Module 5

Web Mining

Introduction

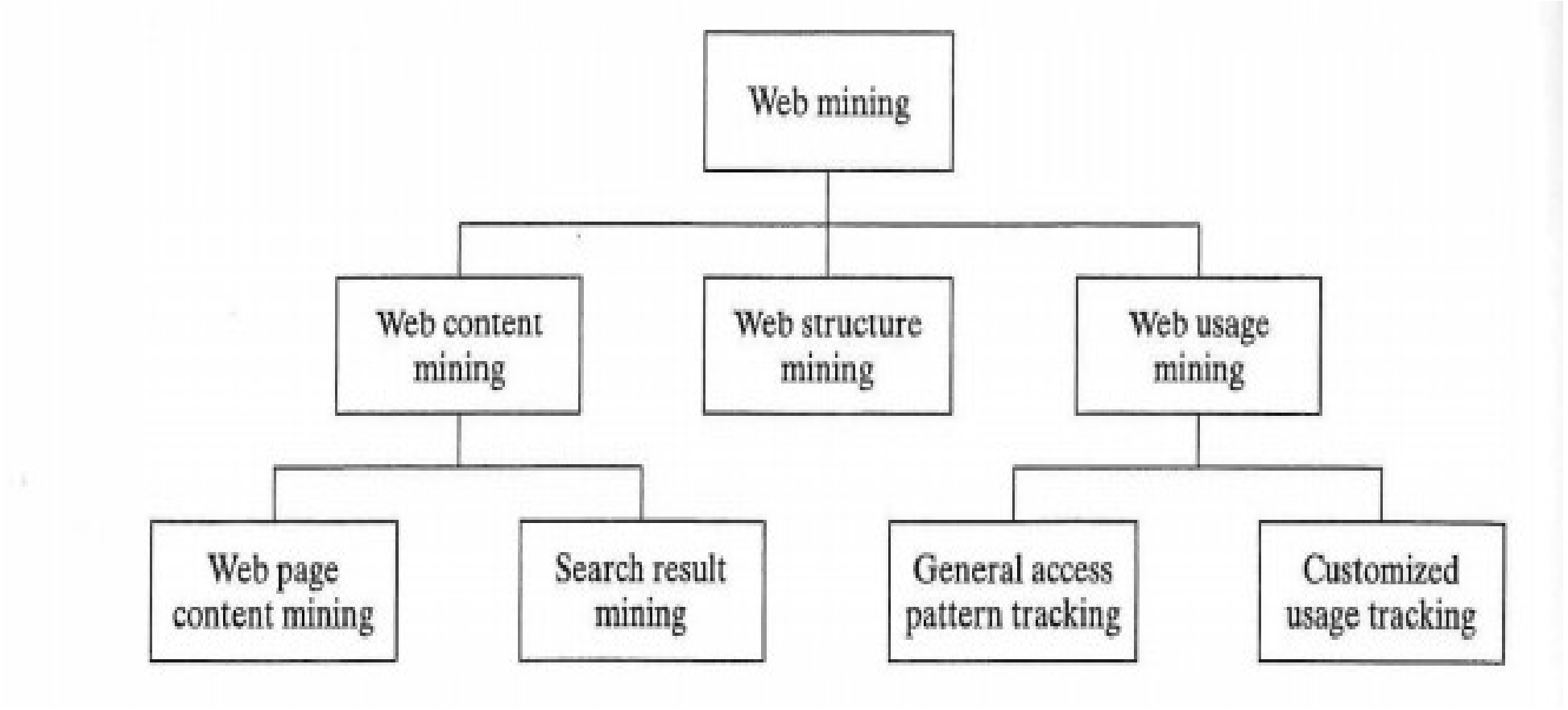
- Web mining is mining of data related to the World Wide Web.
- Web mining is one of the datamining technique used to discover patterns, structures, and knowledge from the web
- Organised into 3 areas:-

a. Web content mining

b. Web structure mining

c. Web usage mining

Web Mining Taxonomy



Web content mining

- Web content mining examines the content of Web pages as well as results of Web searching.
- The content includes text as well as graphics data.
- This type of mining performs scanning of the text, images and group of web pages
- Techniques:-

