Skill Based Subject – III: DATA MINING AND WAREHOUSING - 18BIT55S

UNIT IV: Web Data Mining: Introduction – Web Terminology and Characteristics – Locality and Hierarchy in the Web – Web Content Mining – Web Usage Mining – Web Structure Mining – Web Mining Software. Search Engines: Search Engine Functionality - Search Engine Architecture – Ranking of Web Pages.

TEXT BOOK

G.K Gupta, "Introduction to Data Mining with Case Studies", Prentice Hall of India(Pvt) Ltd, India, 2008.

Prepared by : Mrs. G. Shashikala, Assistant Professor, PG Department of Information Technology

Web mining

- Web mining is the application of data mining techniques to find interesting and potentially useful knowledge from web data. Either the hyperlink structure of the web or the web log data or both is used in the mining process.
- Web mining is divided into several categories

1 Web content mining : it deals with discovering Useful information or knowledge from Web page contents.

2 Web structure mining: It deals with discovering and modelling the link structure of the web

3 Web usage mining: It deals with understanding user behaviour in interacting with the web or with a website

-

- The following are the major differences between searching conventional text and searching the web:
  1 Hyperlink: The text documents do not have hyperlinks, while the links are very important components of web documents

2 Types of information : Webpages differ in structure quality and their usefulnesss. Web pages consist of text frames, multimedia objects, animation and other types of information. Documents mainly consist of text but may have tables, diagrams, figures

3 Dynamics : The text documents do not change unless a new edition of a book appear , while webpages change frequently

4 Quality: The text documents are usually of high quality, but much of the information on the web is of low quality

5 Huge size : Although some of the libraries are very large, the web in comparison is much larger

6 Document use : Compared to the use of conventional documents, the use of web documents is very different

Web terminology and characteristics:

- Some of the web terminology based on W3C are :
- The world wide web (WWW) is the set of all the nodes which are interconnected by hypertext links
- A link expresses one or more relationships between two or more resources. Links may also be established within a document by using anchors
- a web page is a collection of information consisting of one or more web resources and identified by a single URL. A web site is a collection of interlinked web pages, including a homepage residing at the same network location
- In addition to simple text, HTML allows embedding of images, sounds and video streams
- A client browser is the primary user interface to the web. It is a program which allows a person to view the contents of the Web pages, and for navigating from one page to another
- A uniform resource locator (URL) is an identifier for an abstract or physical resource, for example a server and the file path or index . URLs are location dependent and each URL contains four distinct parts namely the protocol types(http), the name of the web server, the directory path and the file name
-

- A Web server serves web pages using http to client machines so that a browser can display them
- A Client is the role adapted by an application when it is retrieving a web resource
- A proxy is an intermediary which acts as both a server and the client for the purpose of retrieving resources on behalf of other clients. Clients using a proxy know that the proxy is present and that it is an intermediary
- A domain name server is a distributed database of name to address mappings
- Yeah cookie is the data sent by a web server to a web client, to be stored locally by the client and sent back to the server on subsequent requests
-

Locality and Hierarchy in the web

- Most social structures tend to organize themselves as hierarchies. The web shows a strong hierarchical structure.
- Web pages can be classified into several types :

  1 Home page or the head page : represents an entry point for the web site of an enterprise

  2 Index page : assists the user to navigate through the enterprise's web site

  3 Reference page : provides some basic information that is used by a number of pages . For ex., link to a page that provides enterprise's privac policy

  4 Content page : provides content and are often the leaf nodes of a tree

Web content mining

- This deals with discovering useful information from the web
- The algorithm proposed is called Dual Iterative Pattern Relation Extraction (DIPRE). It works as follows:

  1 Sample : Start with a sample provided by the user

  2 Occurrences : Find occurrences of tuples starting with those in S. Once tuples are found, the context of every occurrence is save. Let these be O. O→S

  3 Pattern : Generate patterns based on the set of occurrences O. This requires generating patterns with similar contexts. P →O

  4 Match patterns : The web is now searched for the patterns

  5 Stop if enough matches are found. Else, go to Step 2

Web usage mining

- The objective of web usage mining is to understand and predict user behaviour in interacting with the web or with the website in order to improve the quality of service
- using some tools the following information may be obtained
- number of hits
- number of visitors
- visitor referring website
- visitor referral website
- entry point
- Visitor time and duration
- path analysis
- visitor IP address
- browser type
- Platform
- Cookies
- it is decidable to collect information on
- Path Traversed
- conversion rates
- impact of advertising
- impact of promotions
- website design
- customer segmentation
- enterprise search
-

