WEB BASED INFORMATION SYSTEM AND NAVIGATION

7.1. THE STRUCTURE OF THE WEB

The Web is an integral part of our daily lives. Studies report that billions of people are visiting billions of pages on the Web every day. It is reasonable, therefore, that one would like to know what the graph that connects those Web pages looks like.

The Web Graph, as it is known, is the graph comprised of Web pages interconnected through the hyperlinks they contain. For simplicity, let's assume that pages on the Web are only "static", that is, they are composed of text and binary files residing on servers on the Internet.

Bowtie structure of Web

- A global map of web ,using strongly connected component as the basic building blocks
- Classify nodes by their ability to reach and be reached from giant SCC but cannot be reached from giant SCC(Strongly Connected Components)
- IN :nodes that can reach from giant SCC but cannot be reached from it-i.e. Nodes that are "upstream" of it
- Out : nodes that can be reached from giant SCC but cannot reach it- i.e. Nodes are "downstream" of it

A central core contains pages between which users can surf easily. Another large cluster, labelled 'in', contains pages that link to the core but cannot be reached from it. These are often new pages that have not yet been linked to. A separate 'out' cluster consists of pages that can be reached from the core but do not link to it, such as corporate websites containing only internal links. Other groups of pages, called 'tendrils' and 'tubes', connect to either the in or out clusters, or both, but not to the core, whereas some pages are completely unconnected. To illustrate this structure, the researchers picture the web as a plot shaped like a bow tie with finger-like projections.

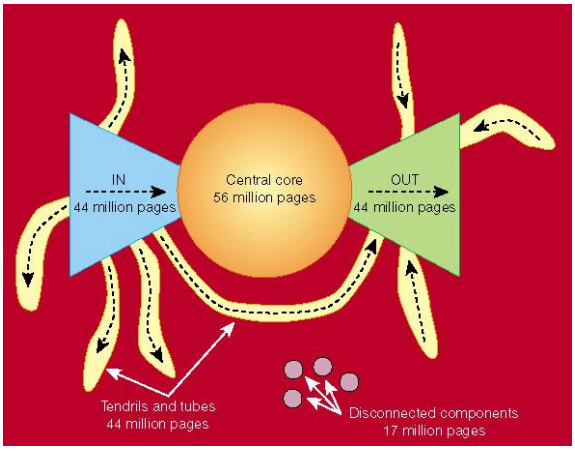


Fig: 7.1 Bowtie Model of Web

7.2 LINK ANALYSIS

Link analysis is a data-analysis technique used to evaluate relationships (connections) between nodes. Relationships may be identified among various types of nodes (objects), including organizations, people and transactions. Link analysis has been used for investigation of criminal activity (fraud detection, counterterrorism, and intelligence), computer security analysis, search engine optimization, market research, medical research, and art.

Link analysis is used for 3 primary purposes:-

1. Find matches in data for known patterns of interest;
2. Find anomalies where known patterns are violated;
3. Discover new patterns of interest (social network analysis, data mining).

WHY?

- Strategy study is an important task of MIS development, The analysis of exterior value link and interior value can be a MIS development strategy
- Web based information system-searching the web.
- Data analysis technique used to evaluate connections between nodes
- Knowledge discovery for strategic planning of Information System

APPLICATION

- Analysis of hyperlinks and graph structure of the web
- Computer Security Analysis
- Search Engine optimization

- Market research
- Criminal Activity

Fraud detection

Counter terrorism

LINK ANALYSIS IN WEB SEARCH

- Used by web search engines in computing a composite score for a web page
- Web site is ranked accordingly for any query
- Useful indicator of what pages(s) to crawl next while crawling the web
- The web is not just a collection of documents – its hyperlinks are important!
- A link from page A to page B may indicate:
    - ✓ A is related to B, or
    - ✓ A is recommending, citing, voting for or endorsing B
- Links are either
    - ✓ Referential – click here and get back home, or
    - ✓ Informational – click here to get more detail

Links effect the ranking of web pages and thus have commercial value.

