UNIT-IV

Chapter 9
Web Mining

Introduction

The World Wide Web has become a very popular medium of publishing. Gathering the information and making sense of the data is difficult because the publication on the web is largely unorganized. In this chapter, the essential features of web mining are identified. This chapter also discusses the sub problems, where these standard techniques can be employed. It will be shown here that text mining research is the most important aspect of web mining.

Web Mining

With the huge amount of information available online, the World Wide Web is a rich area for data mining research. Web mining is the use of data mining techniques to automatically discover and extract information from web documents and services. Web is used for several purposes like

- Finding relevant information
- Discovering new knowledge from the web
- Personalized web page synthesis
- Learning about individual users

Finding relevant information: When we want to retrieve specific information from the web, we either browse or use the search engine. We pose a simple keyword query and the response from the web search engine is a list of pages, ranked based on the query. The related problems associated with the search tools are;

Low precision: This is due to irrelevance of many of the search results. i.e. we may get many web pages that are not relevant to our query

Low recall: This is due to the inability to index all the information available on the web. Some of the relevant pages are not properly indexed; we may not get the required pages through any of the search engine.

Discovering new knowledge on the web: This is a query-triggered process that is retrieval oriented. On the other hand we have data-triggered process that presumes that we have a collection of web data and we want to extract useful knowledge out of it.

Personalized web page synthesis: We may wish to synthesize a web page for different individuals from the available set of web pages. The information providers may wish to create a system by aggregating information from various sources. It is about what the customers do and what they want. This problem consists of sub problems such as mass customizing information according to the consumers need, problems related to marketing, effective web page design, management etc. Thus web mining techniques provide a set of techniques that can be used to solve the above said problems. They provide even direct solutions to the above problems. There are three operations related to web mining. They are

- Clustering
- Association
- Sequential analysis

Mining techniques in the web can be categorized into three areas of interest, based on which part of the web is to be mined. They are

- Web content mining

- Web structure mining
- Web usage mining

Web content Mining

Web content mining describes the discovery of useful information from the web contents.
The web encompasses a wide range of data like information about government, digital libraries, electronic business and services and other web applications.
The web content also consists of several types of data like text, image, audio, video, metadata and hyperlinks. Mining of multi-types of data is termed as multimedia data mining.

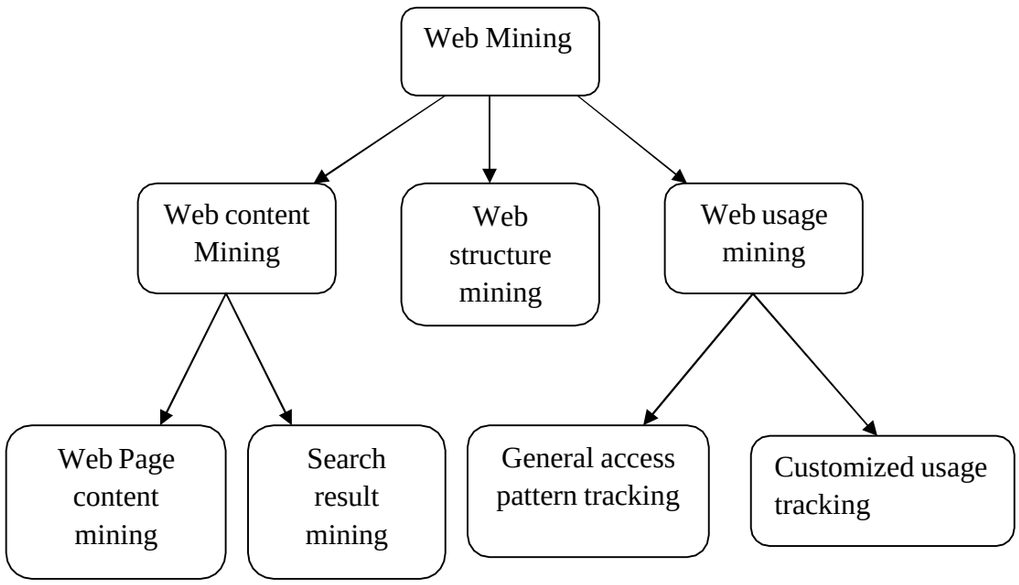

Figure 9.1: Web Mining Tasks

Web Structure Mining

Web structure mining is concerned with discovering the model underlying the link structures of the web. It is used to study the topology of the hyperlinks with or without the description of the links. This model can be used to categorize web pages and is useful to determine the similarity and relationship between different web sites. Web content mining attempts to explore the structure within a document and also within the web itself. We can consider any collection of hyperlinked pages V as a directed graph G=(V,E): the nodes correspond to the pages and a directed edge (p.q)$\in$E indicates the presence of a link from p to q. The out degree of a node p is the number of nodes to which it has links. The in degree of a node p is the number of nodes that have links to it. If W$\subseteq$V is a subset of the pages, we use G[W] to denote the graph induced on W- its pages are in W and its edges correspond to all the links between the pages in W. Some algorithms like HITS, PageRank, CLEVER are designed to calculate the relevancy of each web page.

