textit{\textless \textless when the price of AT&T and Microsoft increase, the price of Sun Microsystems decreases with a probability of 78\%\textgreater \textgreater}.

The presence of context dependent constraints in the above case implies that, one needs to carefully check whether the constraints are satisfied by scanning the transaction table. Incremental algorithms are valuable tools in this setting and make use of available results from previous queries.

\section{1.2 OPEN ISSUES}

Open issues and research avenues in the field of association rule mining can be arranged into the following groups:

- Constraining the exploration, through the development and incorporation of measures of interest, efficient traversal and pruning strategies.
- Reducing I/O through efficient pruning of the datasets, and removing the necessity of large datasets to be memory resident.
- Creating useful data structures to make the analysis more tractable.
- Producing condensed inference sets, allowing the entire results to be inferred from a reduced set of inferences, reducing storage and computing costs.
- Specialization of fundamental association mining algorithms to address specific issues, such as the development of incremental algorithms to facilitate dynamic dataset mining and the inclusion of additional semantics such as time, space and ontologies.
1.3 THESIS CONTRIBUTIONS

The thesis focuses on the development of strategies for frequent itemset mining by exploring certain lattice theoretic principles in order to address I/O, storage and search space pruning issues. Efficient hybrid techniques for mining frequent and maximal frequent itemsets are proposed. Since many of the association rule mining applications are based on dynamic databases, an incremental approach to closed itemset mining is also proposed. Lattice Theory and Formal concept Analysis forms the basic framework for the design of the various strategies. The use of context transformations as a preprocessing mechanism is also presented. The rest of the thesis is organized as follows:

Chapter 2 highlights the connections between Formal Concept Analysis (FCA) and Knowledge Discovery. The use of FCA in knowledge representation and analysis is discussed.

Chapter 3 presents a survey of association rule mining, detailing the evolution from the seminal to the state-of-the-art. The extant literature has been classified based on several criteria.

Chapter 4 discusses the hybrid strategies which have been developed using a lattice framework. Search space decomposition and optimal application of the closure properties to prune the search space are based on lattice theoretic principles. Experimental results on two types of data organizations using synthetic and real datasets are presented and analyzed. Parts of this work has appeared in Kalpana et al. (2007a; 2007b; 2008a).

Chapter 5 presents the Indexed Trie approach to incrementally mine the closed itemsets. Since databases tend to be dynamic, it is observed that the maintenance of the association rules can be directly mapped into the problem of maintaining closed frequent itemsets. The proposed framework emphasizes on
certain fundamental and structural properties of Galois lattice theory to overcome
the limitations of earlier approaches. Experimental results are compared with some
existing strategies. Reducing the constraint on the main memory and a selective
update using the indexed trie approach are the significant contributions. Part of the
work is published in Kalpana et al. (2008b; 2008c).

Context clarification as a preprocessing mechanism has been discussed in
Chapter 6. Results on several synthetic and real datasets is presented. Chapter 7
summarizes the main contributions of the thesis and suggests avenues for future
work.