CITATION ANALYSIS

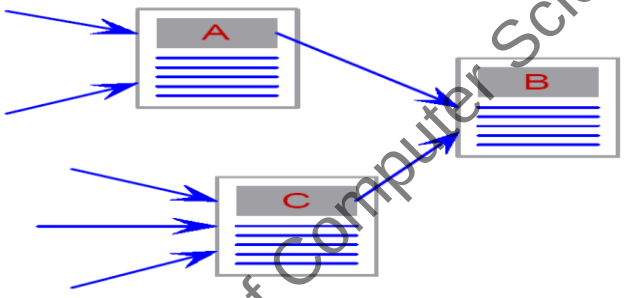
- The **impact factor** of a journal = A/B
    - A is the number of current year citations to articles appearing in the journal during previous two years.
    - B is the number of articles published in the journal during previous two years.

Journal Title (AI)	Impact Factor (2004)
J. Mach. Learn. Res.	5.952
IEEE T. Pattern Anal.	4.352
IEEE T. Evolut. Comp.	3.688
Artif. Intell.	3.570

Mach. Learn.	3.258

PAGERANK - MOTIVATION

- A link from page *A* to page *B* is a **vote** of the author of *A* for *B,* or a **recommendation** of the page.
- The number incoming links to a page is a measure of importance and authority of the page.
- Also take into account the quality of recommendation, so a page is more important if the sources of its incoming links are important.



PAGERANK (*PR*) – DEFINITION

$$PR(W) = \frac{T}{N} + (1-T)\left(\frac{PR(W_1)}{O(W_1)} + \frac{PR(W_2)}{O(W_2)} + \ldots + \frac{PR(W_n)}{O(W_n)}\right)$$

- *W* is a web page
- *Wi* are the web pages that have a link *to W*
- *O*(W*i*) is the number of out links from W*i*
- *T* is the teleportation probability
- *N* is the size of the web

7.3. SEARCHING THE WEB

A **web search engine** is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages (SERPs). The information may be a mix of web pages, images, and other

types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler.

TYPES OF SEARCH ENGINES

- Search engines break down into two type directories and indexes.

1. **Directories**
   SE like yahoo! Are good at identifying general information .They classify websites into categories e.g.: the result of search will be list of websites related to search term

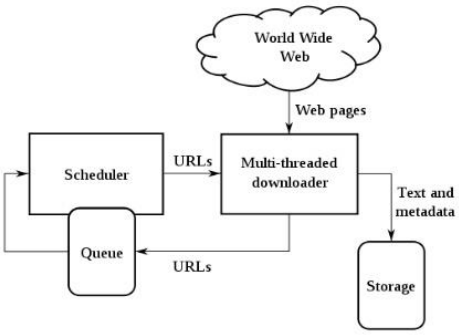
2. **Indexes**
   SE like Google and Bing identify the text on individual pages of a website that match the search criteria, even if the site itself has nothing to do with what is being searched.

HOW SEARCH ENGINE WORKS?

1. Web crawling
2. Indexing
3. Searching

- Spider "crawls" the web to find new documents (web pages, other documents) typically by following hyperlinks from websites already in their database
- Search engines indexes the content (text, code) in these documents by adding it to their databases and then periodically updates this content
- Search engines search their own databases when a user enters in a search to find related documents (not searching web pages in real-time)
- Search engines rank the resulting documents using an algorithm (mathematical formula) by assigning various weights and ranking factors

High Level Architecture of a Web Crawler



Web crawlers are a central part of search engines, and details on their algorithms and architecture are kept as business secrets

Google is the world's most popular search engine, with a market share of 67.49 percent as of September, 2015, comes in at second place. The world's most popular search engines are:

Search engine	Market share in September 2015
Google	69.24%
Bing	12.26%
Yahoo!	9.19%
Baidu	6.48%
AOL	1.11%
Ask	0.24%
Lycos	0.00%

HOW GOOGLE WORKS

- Google has three distinct parts:
- Google bot ,a web crawler
- The indexer
- The query processor

Google Bot: Google Bot is Google's web crawling robot which finds and retrieves pages on the web and hands them off to the Google indexer

Google Bot finds pages into ways: through an and URL form

- www.google.com/addurl.html and through finding links by crawling the web.