Page Rank: Page Rank is defined as follows:
Let us assume that a page A has pages T_1,T_2 …T_n which point to it i.e. they are citations.

The parameter d is called the damping factor which can be set between 0 and 1 and is usually set to 0.85. Out_deg(A) denotes the number of links going out of page A. The page rank of page A is given as follows

$PR(A)=(1-d)+d(\Sigma_{i=1}^{n}(PR(T_i)/out\_deg(T_i)))$

Page rank can be calculated using a simple iterative algorithm. The probability that a surfer visits a page is called its page rank. The damping factor d is to model the probability that at each page the surfer would get bored and request another random page.

Social Network: Social network analysis is another method of studying the web link structures. It studies ways to measure the relative standing or importance of individuals in the network. Web structure mining utilizes the hyperlinks structure of the web to apply to social network analysis. The basic premise here is that if a web page points a link to another page then the former is endorsing the importance of the latter in some sense or other. Standing of a node: for nodes p and q let $P_{pq}^{(r)}$ denote the number of paths of length exactly r from p to q. Let $b<1$ be a constant, chosen to be small enough so that $Q_{pq}=\Sigma_r b^r P_{pq}^{(r)}$converges. One can view as the damping factor. The damping factor varies with the length of the path. The standing of a node q, σ_q is defined as

$\Sigma_p Q_{pq}$

Definitions:

Transverse Links: A link is said to be transverse link if it is between pages with different domain names.

Intrinsic Links: A link is said to be an intrinsic ink if it is between pages with the same domain name.

Index node: An index node is a node whose out-degree is significantly larger than the average out degree of the graph.

Reference node: A reference node is a node whose in-degree is significantly larger than the average in-degree of the graph.

In order to define similar pages, we need to define similarity measure between pages. The two basic similarity functions are

Bibliographic coupling: For a pair of nodes p and q bibliographic coupling is equal to the number of nodes that have links from both p and q.

Co-Citation: For a pair of nodes p and q, the co-citation is the number of nodes that point to both p and q.

Web usage mining:

Web usage mining deals with studying the data generated by the web surfer's sessions or behaviors. Web content or web structure mining utilize with the real or primary data on the web whereas web usage mining mines the secondary data derived from the interactions of the users with the web. Secondary data includes the data from the web server access logs, browser logs, user profiles, bookmark data , mouse clicks, scrolls which are results of user interactions. This data can be accumulated by the web server. There are two main approaches in web usage mining.

General Access Pattern Tracking: This is to learn user navigation patterns (impersonalized). It analyzes the web logs to understand the access patterns and trends.

Customized Usage Tracking: This is to learn user profile or user modeling in adaptive interfaces (personalized). It analyzes individual trends. The information displayed the depth of the site structure and the format of the resources can all be dynamically customized for each user over a

period of time, based on their access patterns. The mining techniques for web usage mining can be classified into two commonly used approaches.
The first approach maps the usage data of the web server into relational tables before a traditional data mining technique is performed.
The second approach uses the log data directly by utilizing the preprocessing techniques. The web usage data can also be represented with graphs.

Text Mining

Text mining corresponds to the extension of data mining approach to textual data and is concerned with various tasks, such as extraction of data implicitly contained in the collection of documents, or similarity based structuring. Although textual data expresses vast information it cannot be viewed as a traditional database. Hence specific techniques called text mining techniques have to be developed to process the unstructured textual data to aid in knowledge discovery. One way is to impose a structure on the textual data base and use any of the known data mining techniques. The other approach would be to develop a very specific technique for mining that exploits the inherent characteristics of textual data base. There is a relationship between areas like information retrieval (IR), information exchange (IE) and computational linguistics with text data mining.

Information Retrieval: IR is concerned with finding and ranking the documents that match the user's information needs. The way of dealing with textual information by the IR community is a keyword based document representation. The body of the text is analyzed by it constituent words and various techniques are used to build the core words of the document. The goals are
To find documents that are similar
To find the right index terms in a collection, so that querying will return the appropriate document.
IR is thus the automatic retrieval of relevant documents while at the same time retrieving as few of the non relevant ones as possible.

