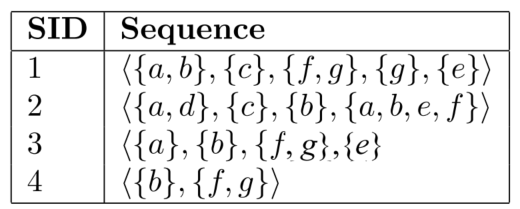
**What is sequential pattern mining?**

[Data mining](http://data-mining.philippe-fournier-viger.com/introduction-data-mining/) consists of extracting information from data stored in databases to understand the data and/or take decisions. Some of the most fundamental data mining tasks are clustering, classification, outlier analysis, and pattern mining. **Pattern mining**consists of discovering interesting, useful, and unexpected patterns in databases  Various types of patterns can be discovered in databases such as [frequent itemsets](http://data-mining.philippe-fournier-viger.com/introduction-frequent-pattern-mining/), associations, [subgraphs](http://data-mining.philippe-fournier-viger.com/introduction-frequent-subgraph-mining/), [sequential rules](http://data-mining.philippe-fournier-viger.com/introduction-to-sequential-rule-mining/), and [periodic patterns](http://data-mining.philippe-fournier-viger.com/an-introduction-to-the-discovery-of-periodic-patterns-in-data/).

The task of **sequential pattern mining** is a data mining task specialized for analyzing **sequential data,**to discover**sequential patterns**. More precisely, it consists of discovering interesting subsequences in **a set of sequences**, where the interestingness of a subsequence can be measured in terms of various criteria such as its occurrence frequency, length, and profit. Sequential pattern mining has numerous real-life applications due to the fact that data is naturally encoded as **sequences of symbols** in many fields such as bioinformatics, e-learning, market basket analysis, texts, and  webpage click-stream analysis.

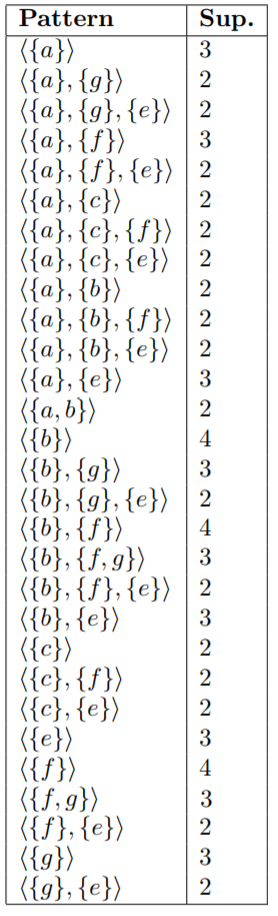
I will now explain the task of**sequential pattern mining** with an example. Consider the following**sequence database**, representing the purchases made by customers in a retail store.

[](http://data-mining.philippe-fournier-viger.com/introduction-sequential-pattern-mining/seq2-2/)

This database contains four sequences.  Each **sequence** represents the items purchased by a customer at different times. A sequence is an ordered list of itemsets (sets of items bought together). For example, in this database, the first sequence (SID 1) indicates that a customer bought some items a and b together, then purchased an item c, then purchased items f and g together, then purchased an item g, and then finally purchased an item e.

**Traditionally, sequential pattern mining**is being used to find subsequences that appear often in a sequence database, i.e. that are common to several sequences. Those subsequences are called the **frequent** **sequential patterns**. For example, in the context of our example, sequential pattern mining can be used to find the sequences of items frequently bought by customers. This can be useful to understand the behavior of customers to take marketing decisions.

To do **sequential pattern mining**, a user must provide a sequence database and specify a parameter called the **minimum support threshold**. This parameter indicates a minimum number of sequences in which a pattern must appear to be considered frequent, and be shown to the user. For example, if a user sets the minimum support threshold to 2 sequences, the task of**sequential pattern mining** consists of finding all subsequences appearing in at least 2 sequences of the input database.  In the example database, 29  subsequences met this requirement. These sequential patterns are shown in the table below, where the number of sequences containing each pattern (called the support) is indicated in the right column of the table.