Indexer: Google bot gives the indexer the full text of the pages it finds. These pages are stored in Google's index database. This index is stored alphabetically by search term, with each index entry storing a list of documents in which the term appears and the locations within the text where it occurs

- To improve search performance ,Google ignores(doesn't index) common words called stop words (such as the is ,on , or, of, how ,why, as well as certain single digits and single letters)

Query Processor

The query processor has several parts, including user interface (search box) the "engine" that evaluates queries and matches them to relevant documents, and the result formatter.

- PageRank is Google's system for ranking web pages. A page with a higher PageRank is deemed more important and is more likely to be listed above a page with a lower PageRank

7.4. NAVIGATING THE WEB

Web navigation refers to the process of navigating a network of information resources in the World Wide Web, which is organized as hypertext or hypermedia. The user interface that is used to do so is called a web browser.

A central theme in web design is the development of a web navigation interface that maximizes usability. A website's overall navigational scheme includes several navigational pieces such as global, local, supplemental, and contextual navigation; all of these are vital aspects of the broad topic of web navigation.

Hierarchical navigation systems are vital as well since it is the primary navigation system. It allows for the user to navigate within the site using levels alone, which is often seen as restricting and requires additional navigation systems to better structure the website.

The global navigation of a website, as another segment of web navigation, serves as the outline and template in order to achieve an easy maneuver for the users accessing the site, while local navigation is often used to help the users within a specific section of the site.

All these navigational pieces fall under the categories of various types of web navigation, allowing for further development and for more efficient experiences upon visiting a webpage.

Principles for good navigation design

A site must:

1. Let me know where all the times am
2. Clearly differentiate hyperlinks from contents
3. Let me know clearly where I can go from here
4. Let me see where I've already been
5. Make it obvious what to do get somewhere
6. Indicate what clicking a link will do

Types of Navigation Systems

The use of website navigation tools allow for a website's visitors to experience the site with the most efficiency and the least incompetence. A website navigation system is analogous to a road map which enables webpage visitors to explore and discover different areas and information contained within the website. There are many different types of website navigation:

- ✓ **Hierarchical Website Navigation**

The structure of the website navigation is built from general to specific. This provides a clear, simple path to all the web pages from anywhere on the website.

- ✓ **Global Website Navigation**

Global website navigation shows the top level sections/pages of the website. It is available on each page and lists the main content sections/pages of the website.

- ✓ **Local Website Navigation**

Local navigation would the links with the text of your web pages, linking to other pages within the website

STYLES OF WEB NAVIGATION

The availability of different navigational styles allows for the information in the website to be delivered easily and directly. This also differentiates between categories and the sites themselves

to indicate what the vital information is and to enable the users' access to more information and facts discussed within the website.

Zheng has summarized and compared some common navigation system designs from an information seeking perspective, including:

- Text Links: The anchor text, link label, link text, or link title is the visible, clickable text in a hyperlink.
- Breadcrumbs: Breadcrumbs or breadcrumb trail is a navigation aid used in user interfaces. It allows users to keep track of their locations within programs or documents. The term comes from the trail of breadcrumbs left by Hansel and Gretel in the popular fairytale.
- Navigation Bar: A navigation bar or (navigation system) is a section of a website or online page intended to aid visitors in travelling through the online document.
- Sitemap: A site map (or sitemap) is a list of pages of a web site accessible to crawlers or users. It can be either a document in any form used as a planning tool for Web design, or a Web page that lists the pages on a Web site, typically organized in hierarchical fashion.
- Dropdown Menu: In computing with graphical user interfaces, a dropdown menu or drop-down menu or drop-down list is a user interface control GUI element ("widget" or "control"), similar to a list box, which allows the user to choose one value from a list.
- Fly out Menu: In computing with graphical user interfaces, a menu that flies out (either down or to the side) when you click or hover (mouseover) some GUI element.
- Named anchor: An anchor element is called an anchor because web designers can use it to anchor a URL to some text on a web page. When users view the web page in a browser, they can click the text to activate the link and visit the page whose URL is in the link

WHERE SHOULD YOU PUT NAVIGATION?