Information Extraction: IE has the goal of transforming a collection of documents, usually with the help of an IR system, into information that is more readily digested and analyzed. IE extracts the relevant facts from the documents, while IR selects relevant documents. Most IE systems use machine learning or data mining techniques to learn the extraction patterns or rules for documents either automatically or semi- automatically. The results of the IE process could be in the form of structured database or could be a summary of the original text or documents. IE is a kind of preprocessing stage in the text mining process, which comes after IR and before the data mining techniques are performed.

Computational Linguistics: Computational linguistics computes statistics over large text collections in order to discover useful patterns. These patterns are used to inform algorithms for various sub problems within natural language processing, such as parts of speech, tagging, word sensing etc.

Unstructured Text

Unstructured documents are free texts such as news stories. For an unstructured document, features are extracted to convert it to a structured form. Some of the important features are

Word occurrences: A bag of words or vector representation takes single words found in the training corpus as features ignoring the sequence in which they occur.

This representation is based on the statistic about single words in isolation. Such a feature is said to be Boolean if we consider whether the word either occurs or doesn't occur in the document.
Stop-words: The feature selection includes removing the case, punctuation, infrequent and stop-words.
Latent Semantic Indexing: LSI transforms the original document vectors to a lower dimensional space by analyzing the co relational structure of the document collection such that similar documents that do not share terms are placed in the same topic.
Stemming: It *is* a Stemming process which reduces words to their morphological roots.
n-Gram: This is using information about word positions in the document.
Part-of-Speech: We can assign number to each of the part of the part of speech.
Positional Collocations: The value of this feature is that, the word that occur one or two position to the right or left of the given word.
Higher order Features: Other features include phrases, terms, hypernyms, dates, email addresses or URLs. These features could still be reduced by further applying some selection techniques.
Once these features are extracted the text becomes a structured text and traditional data mining techniques can be used.

Episode Rule Discovery for Texts

Here the text is considered as sequential data which consists of a sequence of pairs (feature vector, index), where the feature vector is an ordered set of features and the index contains the information about the position of the word in the sequence.
Define a text episode as a pair $\alpha = (V,\leq)$ where V is a collection of feature vectors and $\leq$ is a partial order on V. Given a text sequence S, a text episode $\alpha = (V,\leq)$ occurs within S if there is a way of satisfying the feature vectors in V, using the feature vectors in S so that the partial order $\leq$ is satisfied.
For example:
The text *Pathfinder photographs Mars* can be represented as ((Pathfinder_noun_singular,1),(photographs_verb_singular,2), (Mars_noun_singular,3))
The text *knowledge discovery in databases* can be represented as ((knowledge_ noun_singular,1),(discovery_noun_singular,2)(in_preposition_singular,3), (databases_noun_plural,4))

Instead of considering all occurrences of the episode a restriction is set that the episode must occur within the prespecified window size, w. Thus we examine the substrings S' of S such that the difference of the indices in S' is at most w. For w=2 the subsequence (knowledge_ noun_singular,discovery_ noun_singular) is an episode contained in the window, but the subsequence is (knowledge_ noun_singular, databases_noun_plural) is not contained within the window. The support of α in S is defined as the number of minimal occurrences of α in S.

Hierarchy of Categories

When a user enters a query into a search engine, the system often brings back many different pages. Hence it is necessary to organize the documents into meaningful groups.
One way to group them is to put together all the documents written by the same author or that which are written in the same year or published by the same publisher.They could be grouped

according to their subject too. But the problem of assigning documents with single categories within a hierarchy is that most documents discuss the same topic simultaneously. Hence it is always better to describe a document both in terms of categories as well as hierarchies. Concept hierarchy is a data structure used for this purpose. Concept hierarchy is a directed acyclic graph of concepts where each of the concepts is identified by a unique name. An arc from a to b denotes that is a more general concept than b. We can tag texts with concepts. Each text concept is tagged by a set of concepts that correspond to its contents.

Text Clustering: Text clustering is another important task of text mining. Once the features of unstructured text are identified or the structured data of text is available, any of the clustering techniques can be applied to text. One popular text mining algorithm is Ward's minimum variance method. It is an agglomerative hierarchical clustering technique and it tends to generate compact clusters. Here the Euclidean metric or hamming distance is taken as a measure of dissimilarities between the feature vectors. The clustering method starts with n clusters, one for each text. At any stage two clusters are merged to generate a new cluster. The clusters C_k and C_l are merged to get a new cluster C_{kl} based on the following criteria:

$$V_{kl} = MIN_{i,j} V_{ij}$$

$$V_{ij} = \frac{\| \bar{x}_i - \bar{x}_j \|^2}{\frac{1}{n_i} + \frac{1}{n_j}}$$

where xi is the mean value of the dissimilarity for the cluster Ci and ni is the number of elements in this cluster.