Web structure mining

- The aim of web structure mining is to discover the link structure or the model that is assumed to underlie the web. The Hyperlink Induced Topic Search (HITS) algorithm is used for this .HITS algorithm has two major steps
- 1 sampling step : It collects relevant web pages for a given topic
- 2 Iterative step m: It finds hubs and authorities using the information collected during
- sampling
- step 1 - sampling step
- HITS algorithm expands the root set R into a base set S by using the following algorithm:
- 1 let S = R
- 2 for each page in S, do steps 3 to 5
- 3 let T be the set of all pages S points to
- 4 let F be the set of all pages that point to S
- 5 let S = S + T + some or all of F
- 6 delete all links with the same domain name
- 7 this S is returned
-

- step 2 - finding hubs and authorities
- 1 let a page p have a non-negative authority weight Xp and a non negative hub weight Yp
- 2 the weights are normalized so their squared sum for each type of weight is 1 since only the relative weights are important
- 3 for a page p, the value of Xp is updated to be the sum of Yq over all pages q that link to p
- 4 for a page p the value of Yp is updated to be the sum of Xq over all pages q that p links to
- 5 continue with step 2 unless a termination condition has been reached
- 6 on termination the output of the algorithm is a set of pages with the largest Xp and Yp weights
-

Web mining software

- 123LogAnalyzer
- Analog(from Dr. Stephen Turner)
- Azure web log analyser
- ClickTracks
- Datanautics G2 and Insight 5
- LiveStats.NET
- NetTracker Web Analytics
- Nihuo web log analyser
- Webanalyst from megaputer
- Weblog expert 3.5
- Webtrends 7 from netiq
- WUM – web utilization miner

Search Engines

- Introduction

- The search engines, directories, portals and indexes are the web's "catalogues" allowing a user to search the web for required information.
- Google is the largest global search engine followed by Yahoo! And msn.com
- It is reported that users spend 70% of their online time searching the web

- A web search is different from the text document search because of the following factors:
  - Bulk : the web is much larger than any set of documents used in information retrieval applications.
  - Diversity : the web is very diverse, consisting of text, images, movies, audio, animation and other multimedia content
  - Growth : the web continues to grow exponentially
  - Dynamic : the web changes significantly with time
  - Demanding users : users are very impatient, and they demand immediate result, otherwise they abandon the search and move on to something else.

- Duplication : it is estimated that 30% of the web content is duplicated
- Hyperlinks : web documents contain hypertext links to other web documents
- Index pages : many search results return index pages from various sites providing little content but many links

Search engine functionality

- A search engine carries out a variety of tasks. These include :
  1. Collecting information : A search engine collects web pages or information about them by Web crawling or by human submission of pages
  2. Evaluating and categorizing information : When web pages are submitted to a directory, it has to be evaluated and decided whether the page has to be selected. It has to be categorized based on some ontology used by the search engine.
  3. Creating a database and creating indexes : information collected has to be stored in a database or file system. Indexes must be created to search information efficiently

4. Computing ranks of the web documents : the information used include frequency of key words, value of in-links and out-links from the page and frequency of use of the page.
5. Checking queries and executing them : queries posed by the users has to be checked, for spelling errors and whether words in the query are recognizable
6. Presenting results : the search engine must determine, what results to present and how to display them
7. Profiling the users : search engines carry out user profiling that deals with the way users use search engines

Search engine architecture

- Search engines are different in terms of size, indexing techniques, page ranking algorithms or speed of search.
- The major components in a search engine architecture are :
  - The crawler and the indexer : It collects pages from the web, creates and maintains the index
  - The user interface : It allows users to submit queries and enables result presentation
  - The database and the query server : It stores information about the web pages and processes the query and returns results