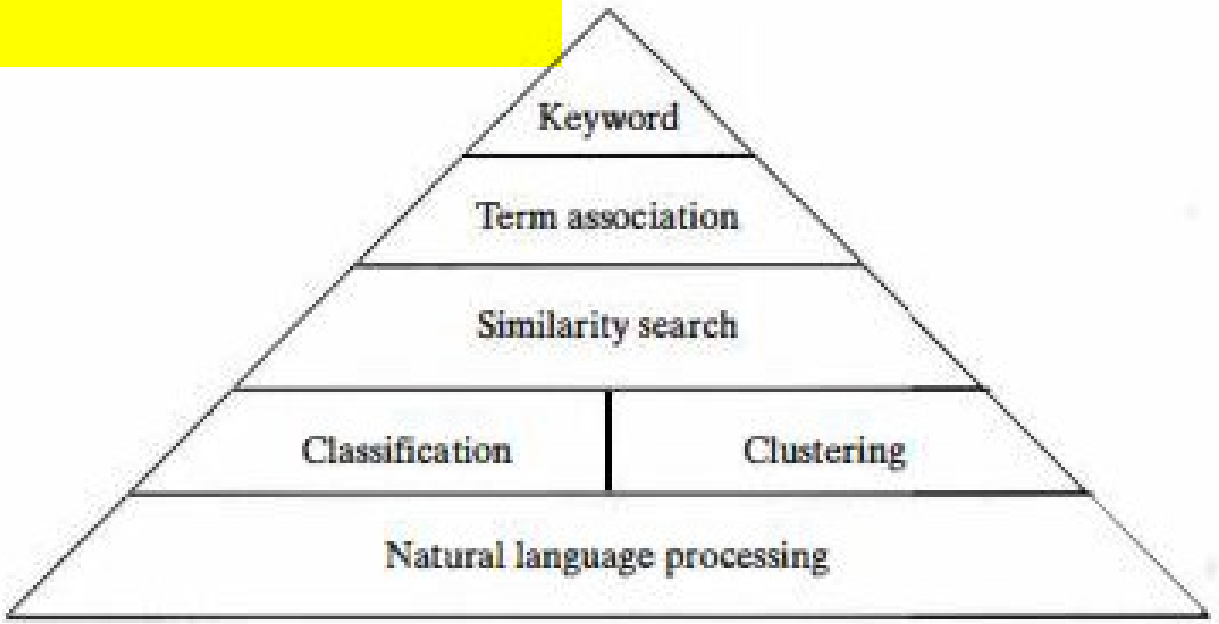
a) Natural Language Processing(NLP)

b) Information Retrieval(IR)

- Web content mining can be thought of as extending the work performed by basic search engines
  - Eg :If a user want to search for particular book then search engine provide list of suggestion
- Most search engines are keyword based
- Traditional search engines must have crawlers to search the Web and gather information
- Data mining techniques can be used to help search engines provide the efficiency, effectiveness, and scalability needed

- One taxonomy of Web mining divided Web content mining into agent based and database approaches
  - Agent based approaches have software systems
  - The database approaches view the Web data as belonging to a database, there have been approaches that view the Web as a multilevel database, and there have been many query languages that target the Web

- Basic content mining is a type of text mining
- Text mining functions can be viewed in a hierarchy with the simplest functions at the top and the more complex functions at the bottom
- Text r



Crawlers

- A robot (spider/crawler) is a program that traverses the hypertext structure in the Web.
- The page (or set of pages) that the crawler starts with are referred to as the seed URLs.
- By starting at one page, all links from it are recorded and saved in a queue.
- These new pages are in turn searched and their links are saved.
- A crawler may visit a certain number of pages and then stop, build an index, and replace the existing index.
- Crawlers are used to facilitate the creation of indices used by search engines.

- Traditional crawlers usually replace the entire index or a section thereof.
- An incremental crawler selectively searches the Web and only updates the index incrementally as opposed to replacing it
- Because of the tremendous size of the Web, it has also been proposed that a focused crawler be used.
- A focused crawler visits pages related to topics of interest.

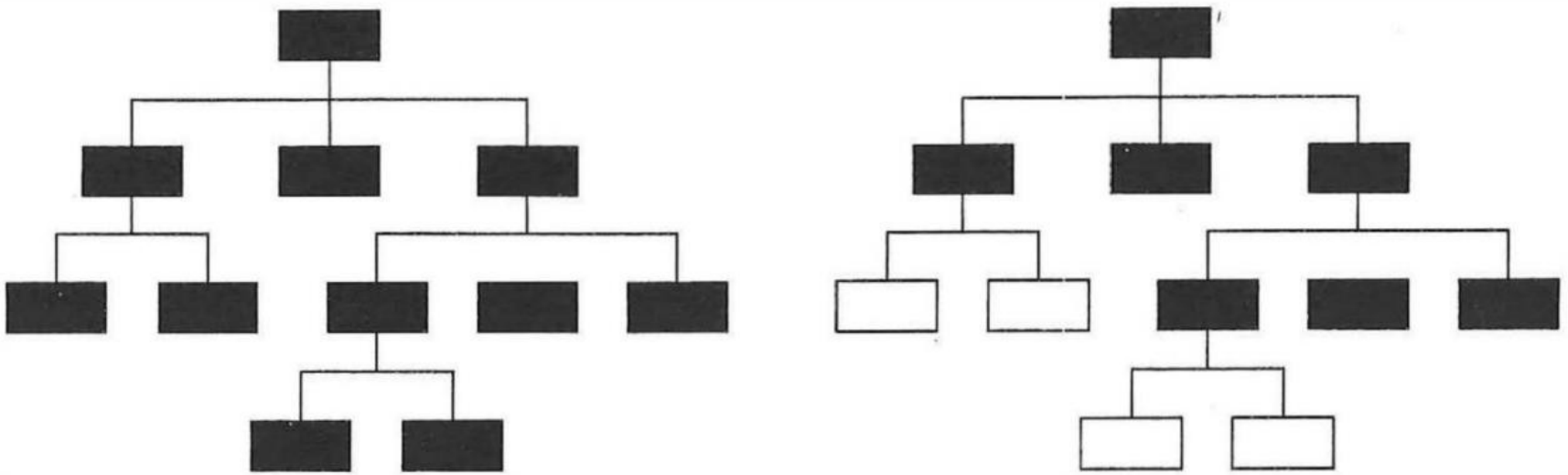


FIGURE 7.3: Focused crawling.