Scatter/Gather: It is a method of grouping documents using clustering. This uses text clustering to group documents according to their overall similarities in their content. It is named so because it allows the user to scatter the documents in the form of clusters and then gather a subset of these groups and re-scatter them to form new groups. Each cluster is represented by a list of words that attempt to give the user a gist of what the documents in the cluster are all about. The user can look at the titles of the documents in each group. If a cluster has too many documents regrouping of the documents into smaller subsets can also be done.

Assignment-9

Short Answer Questions

1. What is page rank? (M.U. April/May 2012, 2011, 2008, M.U. Oct./Nov. 2013)
2. What is stemming? (M.U. April/May 2012, 2010)
3. What is text mining? (M.U. April/May 2012)
4. What are transverse and intrinsic link? (M.U. April/May 2010, M.U. Oct./Nov. 2013)
5. What is co-citation and bibliographic coupling? (M.U. April/May 2008)
6. What is low precision and low recall?
7. List out the main four purposes of Web.
8. What is text clustering? (M.U. Oct./Nov. 2013)
9. What are the three categories of web mining techniques?
10. What are three operations related to web mining?
11. What is web content mining? (M.U. Oct./Nov. 2013)
12. Define web usage mining.
13. Define web structure mining.

14. Differentiate between index node and reference node.
15. Differentiate between General Access Pattern Tracking and Customized Usage Tracking.
16. What do you mean by computational linguistics?
17. List any four features of unstructured text.
18. Differentiate between n-gram and part of speech.
18. What is Scatter/Gather?

Long Answer Questions

1. What are the different features of unstructured text? Explain. (5 Marks- M.U. April/May 2012, 2011, 2010, 2009, M.U. Oct./Nov. 2013)
2. What is web usage mining? Explain the types of web usage mining? (5 Marks- M.U. April/May 2012, 2011)
3. Explain the purpose of web mining? (5 Marks- M.U. April/May 2012, 2011, 2009)
4. What is text clustering? Explain? (5 Marks- M.U. April/May 2012)
5. What is page rank? How is it computed? Explain. (5 Marks- M.U. April/May 2010)
6. How is web usage mining different from web structure mining? Explain. (5 Marks- M.U. April/May 2010)
7. Explain the episode rule discovery for texts. (5 Marks- M.U. April/May 2010)
8. Explain the different types of web mining. (6 Marks- M.U. April/May 2009)
9. What is web content mining? (5 Marks- M.U. April/May 2008)
10. Write a note on web usage mining.(5 Marks, M.U. Oct./Nov. 2013)
11. Write a note on web structure mining. (5 Marks, M.U. Oct./Nov. 2013)
12. Write a note on social network. (5 Marks)
13. Write a note on web text mining. (5 Marks, M.U. Oct./Nov. 2013)
14. Write a note on hierarchy of categories (5 Marks)
15. Write a short note on Scatter/Gather. (4 Marks)

UNIT-IV

Chapter 10

Temporal and Spatial Data mining

Introduction

The widespread use of GIS by local and federal governments and other institutions, necessitate the development of adequate mining tools for geo-referenced data. Perhaps, the second generation of data mining techniques would contribute in a major way to spatial and temporal data mining. Temporal and spatial mining are important in other applications such as DNA sequencing. The motivation for the special-purpose algorithms for spatial data mining is almost the same as that of temporal mining. Some of the techniques of temporal mining can be trivially extended to spatial data. There are also certain similar mining tasks in these approaches. One of the major temporal data mining techniques is episode mining.

Temporal Data Mining

Temporal Data Mining is an important extension of data mining and it can be defined as the non-trivial extraction of implicit, potentially useful and previously unrecorded information with an implicit or explicit temporal content, from large quantities of data. It has the capability to infer causal and temporal proximity relationships, and this is something that non-temporal data mining cannot do.

Temporal rules cannot be mined from a database which is free of temporal components by traditional (non-temporal) data mining techniques. Thus, the underlying database must be a temporal one and specific temporal data mining techniques are also necessary. Consider, for example, associations rule which looks like-"Any person who buys a car also buys- steering lock". But if we take the temporal aspect into this rule would be-"Any person who buys a car also buys a steering lock after that". The concepts of during and after are explicitly temporal. Some other examples of temporal knowledge discovery are-"A drop in atmospheric pressure precedes rainfall in 60% of the cases"; "The sequence {semester examination, grading, result processing} occurs every semester"; Thus, temporal data mining aims at mining new and hitherto unknown knowledge, which takes into account the temporal aspects of the data.