significantly. We

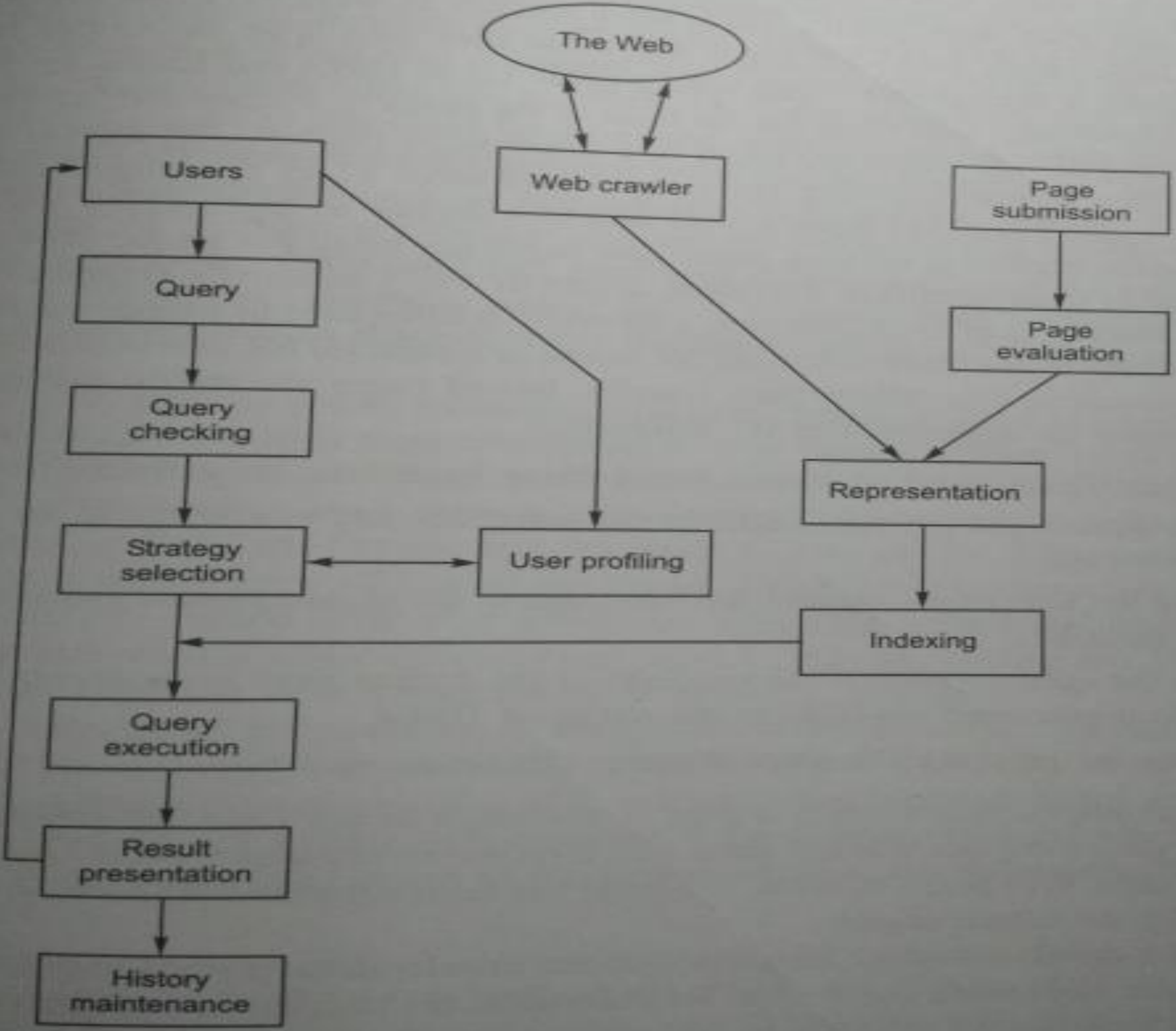

Figure 6.2 Typical architecture of a search engine.

- All search engines include a crawler, indexer, and a query server
- The Crawler
  - The crawler(or spider or robot or bot)is an application program that carries out a task similar to graph traversal
  - It is given a set of starting URLs that it uses to traverse the web by retrieving a page, initially from the starting set
  - Crawlers tend to return to each site on a regular basis, to look for changes.
  - Frequently changing sites like newspaper sites are visited even every few hours
- Crawling is bandwidth-bound
- It is given a set of starting URLs and fetches the pages
- The crawler then reads the out-links of the pages and fetches those pages. This continues until no new pages are found or a threshold is reached.
- Each page found by the crawler is not stored as a separate file and lots of pages are stuffed into one file

- The algorithm followed by crawlers is :
  - Find base URLs- a set of known and working hyperlinks are collected
  - Build a queue- put the base URLs in the queue and add new URLs to the queue as more are discovered
  - Retrieve the next page – retrieve the next page in the queue, process and store in the search engine database
  - Add to the queue – check if the out-links of the current page have already been processed. Add the unprocessed out-links to the queue of URLs
  - Continue the process until some stopping criteria is met

Ranking ofweb pages

- Google has the ranking algorithm, called the Page Rank algorithm, and it is based on the hyperlinks as indicators of a page's importance.
- Yahoo! Web rank
  - Yahoo! Has developed its own page ranking algorithm called Web Rank
  - Here the rank is calculated by analysing the Web page text, title and description , its associated links and other unique document characteristics
  - So if many users visit a particular site, that might be a factor in helping the site get a better web rank score