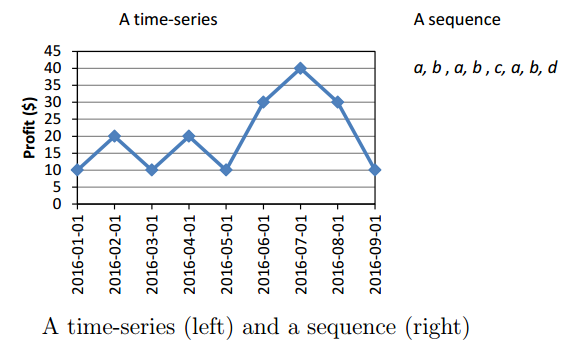
[](http://data-mining.philippe-fournier-viger.com/introduction-sequential-pattern-mining/results/)

For example, the patterns <{a}> and <{a}, {g}> are frequent and have a support of 3 and 2 sequences, respectively. In other words, these patterns appears in 3 and 2 sequences of the input database, respectively.  The pattern <{a}> appears in the sequences 1, 2 and 3, while the pattern <{a}, {g}> appears in sequences 1 and 3.   These patterns are interesting as they represent some behavior common to several customers. Of course, this is a toy example. Sequential pattern mining can actually be applied on database containing hundreds of thousands of sequences.

Another example of application of sequential pattern mining is text analysis. In this context, a set of sentences from a text can be viewed as sequence database, and the goal of sequential pattern mining is then to find subsequences of words frequently used in the text. If such sequences are contiguous, they are called “ngrams” in this context. If you want to know more about this application.

**Can sequential pattern mining be applied to time series?**

Besides sequences, **sequential pattern mining** can also be applied to**time series** (e.g. stock data), when discretization is performed as a pre-processing step.  For example, the figure below shows a **time series**  (an ordered list of numbers) on the left. On the right, a **sequence** (a sequence of symbols) is shown representing the same data, after applying a transformation.   Various transformations can be done to transform a time series to a sequence such as the popular SAX transformation. After performing the transformation, any sequential pattern mining algorithm can be applied.

[](http://data-mining.philippe-fournier-viger.com/introduction-sequential-pattern-mining/seq/)

**Where can I get Sequential pattern mining implementations?**

To try sequential pattern mining with your datasets, you may try the open-source**SPMF data mining software**, which provides implementations of numerous**sequential pattern mining algorithms:**<http://www.philippe-fournier-viger.com/spmf/>

It provides implementations of several algorithms for sequential pattern mining, as well as several variations of the problem such as discovering **maximal sequential patterns**, **closed sequential patterns** and sequential rules.[Sequential rules](http://data-mining.philippe-fournier-viger.com/introduction-to-sequential-rule-mining/)are especially useful for the purpose of performing predictions, as they also include the concept of confidence.

**What are the current best algorithms for sequential pattern mining?**

There exists several sequential pattern mining algorithms. Some of the classic algorithms for this problem are **PrefixSpan, Spade, SPAM,** and **GSP**. However, in the recent decade, several novel  and more efficient algorithms have been proposed such as **CM-SPADE**  and **CM-SPAM** (2014), **FCloSM**and**FGenSM** (2017), to name a few.  Besides, numerous algorithms have been proposed for extensions of the problem of sequential pattern mining such as finding the sequential patterns that generate the most profit (high utility sequential pattern mining).

# Data Mining - Applications & Trends

Data mining is widely used in diverse areas. There are a number of commercial data mining system available today and yet there are many challenges in this field. In this tutorial, we will discuss the applications and the trend of data mining.

Data Mining Applications

Here is the list of areas where data mining is widely used −

* Financial Data Analysis
* Retail Industry
* Telecommunication Industry
* Biological Data Analysis
* Other Scientific Applications
* Intrusion Detection

Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows −

* Design and construction of data warehouses for multidimensional data analysis and data mining.
* Loan payment prediction and customer credit policy analysis.
* Classification and clustering of customers for targeted marketing.
* Detection of money laundering and other financial crimes.

Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web.

Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in the retail industry −

* Design and Construction of data warehouses based on the benefits of data mining.
* Multidimensional analysis of sales, customers, products, time and region.
* Analysis of effectiveness of sales campaigns.
* Customer Retention.
* Product recommendation and cross-referencing of items.