Types of Temporal Data: There can be four different levels of temporality

Static: Static data are free of any temporal reference and the inferences that can be derived from this data are also free of any temporality.

Sequences (ordered sequences of events): If a transaction appears in the database before another transaction, it implies that the former transaction has occurred before the latter. There may not be any reference to quantitative temporal relationships. While most collections are often limited to the sequence relationships before and after, this category also includes the richer relationships, such as during, meet, overlap, etc. Such relationships are called qualitative relationships between time events. Sequence mining is one of the major activities in temporal data mining.

Timestamped: In this category the temporal information is explicit. Note that the relationship can be quantitative, in the sense that we can not only say that one transaction occurred before another, but also the exact temporal distance between the data elements.

Some examples include census data, land-use data and satellite meteorological data.

Fully Temporal: In this category, the validity of the data element is time-dependent. The inferences are necessarily temporal in such cases.

Temporal data Mining Tasks: Some of the conventional mining tasks can be extended with some additional temporal information as described below.

Temporal Association: The association rule discovery can be extended to temporal association. In static association rule discovery tasks, we were trying to find static associations between two non-temporal itemsets. In the temporal association discovery, we attempt to discover temporal association between non-temporal itemsets.

Temporal Classification: We can cluster the data items along temporal dimensions. For example, we can identify a set of people who go for a walk in the evening and a set of people who go for a walk in the morning. We can categorize sets of patients based on their visit sequence to different medical experts.

Temporal Characterization: An interesting experiment would be to extend the concept of decision tree construction on temporal attributes.

Trend Analysis: The analysis of one or more time series of continuous data may show similar trends i.e., similar shapes across the time axis. For example, "The deployment of the Data Mining system is increasingly becoming popular in the banking industry".

Sequence Analysis: Events occurring at different points in time may be related by causal relationships, in that an earlier event may appear to cause a later one. To discover such relationships, sequences of events must be analyzed to discover common patterns. This category includes the discovery of frequent events and also the problem of event prediction.

Temporal Association Rules

Association rules identify whether a particular subset of items are supported by an adequate number of transactions. The association may not indicate any causal relationship, unless the temporality in the association is brought out. One can extend the association rule discovery to incorporate temporal aspects too. It should be noted that the presence of a temporal association rule may suggest a number of interpretations, such as

- The earlier event plays some role in causing the later event.
- There is a third set of a reason that causes both events.
- The confluence of events is coincidental.

Temporal association rules are sometimes viewed in the literature as causal rules. Causal rules describe relationships, where changes in one event cause subsequent changes in other parts of the domain. The static properties, such as gender, and the temporal properties, such as medical treatments, are taken into account during mining. While the concept of association rule discovery is the same for temporal and non-temporal rules, algorithms designed for conventional rules cannot be directly applied to extract temporal rules.

Sequence Mining

An efficient approach to mining causal relations is sequence mining. As observed earlier, sequence mining is a topic in its own right and many application domains such as DNA sequence, signal processing, and speech analysis require mining of sequence data, even though there is no explicit temporality in the data. Discovering sequential patterns from a large database of sequences has been recognized as an important problem in the field of knowledge discovery and data mining. To put it briefly, given a set of data sequences, the problem is to discover subsequences that are frequent, in the sense that the percentage of data sequences containing them exceeds a user-specified minimum support.

Sequence Mining Problem: The most general form of the sequence mining problem [Zaki, 1998] can be stated as follows:

Let $\sum = \{i_1, i_2 \ldots , i_m\}$ be a set of m distinct items comprising the alphabet.

An event is a non-empty, disordered collection of items. Without any loss of generality, we write the items in an event in some predefined order. An event is denoted as $\{i_1, i_2 \ldots , i_m\}$, where i_j, is an item in $\sum$. Often, we drop the ',' and parentheses for notational convenience.

Any event that is given as input will also be called a transaction. Thus, transactions and events

have the same structure, except that a transaction is known to us prior to the process and an event is generated during the algorithm.

Definitions of some terms related to sequence mining are given below

Sequence: A sequence is an ordered list of events. A sequence, a, is denoted as $(\alpha_1 \rightarrow \alpha_2 \rightarrow \ldots \rightarrow \alpha_q \rightarrow)$, where α_i is an event. A sequence is called a k-sequence, if the sum of the cardinalities of α_i is k.

Subsequence: The sequence $s = (\alpha_1 \rightarrow \alpha_2 \rightarrow \ldots \alpha_q \rightarrow)$, is said to be a subsequence of $S^{|} = (\beta_1 \rightarrow \beta_2 \rightarrow \ldots\ldots \rightarrow \beta_r)$, if there exist indices $t_l < t_2 < \ldots < t_q$ of $s^{|}$, such that $\alpha_1 \subseteq \beta_{t1}$, $\alpha_1 \subseteq \beta_{t2}$, …….. $\alpha_q \subseteq \beta_{tq}$,

A sequence $S^{|}$ is said to support another sequence s if s is a subsequence of $S^{|}$.