- The shaded boxes represent pages that are visited.
- With focused crawling, if it is determined that a page is not relevant or its links should not be followed, then the entire set of possible pages underneath it are pruned and not visited.
- With thousands of focused crawlers, more of the Web can be covered than with traditional crawlers.
- This facilitates better scalability as the Web grows.
- The focused crawler architecture consists of three primary components:

1. A hypertext classifier that associates a relevance score for each document with respect to the crawl topic.
2. A distiller determines which pages contain links to many relevant pages. These are called hub pages.
3. The crawler performs the actual crawling on the Web.

The pages it visits are determined via a priority-based structure governed by the priority associated with pages by the classifier and the distiller.

To use the focused crawler,

- User identifies the documents that are of interest.
- These are then classified based on a hierarchical classification tree, and nodes in the tree are marked as good, indicating that this node in the tree has associated with it document(s) that are of interest.
- These documents are then used as the seed documents to begin the focused crawling.
- Each document is classified into a leaf node of the taxonomy tree.

- One proposed approach, hard focus, follows links if there is an ancestor of this node that has been marked as good.
- Another technique, [illegible] identifies the probability that a pag[illegible]nt as

$$R(d) = \sum_{good(c)} P(c \mid d)$$

- Here c is a node in the tree (thus a page) and good(c) is the indication that it has been labeled to be of interest.
- The priority of visiting a page not yet visited is the maximum of the relevance of pages that have been visited and point to it.

- The hierarchical classification approach uses a hierarchical taxonomy and a naive
  Bayes classifier.
- A hierarchical classifier allows the classification to include information contained in the document as well as other documents near it (in the linkage structure).
- The objective is to classify a document d to the leaf node c in the hierarchy with
  the highest posterior probability P(c I d).
- Based on statistics of a training set, each node c in the taxonomy has a probability.
- The probability that a document can be generated by the root topic, node q,
  obviously is 1.
- Following the argument found in , suppose q, ... , Ck = c be the path from the root node to the leaf c.

- We thus know

$$P(c_i \mid d) = P(c_{i-1} \mid d) P(c_i \mid c_{i-1}, d)$$

- Using Bayes rule, we have

$$P(c_i \mid c_{i-1}, d) = \frac{P(c_i \mid c_{i-1}) P(d \mid c_i)}{\sum_{s \text{ is a sibling of } c_i} P(d \mid s)}$$

- P (d I c;) can be found using the Bernoulli model, in which a document is seen as a

  bag of words with no order.

- More recent work on focused crawling has proposed the use of context graphs.

The context focused crawler (CFC) performs crawling in two steps.

In the first phase, context graphs and classifiers are constructed using a set of seed documents as a training set.

In the second phase, crawling is performed using the classifiers to guide it.

In addition, the context graphs are updated as the crawl takes place.

This is a major difference from the focused crawler, where the classifier is static after the learning phase.

- The CFC approach is designed to overcome problems associated with previous crawlers:
  - There may be some pages that are not relevant but that have links to relevant pages. The links out of these documents should be followed.
  - Relevant pages may actually have links into an existing relevant page, but no links into them from relevant pages. However, crawling can really only follow the links out of a page. It would be nice to identify pages that point to the current page. A type of backward crawling to determine these pages would be beneficial.

- The CFC approach uses a context graph, which is a rooted graph in which the root represents a seed document and nodes at each level represent pages that have links to a node at the next h