Telecommunication Industry

Today the telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business.

Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list of examples for which data mining improves telecommunication services −

* Multidimensional Analysis of Telecommunication data.
* Fraudulent pattern analysis.
* Identification of unusual patterns.
* Multidimensional association and sequential patterns analysis.
* Mobile Telecommunication services.
* Use of visualization tools in telecommunication data analysis.

Biological Data Analysis

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis −

* Semantic integration of heterogeneous, distributed genomic and proteomic databases.
* Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
* Discovery of structural patterns and analysis of genetic networks and protein pathways.
* Association and path analysis.
* Visualization tools in genetic data analysis.

Other Scientific Applications

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy, etc. A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modeling, chemical engineering, fluid dynamics, etc. Following are the applications of data mining in the field of Scientific Applications −

* Data Warehouses and data preprocessing.
* Graph-based mining.
* Visualization and domain specific knowledge.

Intrusion Detection

Intrusion refers to any kind of action that threatens integrity, confidentiality, or the availability of network resources. In this world of connectivity, security has become the major issue. With increased usage of internet and availability of the tools and tricks for intruding and attacking network prompted intrusion detection to become a critical component of network administration. Here is the list of areas in which data mining technology may be applied for intrusion detection −

* Development of data mining algorithm for intrusion detection.
* Association and correlation analysis, aggregation to help select and build discriminating attributes.
* Analysis of Stream data.
* Distributed data mining.
* Visualization and query tools.

Data Mining System Products

There are many data mining system products and domain specific data mining applications. The new data mining systems and applications are being added to the previous systems. Also, efforts are being made to standardize data mining languages.

Choosing a Data Mining System

The selection of a data mining system depends on the following features −

* **Data Types** − The data mining system may handle formatted text, record-based data, and relational data. The data could also be in ASCII text, relational database data or data warehouse data. Therefore, we should check what exact format the data mining system can handle.
* **System Issues** − We must consider the compatibility of a data mining system with different operating systems. One data mining system may run on only one operating system or on several. There are also data mining systems that provide web-based user interfaces and allow XML data as input.
* **Data Sources** − Data sources refer to the data formats in which data mining system will operate. Some data mining system may work only on ASCII text files while others on multiple relational sources. Data mining system should also support ODBC connections or OLE DB for ODBC connections.
* **Data Mining functions and methodologies** − There are some data mining systems that provide only one data mining function such as classification while some provides multiple data mining functions such as concept description, discovery-driven OLAP analysis, association mining, linkage analysis, statistical analysis, classification, prediction, clustering, outlier analysis, similarity search, etc.
* **Coupling data mining with databases or data warehouse systems** − Data mining systems need to be coupled with a database or a data warehouse system. The coupled components are integrated into a uniform information processing environment. Here are the types of coupling listed below −
  + No coupling
  + Loose Coupling
  + Semi tight Coupling
  + Tight Coupling
* **Scalability** − There are two scalability issues in data mining −
  + **Row (Database size) Scalability** − A data mining system is considered as row scalable when the number or rows are enlarged 10 times. It takes no more than 10 times to execute a query.
  + **Column (Dimension) Salability** − A data mining system is considered as column scalable if the mining query execution time increases linearly with the number of columns.
* **Visualization Tools** − Visualization in data mining can be categorized as follows −
  + Data Visualization
  + Mining Results Visualization
  + Mining process visualization
  + Visual data mining
* **Data Mining query language and graphical user interface** − An easy-to-use graphical user interface is important to promote user-guided, interactive data mining. Unlike relational database systems, data mining systems do not share underlying data mining query language.

Trends in Data Mining

Data mining concepts are still evolving and here are the latest trends that we get to see in this field −

* Application Exploration.
* Scalable and interactive data mining methods.
* Integration of data mining with database systems, data warehouse systems and web database systems.
* SStandardization of data mining query language.
* Visual data mining.
* New methods for mining complex types of data.
* Biological data mining.
* Data mining and software engineering.
* Web mining.
* Distributed data mining.
* Real time data mining.
* Multi database data mining.
* Privacy protection and information security in data mining.