Frequency: The frequency of a sequence s, with respect to this database D, is the total number of input sequences in D that support it.

Frequency Sequence: A frequent sequence is a sequence whose frequency exceeds some user-specified threshold. A frequent set is maximal if it is not a subsequence of another frequent sequence. The rationale behind frequent sequences lies in detecting precedence and causal relationships that make them statistically remarkable.

The GSP algorithm

The algorithms for solving sequence mining problems are mostly based on the A priori (level-wise) algorithm. One way to use the level-wise paradigm is to first discover all the frequent items in a level-wise fashion. It simply means counting the occurrences of all singleton elements in the database. Then, the transactions are filtered by removing the non-frequent items. At the end of this step, each transaction is a modified transaction consisting of only the frequent elements it contains.

We use this modified database as an input to the GSP algorithm. This process requires one pass over the whole database. GSP makes multiple passes over the database. In the first pass, all single items (1-sequences) are counted. From the frequent items, a set of candidate 2-sequences are formed, and another pass is made to gather their support. The frequent 2-sequences are used to generate the candidate 3-sequences, and this process is repeated until no more frequent sequences are found. There are two main steps in the algorithm.

Candidate Generation: Given the set of frequent (k- l)-frequent sequences $F_{(k-1)}$, the candidates for the next pass are generated by joining $F_{(k-l)}$ with itself. A pruning phase eliminates any sequence, at least one of whose subsequences is not frequent.

Support Counting: Normally, a hash tree-based search is employed for efficient support counting. Finally non-maximal frequent sequences are removed.

GSP Algorithm
F_1 = the set of frequent 1-sequence
k=2,
do while $F_{k-1} \neq \emptyset$;
generate candidate sets C (Set of candidate k-sequences);
for all input sequence s in the database D do
increment count of all a in Ct if s supports a
F_k = {a $\in$ C_k such that its frequency exceeds the threshold}
k=k+1
set of all frequent sequences is the union of all F_ks
end do.

The above algorithm looks like the a priori algorithm. One main difference is however the generation of candidate sets. Let us assume that A $\rightarrow$B and A$\rightarrow$ C are two frequent 2-sequences. The items involved in these sequences are (A, B) and (A, C), respectively. The candidate generation in the usual a priori style would give (A, B, C) as a 3-itemset, but in the present context we get the following 3-sequences as a result of joining the above 2-sequences
A$\rightarrow$B$\rightarrow$C, A$\rightarrow$C$\rightarrow$C, and AB$\rightarrow$C . The candidate-generation phase takes this into account.

Episode Discovery

Another important temporal data mining problem is the discovery of episodes that occur frequently within sequences. Episode discovery is similar to sequence mining, but for the following special assumptions:

- The input sequence is a single long input sequence, unlike in the case of sequence mining where we have a set of data sequences.
- The events (in this context, referring to a transaction as an event is more appropriate) are typically single item events.
- An episode is a subsequence.

The frequent episode discovery problem is to find all episodes that occur frequently in the event sequence within a time window. Let us define some basic concepts that are necessary in the present context.

Definitions of some terms related to episode discovery are given below

Event: An event is a pair {A, t}, where A is an single item event, and t is an integer timestamp of the occurrence of A.

Event Sequence: An event sequence Ev_Seq is a triplet (Seq, T_slart, T_end), with T_start and T_end denoting the start and end time of the sequence; and Seq = ({A_1, t_1}, {A_2, t_2}, ... {A_k, t_k}) is an ordered sequence of events.
Time window: A time window W for Ev_Seq(Seq, T_start, T_end}, is an event sequence (W, t_s, t_e), where $t_s \geq$T_start and $t_e \leq$ T_end and (t_e- t_s) is said to be width of the window.

We shall represent the data in the form of a graph, where each event corresponds to a node. The precedence relationships among nodes represent the temporal precedence among events.

Episode: An episode a is a triple (V, ≤, g), where V is a set of nodes, ≤ is a partial order on V, and g: V→I is a mapping associating each node with an event satisfying the partial order.

Parallel and Serial Episodes: If the partial order ≤, is a trivial partial order, the episode is called a parallel episode. If the partial order is a total ordering, then the episode is called serial episode. An episode A → B → C is a serial episode. A serial episode occurs in a given sequence only if A, B and C occur in this order relatively close. There can be other events occurring between these three. In a parallel episode, there is no constraint on the relative order of the events.