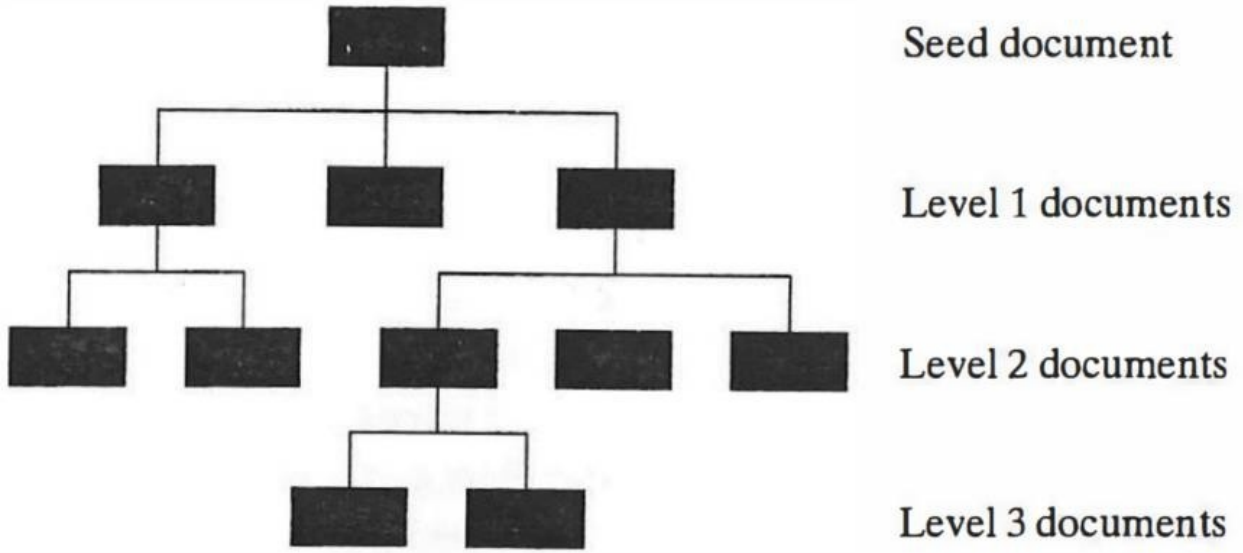


FIGURE 7.4: Context graph.

- The number of levels in a context graph is dictated by the user.

- A node in the graph with a path of length n to the seed document node represents a document that has links indirectly to the seed document through a path of length n.
- The number of back links followed is designated as input to the algorithm.
- Here n is called the depth of the context graph.
- The context graphs created for all seed documents are merged to create a merged context graph.
- The context graph is used to gather information about topics that are related to the topic being explored.
- Backward crawling finds pages that are not pointed to by relevant documents but are themselves relevant.
- These types of pages may be new and may not yet have been discovered and linked to from other pages.

- Although backward links do not really exist in the Web, a backward crawl can be performed relatively easily because most search engines already maintain information about the back links.
- This type of information is similar to that often used by commercial citation servers, which find documents that cite a given document.
- The value of the Science Citation Index in performing traditional literature
  searches is well known.
- The use of backlinks on the Web can provide similar benefits.
- CFC performs classification using a term frequency-inverse document frequency
- (TF-IDF) technique. The vocabulary used is formed from the documents in the seed set and is shown in the merged context graph.
- Each document is represented by a TF-IDF vector representation and is assigned to a particular level in the merged context graph.

Harvest System

- Harvest system is based on the use of caching, indexing and crawling.
- Harvest is actually a set of tools that gather information from diverse sources.
- It is centered around the use of gatherers and brokers.
- Gatherers obtains information for indexing from an Internet Service Provider.
- Brokers provide index and query interface.
- Harvest gatherers use the essence system to assist in collecting data.
- Essence classifies documents by creating a semanticindex.
- Semantic indexing generates different types of information for different types of files and then creates indices on this information.
- This process first classify files based on type and then summarize the files based on keyword.

Virtual Web View

- Handling the large amount of unstructured data on the Web is to create a Multiplelayered DataBase(MLDB) on the top of the data in the cells.
- Each layer of this data base is more generalized than the large beneath it.
- The upper levels are structured and can be accessed by an SQL like querylanguage.
- A view of the MLDB called Virtual WebView(VWV) can beconstructed.
- Extraction and translation tools are proposed for the creation of the first layer of theMLDB.
- Translationtools are used to convert Web Documents to XML.
- Extraction tools extract the desired information from the web page and insert it into first layerof MLDB.
- The layer one data can be viewed as a massive distribution database.
- The higher levels of the database becomes less distributed and more summarized as they move up hierarchy.

- Generalization tools are proposed and concept hierarchies are used in the generalization processfor constructing the higher levels of the MLDB.
- These hierarchies can be created using the Word Net Semantic Network. Word Net is a databaseof English language.
- Web Data Mining Query Language, Web ML perform data mining operation on the MLDB. It is an extension of DMQL.
- A major feature of Web ML are four primitive operation based on the use of concepthierarchies.
  - COVERS: One Concept Covers another if it is higher in the hierarchy.
  - COVERED BY: It is the reverse of COVERS.
  - LIKE: Concept is a synonym.
  - CLOSE TO: One Concept is close to another if it is sibling in the hierarchy.

  Eg: illustrates WebML. Query find all the document at the level of www.engo.smu.edu that have a keyword that covers the keyword cat.
  - SELECT: From document in www.engr.smu.eduWHERE one of keywords COVERS “cat”.
  - WHERE: indicates selection based on the links found in the page, keywords
    and information about the domains where the document is found.

- Personalization : Example of web content mining.
- Web access or web contents tuned to better fit the desires of each user.
- With personalization, advertisements to be sent to the customer based on the specific knowledge.
- Targeting, personalization may be performed on the target Web page.
- Goal – Make the customer to purchase something.
- With targeting, businesses display advertisements at other sites visited by their users.
- With personalization, when a particular person visits a Web site, the advertising can be designed specifically for that person.
- Personalization can be viewed as a type of clustering, classification, prediction.
    - >Classification - the desires of a user are determined based on those for the class.
    - >Clustering - the desires are determined based on those users to determined to be similar.
    - >Prediction - used to predict what the user really wants to see.

