1. The World Wide Web is officially defined as a “wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.”
2. In simpler terms, the Web is an Internet-based computer network that allows users of one computer to access information stored on another through the world-wide network called the Internet.
3. In this model, a user relies on a program (called the client) to connect to a remote machine (called the server) where the data is stored. Navigating through the Web is done by means of a client program called the browser
4. To view these documents, one simply follows the links (called hyperlinks).
5. The idea of hypertext was invented by Ted Nelson in 1965 [14], who also created the well known hypertext system Xanadu (http://xanadu.com/). Hypertext that also allows other media (e.g., image, audio and video files) is called hypermedia.
6. **Creation of the Web:** The Web was invented in 1989 by Tim BernersLee, who, at that time, worked at CERN (Centre European pour la Recherche Nucleaire, or European Laboratory for Particle Physics) in Switzerland.
7. The proposal called for a simple protocol that could request information stored in remote computer systems through networks, and for a scheme by which information could be exchanged in a common format and documents of individuals could be linked by hyperlinks to other documents.
8. HyperText Transfer Protocol (HTTP)
9. the HyperText Markup Language (HTML) used for authoring Web documents
10. the Universal Resource Locator (URL)
11. In 1973, Vinton Cerf and Bob Kahn started to develop the protocol later to be called TCP/IP (Transmission Control Protocol/Internet Protocol).
12. The next significant event in the development of the Web was the arrival of Mosaic.
13. In mid-1994, Silicon Graphics founder Jim Clark collaborated with Marc Andreessen, and they founded the company Mosaic Communications (later renamed as Netscape Communications)
14. The Internet Explorer from Microsoft entered the market in August, 1995 and began to challenge Netscape
15. The Internet started with the computer network ARPANET in the Cold War era
16. Advanced Research Projects Agency (ARPA)
17. In 1973, Vinton Cerf and Bob Kahn started to develop the protocol later to be called TCP/IP (Transmission Control Protocol/Internet Protocol).
18. Google was launched in 1998 by Sergey Brin and Larry Page based on their research project at Stanford University.
19. launched the MSN search engine in spring 2005 (which is now called Bing).
20. W3C (The World Wide Web Consortium)

Web data mining
Characteristics of data web mining
Web is huge and still growing
Data of all types exist on the Web
Web is heterogeneous.
Web is linked

Web is noisy
The Web is also about businesses and commerce
The Web is dynamic
The Web is a virtual society

Data mining is also called knowledge discovery in databases (KDD).
It is commonly defined as the process of discovering useful patterns or knowledge from data sources.
Some of the common ones are supervised learning (or classification), unsupervised learning (or clustering), association rule mining, and sequential pattern mining.
A data mining application usually starts with an understanding of the application domain by data analysts (data miners).
data mining can be performed, which is usually carried out in three main steps:

- Pre-processing
- Data mining
- Post-processing

The whole process (also called the data mining process) is almost always iterative.
Web mining and text mining are becoming increasingly important and popular.

Web mining tasks can be categorized into three types:
Web structure mining,
Web content mining
Web usage mining.

The Web mining process is similar to the data mining process.

Association rules are an important class of regularities in data.
Its objective is to find all co-occurrence relationships, called associations.
The classic application of association rule mining is the market basket data analysis.
Such patterns are useful in Web usage mining for analyzing clickstreams in server logs.
They are also useful for finding language or linguistic patterns from natural language texts.

The problem of mining association rules can be stated as follows: Let I = {i1, i2, …, im} be a set of items.
Let T = (t1, t2, …, tn) be a set of transactions (the database).
An association rule is an implication of the form, $X \to Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \varnothing$. X (or Y) is a set of items, called an itemset.

that have support and confidence greater than or equal to the user-specified minimum support (denoted by minsup) and minimum confidence (denoted by minconf).

Confidence thus determines the predictability of the rule.

The Apriori algorithm works in two steps:

Generate all frequent itemsets

Generate all confident association rules from the frequent itemsets

The Apriori algorithm relies on the apriori or downward closure property to efficiently generate all frequent itemsets.

lexicographic order (a total order).

Candidate-gen function: The candidate generation function is given in consists of two steps,

the join step

the pruning step

minimum item support (MIS).

This dilemma is called the rare item problem.

Mining with Multiple Minimum Supports in two ways:

Multiple minimum class supports

Multiple minimum item supports

Basic Concepts of Sequential Patterns

A sequence is an ordered list of itemsets.

Recall an itemset X is a non-empty set of items $X \subseteq I$. We denote a sequence s by ⟨a1a2…ar⟩, where ai is an itemset, which is also called an element of s.

We assume without loss of generality that items in an element of a sequence are in lexicographic order.

The size of a sequence is the number of elements (or itemsets) in the sequence.

The length of a sequence is the number of items in the sequence.

A sequence of length k is called a k-sequence.

A sequence s1 = ⟨a1a2…ar⟩ is a subsequence of another sequence s2 = ⟨b1b2…bv⟩, or s2 is a supersequence.

The new algorithm, called MS-GSP.

The PrefixSpan algorithm can be adapted to mine with multiple minimum supports.

Generating Rules from Sequential Patterns

This section intro duces only three types:

sequential rules,
label sequential rules
class sequential rules

A sequential rule (SR) is an implication of the form, X → Y, where Y is a sequence and X is a proper subsequence of Y, i.e., X is a subsequence of Y and the length Y is greater than the length of X.
The support of a sequential rule, X → Y, in a sequence database S is the fraction of sequences in S that contain Y.
The confidence of a sequential rule, X → Y, in S is the proportion of sequences in S that contain X also contain Y.

Sequential rules may not be restrictive enough in some applications. We introduce a special kind of sequential rules called label sequential rules.
label sequential rule (LSR).
A wildcard is denoted by an "*" which matches any item.
These replaced items are usually very important and are called labels.

Class sequential rules (CSR) are analogous to class association rules (CAR).
0