Subepisode: An episode is to be a sub episode if it is obtained by deleting some events from an episode.

Occurrence of Episode in an Event Sequence: An episode is said to be occurring in a sequence if the events corresponding to the nodes of the episode appear in the sequencing, preserving the partial order of the episode.

Frequency of an Episode: The frequency of an episode with respect to a given window width in a sequence, is the fraction of all the windows in the sequence of the specified width in which the episode occurs. Let us assume that W(ω) is the set of total number of time windows of width ω in the given sequence, and W(ω, α) is the set of windows in W(ω) in which the episode a occurs. Then, the frequency is the ratio of W(ω, α) to W (ω). A frequent episode is the episode that has a frequency above a user-specified threshold.

The episode discovery problem can be stated as follows. Given a sequence Ev _seq, a class C of episodes, a window width w, and a frequency threshold σ, it is to find all frequent episodes with respect to ω and σ in Ev_seq.

Episode discovery Process: Like many other mining problems, the discovery of a frequent episode is also influenced by the level-wise algorithm. The algorithm makes multiple passes and the candidate sets of (k+ 1)-episodes are generated from frequent k-episodes. This is done by finding the pair of frequent k-episodes having a common prefix of size k-l and preserving the partial ordering.

Event Prediction Problem

The discovery of frequent sequences is inappropriate for many applications where sequence pattern discovery is relevant. The event prediction problem is to predict a type of future event, the target event, based on past events. More specifically, the problem is to find a prediction rule that successfully predicts future target events by taking the input as time stamped records with categorical items.

This problem is particularly interesting in situations where the target event occurs infrequently in the event sequence. When we try to predict an event, we also keep in mind the target event's prediction period, which means it must occur at least warning time units before the target event and no more than monitoring time units before the target event. It is interesting to note that in event prediction problems a target event is said to be correctly predicted if at least one prediction is made within its prediction period, regardless of any subsequent negative predictions. Thus, the reliability of a positive prediction is not affected by the presence of negative predictions.

Every event is represented by a tuple consisting of a timestamp field and a number of features. We introduce a wildcard in the event as '?'. Each feature in an event is permitted to take on any of a predefined list of valid feature values, as well as the wildcard ("?") value, which matches any feature value.

Time Series Analysis

Time series are an important class of complex data objects; they arise in many applications. For example, stock price indices, volume of product sales, telecommunication data, one-dimensional medical signals, audio data, and environmental measurement sequences are all time-series databases. Time-series data are sequences of real numbers representing measurements at uniformly-spaced temporal instances. Time-series analysis, like all other forms of data analysis, is used to characterize or explain the reasons for the behaviour of a system and/or to predict future behaviour.

The most important aspects of time-series data are that these are sequence data but uniformly spaced on the temporal attribute. Analysis tasks of time series include feature extraction of time-series data; computation of similarity measure among time series data set; segmentation of data set; matching two time series data; clustering and classifying time-series data.

Definitions of some terms related to time series analysis are given below

n-series: An n-series X is a sequence $\{x_1, x_2, \ldots, x_n\}$ of real numbers. Each n-series X has an average $\alpha(X)$ and a deviation $\sigma(X)$:

$$\alpha(X) = - \sum \quad \text{and} \quad \sigma(X) = - \sum - \alpha(X))2$$

Similarity Function: While mining a time-series database we need an essential primitive operation, called the similarity function. It is often necessary to search within a series database for those series that are similar to a given query series. This primitive is needed, for example, for prediction and clustering purposes. For instance, we may be interested in finding the months in five years having similar sales patterns; or in classifying companies based on similar stock price

fluctuations. While the statistical literature on time-series is vast, it has not studied similarity notions that would be appropriate for data mining applications. Typically, the task is to define a function Sim(X; Y), where X and Yare two time-series, and the function value represents how "similar" they are to each other.

Similarity: We say that two time-series X and Yare similar, if there exist a > 0 and b, such that Y, = ax_i+b, for all i.

Distance Metric: Distance metric plays a major role in the measurement of similarity. We say that X and Y are approximately similar with respect to a distance metric d, if d(X, Y) ≤ ε for some tolerance ε. The distance metric can be defined such a way that the scale and shift invariances are taken care of within d.

Normal Series: An n-series X is normal if the average $\alpha(X) = 0$ and deviation σ (X)=1.

There exists a unique normal series for X^* and hence the normal form representation of all elements of X^*, is said to be the corresponding normal denoted by n(X). The normal form can play an important role in determining similarity. Two objects are similar with regard to scale and shift, if their corresponding normal forms are similar.