Three basic types of Web page: personalization

I. Manual techniques - Identifies user preferences based on profiles or demographics.
II. Collaborative filtering - Identifies preference based on ratings from similar users.
III. Content based filtering -Retrieves pages based on similarity between pages and user profile.

- Personalization using data mining techniques to determine user preferences.

Automated personalization technique

predicts future needs based on past needs or the needs of similar users.

i. News Dude

- uses the interestingness of a document to determine if a user is interested in it.
- interestingness is based on the similarity between the document and that of what the user wishes.
- Similarity is measured by the co-occurrence of words in the documents and a user profile created for the user.
- News Dude uses a two-level scheme to determine interestingness.
    i. One level is based on recent articles the user has read
    ii. Second level is a more long-term profile of general interests.

ii. Firefly

- Firefly is based on the concept that humans often base decisions on what they hear from others.
- viewed as a type of clustering.
- This approach to Web mining is referred to as collaborative filtering.

iii. Web Watcher

- Web Watcher prioritizes links found on a page based on a user profile and the results of other users with similar profiles who have visited this page.

Web structure mining

- Web structure mining is the process of discovering structure information from the web.
- It extract patterns from hyperlink and document structures.
- Webpages –as nodes
- Hyperlinks –as connection between two related pages. (nodes)
- Techniques

a) Page rank technique by Google

b) CLEVER technique by giving weights to the pages

1.PageRank

- PageRank is used to measure the importance of a page .
- The PageRank value for a page is calculated based on the number of pages that point to it.
- This is actually a measure based on the number of backlinks to a page. A backlink is a link pointing to a page rather than pointing out from a page.

- Given a page *p*, we use to be the set of pages that point to *p*, and to be the set of links out of *p*. The PageRank of a page *p* is defined as

$$\mathrm{PR}(p) = c \sum_{q \in B_p} \frac{\mathrm{PR}(q)}{N_q}$$

- Here =| |. The constant c is a value between 0 and 1 and is used for normalization.

- A problem, called rank sink, that exists with this PageRank calculation is that when a cyclic reference occurs (page A points to page B and page B points to page A), the PR value for these pages increases.
- This problem is solved by adding an additional term to the formula: -

$$PR'(p) = c \sum_{q \in B_p} \frac{PR(q)}{N_q} + cE(v)$$

- where c is maximized. Here E (v) is a vector that adds an artificial link.
- This simulates a random surfer who periodically decides to stop following links and jumps to a new page.

2. Clever

- Clever, is aimed at finding both authoritative pages and hubs
- The authors define an authority as the "best source" for the requested information
- In addition, a hub is a page that contains links to authoritative pages.
- The Clever system identifies authoritative pages and hub pages by creating weights.
- Hyperlink-induced topic search (HITS) finds hubs and authoritative pages

- The HITS technique contains two components:-

a) Based on a given set of keywords, a set of relevant pages is found.

b) Hub and authority measures are associated with these pages. Pages with the highest values are returned.

Input:

```
W          //WWW viewed as a directed graph
q          //Query
s          //Support
```

Output:

```
A          //Set of authority pages
H          //Set of hub pages
```

HITS algorithm

$R = SE(W, q)$

$B = R \cup \{\text{pages linked to from } R\} \cup \{\text{pages that link to pages in } R\}$;

$G(B, L) =$ Subgraph of W induced by B;

$G(B, L^1) =$ Delete links in G within same site;

$x_p = \sum_q$ where $(q,p) \in L^1$ y_q; // Find authority weights;

$y_p = \sum_q$ where $(p,q) \in L^1$ x_q; // Find hub weights;

$A = \{p \mid p$ has one of the highest $x_p\}$;

$H = \{p \mid p$ has one of the highest $y_p\}$;

Web usage mining

- Web usage mining looks at logs of Web access.
- It is classified as General access pattern tracking and Customized tracking
- General access pattern tracking is a type of usage mining that looks at a history of Web pages visited
- Customized tracking:- targeted to specific usage or users
- Techniques:-

a) Pattern discovery tools

b) Pattern analysis tools

- Web usage mining performs mining on Web usage data, or Web logs.
- A Web log is a listing of page reference data. Sometimes it is referred to as clickstream data because each entry corresponds to a mouse click.
- These logs can be examined from either a client perspective or a server perspective.
- Server perspective: mining uncovers information about the sites where the service resides. It can be used to Improve the design of the sites.
- Client perspective: information about a user (or group of users) is detected. This could be used to perform prefetching and caching of pages.

- The webmaster at ABC Corp. learns that a high percentage of users have the following pattern of reference to pages: (A, B,A, C). This means that a user accesses page A, then page B, then back to page A, and finally to page C. Based on this observation, he determines that a link is needed directly to page C from page B. He then adds this link.

Use of Web usage mining

- Personalization for a user can be achieved by keeping track of previously accessed pages.
- By determining frequent access behavior for users, needed links can be identified to improve the overall performance of future accesses.
- Identifying common access behaviors can be used to improve the actual design of Web pages and to make other modifications to the site.
- Web usage patterns can be used to gather business intelligence to improve sales and advertisement.
- Gathering statistics concerning how users actually access Web pages may or may not be viewed as part of mining.

- Web usage mining actually consists of three separate types of activities
- Preprocessing activities center around reformatting the Web log data before processing.
- Pattern discovery activities form the major portion of the mining activities because these activities look to find hidden patterns within the log data.
- Pattern analysis is the process of looking at and interpreting the results of the discovery activities.

- There are many issues associated with using the Web log for mining purposes:
- Identification of the exact user is not possible from the log alone.
- With a Web client cache, the exact sequence of pages a user actually visits is difficult to uncover from the server site. Pages that are re-referenced may be found in the cache.
- There are many security, privacy, and legal issues yet to be solved.

Web Mining Applications

- Targeted advertising: Targeting is any technique that is used to direct business marketing or advertising to the most beneficial subset of the total population. Thus , advertising costs can be reduced.
- Web mining helps to improve the power of web search engine by classifying the web documents and identifying the web pages.
- Web mining is used to predict user behaviour
- Web mining can be used for improving and enhancing the process of E-learning environments, digital libraries.
- Web mining techniques are also used for protection of user system or logging information against cybercrimes.
- Web mining techniques can support a web enabled electronic business to improve on marketing, customer support and sales operations.

Preprocessing

- The Web usage log probably is not in a format that is usable by mining applications.
- As With any data to be used in a mining application, the data may need to be reformatted and cleansed.
- Some issues specifically related to the use of Web logs.
- Steps that are part of the preprocessing phase include cleansing, user identification ,session identification, path completion, and formatting.

- **Let P be a set of literals, called pages or clicks, and U be a set of users. A log is a set of triples {(U1, p1 , t1), ... , (un , Pn , tn)} where ui E U pi E P, and ti is a timestamp**.

- Standard log data consist of the following: source site, destination site, and timestamp.
- The source and destination sites could be listed as a URL or an IP address.
- The definition assumes that the source site is identified by a user ID and the destination site is identified by a page ID.
- Additional data such as Web browser information also may be included.
- Before processing the log, the data may be changed in several ways.
- For security or privacy reasons, the page addresses may be changed into unique page identifications.
- This conversion also will save storage space.
- The data may be cleansed by removing any irrelevant information. As an example, the log entries with figures (gif, jpg, etc.) can be removed

- Data from the log may be grouped together to provide more information.
- All pages visited from one source could be grouped by a server to better understand

  the patterns of page references from each user .
- References to the same site may be identified and examined to better understand who visits this page.
- A common technique is for a server site to divide the log records into sessions.
- **Let L be a log. A session S is an ordered list of pages accessed by a user, i.e., S = ((p1, t1), (p2 , t2), ... , (pn , tn)), where there is a user ui E U such that { (u1 , p1 , t1), (u2 , p2 , t2), ... , (un, Pn , tn")} subset of L. Here ti <=tj iff i <= j. Since only the ordering of the accesses is our main interest, the access time is often omitted. Thus, we write a session S as (P1 , P2 , ... , Pn).**

- A session is a set of page references from one source site during one logical period.
- A session would be identified by a user logging into a computer, performing work, and then logging off.
- The login and logoff represent the logical start and end of the session. With Web log data, this is harder to determine.
- Several approaches can be used to identify these logical period:

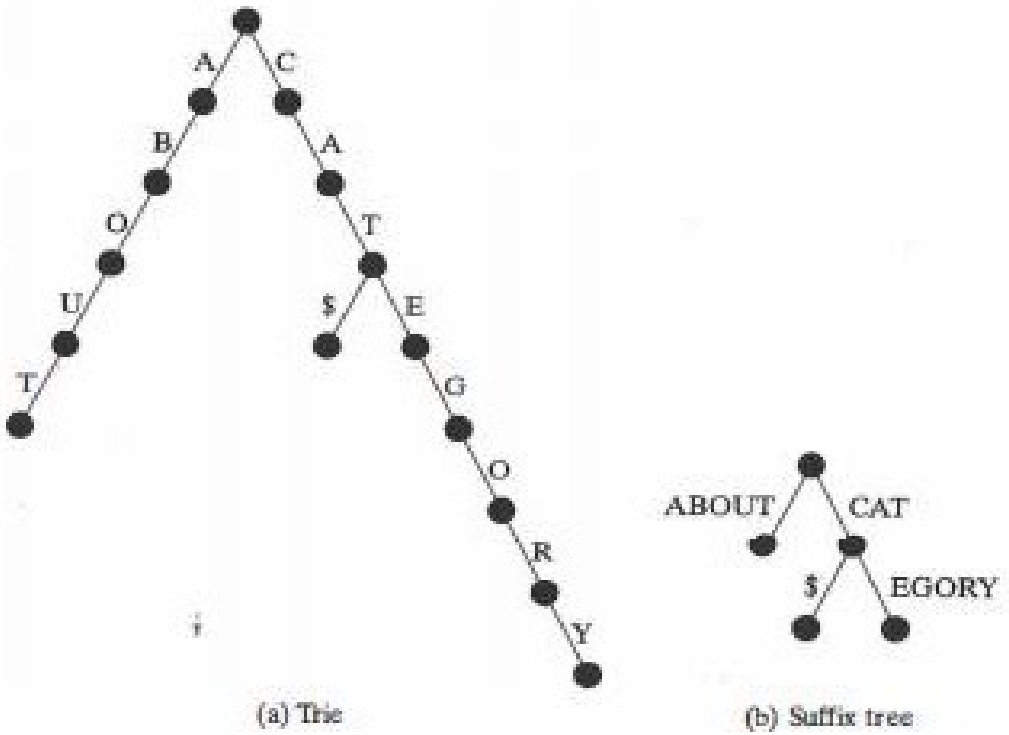
  - ❖ Combine all records from the same source site that occur within a time period.
  - ❖ Add records to a session if they are from the same source site and the time between two consecutive timestamps is less than a certain threshold value.