Discrete Fourier Transform (DFT): The Discrete Fourier Transform of a n-sequence X is defined to be a sequence of n complex numbers $DFT_m(X)$, m = 0, ... , (n-1), given by

$$DFT_m(X) = \frac{}{\sqrt{}} \sum \qquad \frac{\pi i\ m}{}$$

Where i is the imaginary unit, $\sqrt{(-1)}$. Note that $DFT_m(X)$ is a complex number.

Fingerprint: A fingerprint F(X) of an n-series X is the sequence of first few DFT coefficients of X. In other words,

$F(X) = (DFT_1(X), DFT_2(X) \qquad DFT_k(X))$

Where k assumes a very low value, between 3 and 5

Longest Common Sequence: A more reasonable similarity notion is based on the longest common subsequence concept, where intuitively X and Yare considered similar if they exhibit similar behaviour for a large part of their length.

Feature Extraction from Time Series: We can extract k features from every sequence and every sequence is then represented as a unique point in k-dimensional space. Then we can use a multidimensional index to store and search these points. Different multidimensional indexing methods that are currently popular are R*-tress, k-d Trees, Linear Quad-trees etc. The major problems with these features extraction techniques are:

- Completeness of feature extraction: Completeness ensures that similarity between two objects is preserved in the feature space. That is, if two objects are similar in some

distance metric, their corresponding maps in k-dimensional feature space should not be grossly dissimilar.

- Dimensionality Curse: Most multidimensional indexing scale up for high dimensionality. Such spatial indices do not work well for very high dimensional.

Other Methods of Analyzing Time Series: The original approach to time series analysis was the establishment of a mathematical model describing the observed system. While linear models were the dominant paradigm for several decades, non-linear models emerged later to deal with systems, showing properties for which linear models are less appropriate. Since the eighties, time-series analysis research has looked towards the exploitation of machine learning algorithms. Those algorithms analyze an unfamiliar time series by learning, that is, by emulating its structure.

Spatial Mining

Spatial data mining is the branch of data mining that deals with spatial (location, or geo-referenced) data. Consider a map of the city of Hyderabad containing various natural and man-made geographic features, and clusters of points (where each point marks the location of a particular house). The houses might be noteworthy because of their size, historical interest, or their current market value. Clustering algorithms exist to assign each point to exactly one cluster, with the number of clusters being defined by the user. We can mine varieties of information by identifying likely relationships. For example, "the land-value of the cluster of residential area to the east of 'Cyber- Tower' is high" or, "70% the Banjara migrants settle in the city around the market area". Such information could be of value to realtors, investors, or prospective home buyers, and also to other domains such as satellite images, photographs, oil and gas explorations. This problem is not trivial-there may be a large number of features to consider. We need to be able to detect relationships among large numbers of geo- referenced objects without incurring significant overheads.

As in the case of temporal data mining, conventional data mining techniques cannot fully exploit the spatial characteristics of data. It is necessary to devise algorithms that take this aspect into consideration

Assignment-10

Short Answer Questions (2 marks each)

1. What is temporal data mining? (M.U. Oct./Nov. 2013)
2. List out the different types of temporal data.
3. Differentiate between temporal Association and temporal classification.
4. What is trend analysis?
5. What is sequence analysis?
6. Define sequence and sub sequence.
7. Define frequency and frequency subsequence.
8. What is episode discovery?
9. Differentiate between event and event sequence.
10. Define time window and episode.
11. What is frequency of an episode?
12. What is a parallel and serial episode?

13. Define similarity and distance metric.
14. What is discrete Fourier transform?
15. What is fingerprint?
16. What is longest common sequence?
17. What is spatial mining?

Long Answer Questions

1. Write a short note on temporal data mining. (5 Marks)
2. What are the types of temporal data mining? Explain. (5 Marks, M.U. Oct./Nov. 2013)
3. Write a short note on temporal data mining tasks. (5 Marks)
4. Write a short note on temporal association rules. (5 Marks)
5. Explain sequence mining problem. (5 Marks)
6. Explain GSP algorithm? (5 Marks, M.U. Oct./Nov. 2013)
7. Explain episode discovery. (7 Marks)
8. Explain on event prediction problem? (7 Marks)
9. Write a short note on event prediction problem. (5 Marks)
10. Write a short note on time series analysis. (5 Marks)
11. Explain spatial mining. (5 Marks)
12. Define following terms (10 Marks)
    i. Discrete Fourier Transform
    ii. Fingerprint
    iii. Normal series
    iv. Distance Metric
    v. Similarity
13. Define following terms (10 Marks)
    i. Event
    ii. Event sequence
    iii. Time window
    iv. Episode
    v. Parallel and serial episodes