- Associated with each session is a unique identifier, which is called a session ID.
- The length of a session S is the number of pages in it, which is denoted as len(S).

- There are many problems associated with the preprocessing activities, and most of these problems center around the correct identification of the actual user.
- User identification is complicated by the use of proxy servers, client side caching, and corporate firewalls.
- Tracking who is actually visiting a site (and where they come from) is difficult.
- Even though a visit to a Web page will include a source URL or IP address that indicates the source of the request, this may not always be accurate in determining the source location of the visitor.
- Users who access the Internet through an Internet service provider (ISP) will all have the source location of that provider. It is not unique to the individual.
- In addition, the same user may use different ISPs.
- there will be many users accessing the Web at the same time from one machine.
- Cookies can be used to assist in identifying a single user regardless of machine used to access the Web.
- A cookie is a file that is used to maintain client-server information between accesses that the client makes to the server.
- The cookie file is stored at the client side and sent to the server with each access.

Data Structures

- Several unique data structures have been proposed to keep track of patterns identified during the Web usage mining process.
- A basic data structure that is one possible alternative is called a trie.
- A trie is a rooted tree, where each path from the root to a leaf represents a sequence.
- Tries are used to store strings for pattern-matching applications.
- Each character in the string is stored on the edge to the node.
- Common prefixes of strings are shared.
- A problem in using tries for many long strings is the space required.
- This is shown in Figure 7.5(a), which shows a standard trie for the three strings {ABOUT, CAT, CATEGORY}.
- Note that there are many nodes with a degree of one.
- This is a waste of space that is solved by compressing nodes together when they have degrees of one.



(a) Trie (b) Suffix tree

FIGURE 7.5: Sample tries.

- Ftgure 7.5(b) shows a compressed version of this trie.
- Here a path consisting of nodes with single children is compressed to one edge.
- Note in both trees the extra edge labeled "$." This symbol (or any symbol that is not in the alphabet and is used to construct the strings) is added to ensure that a string that is actually a prefix of another (CAT is a prefix of CATEGORY) terminates in a leaf node.
- The compressed trie is called a suffix tree.

- A suffix tree has the following characteristics:

  - Each internal node except the root has at least two children.
  - Each edge represents a nonempty subsequence.
  - The subsequences represented by sibling edges begin with different symbols.

- With the help of a suffix tree, it is efficient not only to find any subsequence in a sequence, but also to find the common subsequences among multiple sequences.
- A suffix tree can also be constructed from a sequence in time and space linear in the length of the sequence.
- When given one session of page references, many different patterns may be found.
- The exact number of patterns depends on the exact definition of the pattern to be
  found.

PATTERN DISCOVERY IN WEB USAGE MINING

Web usage mining

- Web usage mining is designed to discover useful patterns in Web usage data.

Web logs.

- Web logs record the user's browsing of a Web site, and the patterns provide useful information about the user's browsing behavior.
- Such patterns can be used for Web design, improving Web server performance, personalization, etc.
- Pattern discovery applies methods and algorithms from different fields such as statistics, data mining, machine learning, etc., to the prepared data.

- By applying the statistical techniques to the Web usage data, some useful statistics about the users' browsing behavior can be obtained.(e.g., the average view time of a page, and the most frequently accessed pages, etc)
- By applying clustering (unsupervised learning) to the usage data, groups of users which exhibit similar browsing behavior can be found.
- Such knowledge is useful for market segmentation and personalization.
- The focus here is on discovering traversal patterns from the Web usage data.
- A traversal pattern is a list of pages visited by a user in one session

TRAVERSAL PATTERNS AND DISCOVERING METHODS

Five different types of traversal patterns that can be mine in the Web usage data, namely, Association Rules Sequential Patterns, Frequent Episodes, Maximal Frequent Forward Sequences, and Maximal Frequent Sequences are described.

TABLE 7.1: Comparison of Different Types of Traversal Patterns (from [XD01a])

	Ordering	Duplicates	Consecutive	Maximal	Support
Association rules	N	N	N	N	$\frac{\text{freq}(X)}{\text{\# transactions}}$
Episodes	Y[1]	N	N	N	$\frac{\text{freq}(X)}{\text{\# time windows}}$
Sequential patterns	Y	N	N	Y	$\frac{\text{freq}(X)}{\text{\# customers}}$
Forward sequences	Y	N	Y	Y	$\frac{\text{freq}(X)}{\text{\# forward sequences}}$
Maximal frequent sequences	Y	Y	Y	Y	$\frac{\text{freq}(X)}{\text{\# clicks}}$

[1] Serial episodes are ordered, parallel episodes are not, and general episodes are partially ordered.

Ordering: the pages in a traversal pattern can be ordered or not.

Duplicates: which indicate whether backward traversals are allowed in the traversal pattern.

Contiguity: the page references in a traversal pattern may be contiguous or not.

Maximality: a frequent pattern is maximal if it is not contained in any other frequent pattern. A pattern could be maximal or not.

Association Rules

- Association rules were originally proposed for market basket .
- Association rules describe the associations among items bought by customers in the same transaction, e.g., 80% of customers who bought diapers also bought beer in some store.
- Here we are finding large itemsets.
- A page is regarded as an item, and a session is regarded as a transaction with both duplicates and ordering ignored.
- The support is defined to be the number of occurrences of the itemset divided by the number of transactions or sessions

- To mine association rules from the transactions, there are two steps:

  First finding the frequent item sets, and then generating association rules from the frequent item sets.

- An item set is frequent, if the support for the itemset is not less than some predefined threshold.
- The support for an itemset in a database of transactions is defined as the percentage of the transactions that contain the itemset.

Sequential Patterns

- Sequential patterns were also originally proposed for market basket data. For example, customers buy a digital camera, then a photo printer, and then photo papers.
- Such sequential patterns capture the purchasing behavior of customers over time.
- The sessions are ordered by the user id and the access time.
- As for association rules, the duplicate pages are discarded.
- Then for each user, there is a user sequence, which consists of all sessions of the user.

- A sequential pattern is a maximal sequence of itemsets whose support is not less than some predefined threshold.
- A sequence is maximal if it is not contained in any other sequence.
- The support of a sequence is the percentage of users who have the pattern.
- Algorithm AprioriAll was proposed for finding all sequential patterns given some support threshold. AprioriAll was then improved by Generalized Sequential Patterns (GSP).

Frequent Episodes

- Frequent episodes were originally proposed for telecommunication alarm analysis.
- The clicks (pages) correspond to events.
- Episodes are collections of events, which occur together within some time window.
- In general, they are partially ordered sets of events. There are two special types of episodes: parallel episodes and serial episodes.
- They differ in whether the events in the episodes are ordered.
- In parallel episodes the events are not ordered, while in serial episodes the events are totally ordered .

- General episode is one where the events satisfy some partial order. (Note that even though these seem similar to the idea of sequential patterns and association rules, the added constraint of a time window does make an episode different from either of these.)

- An episode is frequent if it occurs in the event sequence not less than some predefined threshold.

- They are ordered by the access time, and usually the users need not be identified, i.e., there are no sessions.

Maximal Frequent Forward Sequences

- Maximal Frequent Forward Sequences (MFFS for short) was referred to as large reference sequence.
- An MFFS describes the path traversal behavior of the user in a distributed information-providing environment like World Wide Web.
- There are two steps to mine MFFSs from the sessions.
- First each session is transformed into maximal forward sequences (i.e., the backward traversals are removed).
- The MFFSs are then mined using level-wise algorithms from the maximal forward sequences.

- **In the raw sessions,** there are often backward traversals made by the user

- A backward traversal means revisiting a previously visited page in the same user session. It is assumed that such backward traversals happen only because of the structure of the Web pages, not because the user wants to do this. When a backward traversal occurs, a forward traversal path terminates.
- This resulting forward traversal path is called maximal forward sequence

- An MFFS is a traversal sequence (consecutive subsequence of a maximal forward sequence) that appears not less than some predefined threshold in the set of maximal forward sequences.

- The pages in an MFFS are required to be consecutive in the maximal forward sequences, and an MFFS is also maximal, which means that it is not a subsequence of any other frequent traversal sequence.

Input:

$D = \{S_1, S_2, \ldots, S_k\}$ //Database of sessions

s //Support

Output:

Maximal reference sequences

Maximal frequent forward sequences algorithm:

find maximal forward references from D;

find large reference sequences from the maximal ones;

find maximal reference sequences from the large ones;

Maximal Frequent Sequences

- In contrast to maximal frequent forward sequences, MFSs do not remove backward traversals from the sessions.
- It was argued in that such backward traversals are useful for discovering the structures of the Web pages.

For example, if a pattern <A, B, A, C> is found frequent, it may suggest that a direct link from page B to page C is needed, while the resulting maximal forward sequences <A, B> and <A, C> lose such information.

- **An MFS is a traversal sequence (consecutive subsequence of a session)** that appears not less than some predefined threshold. Since the backward traversals are kept in the sessions, a traversal sequence may occur in a session more than once.
- In order to measure the actual number of occurrences of a traversal sequence, the support of an MFS is defined as the ratio of the actual number of occurrences to the total length of all sessions.
- The length of a session is the number of clicks in the session. The pages in an MFS are required to be consecutive in the sessions, and an MFS is also maximal.