- Depends on the type of navigation
- The golden rules are:

-Put the most useful navigation where it's closest to hand

-Put navigation where the user is likely to look for it

DESIGN OF WEB NAVIGATION

- ✓ Interesting Navigation Designs
- ✓ Beautiful Vertical Navigation Designs
- ✓ Modern Navigation Design Trend
- ✓ Designing Drop-Down Menus

FUTURE OF WEB NAVIGATION

- ✓ Adaptive Website Navigation
- ✓ Browser Integrated Web Navigation

7.5. WEB USES MINING

Web mining - is the application of data mining techniques to discover patterns from the World Wide Web. Web mining can be divided into three different types – Web usage mining, Web content mining and Web structure mining.

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based

applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site.
Web usage mining itself can be classified further depending on the kind of usage data considered:

- **Web Server Data**: The user logs are collected by the Web server. Typical data includes IP address, page reference and access time.
- **Application Server Data**: Commercial application servers have significant features to enable e-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.
- **Application Level Data**: New kinds of events can be defined in an application, and logging can be turned on for them thus generating histories of these specially defined events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the categories above.

PROS OF WEB USES MINING

- ✓ Web usage mining essentially has many advantages which makes this technology attractive to corporations including the government agencies.
- ✓ This technology has enabled e-commerce to do personalized marketing, which eventually results in higher trade volumes.
- ✓ Government agencies are using this technology to classify threats and fight against terrorism. The predicting capability of mining applications can benefit society by identifying criminal activities.
- ✓ The companies can establish better customer relationship by giving them exactly what they need. Companies can understand the needs of the customer better and they can react to customer needs faster.
- ✓ The companies can find, attract and retain customers; they can save on production costs by utilizing the acquired insight of customer requirements.
- ✓ They can increase profitability by target pricing based on the profiles created. They can even find the customer who might default to a competitor the company will try to retain the customer by providing promotional offers to the specific customer, thus reducing the risk of losing a customer or customers.

CONS OF WEB USES MINING

- ✓ Web usage mining by itself does not create issues, but this technology when used on data of personal nature might cause concerns. The most criticized ethical issue involving web usage mining is the invasion of privacy. Privacy is considered lost when information concerning an individual is obtained, used, or disseminated, especially if this occurs without their knowledge or consent.
- ✓ Another important concern is that the companies collecting the data for a specific purpose might use the data for a totally different purpose, and this essentially violates the user's interests.
- ✓ The growing trend of selling personal data as a commodity encourages website owners to trade personal data obtained from their site. This trend has increased the amount of data being captured and traded increasing the likeliness of one's privacy being invaded. The companies which buy the data are obliged make it anonymous and these companies are considered authors of any specific release of mining patterns. They are legally responsible for the contents of the release; any inaccuracies in the release will result in serious lawsuits, but there is no law preventing them from trading the data.
- ✓ Some mining algorithms might use controversial attributes like sex, race, religion, or sexual orientation to categorize individuals. These practices might be against the anti-discrimination legislation. The applications make it hard to identify the use of such controversial attributes, and there is no strong rule against the usage of such algorithms with such attributes

WEB USAGE MINING: DATA SOURCE

- ✓ Web structure data (sites,map,links,etc.)
- ✓ Web content data
- ✓ User profile (May Not Be Available)
- ✓ Web Log (Web Usage Data,clickstream Data)

Examples of data mining uses-

Games, Business, Science and Engineering, Human Rights, Medical data mining, music data mining etc.

7.6. COLLABORATIVE FILTERING

Collaborative filtering (**CF**) is a technique used by some recommender systems (**Recommender systems** or **recommendation systems** sometimes replacing "system" with a synonym such as platform or engine) are a subclass of information filtering system that seek to predict the 'rating' or 'preference' that a user would give to an item.

Collaborative filtering has two senses, a narrow one and a more general one. In general, collaborative filtering is the process of filtering for information or patterns using techniques involving collaboration among multiple agents, viewpoints, data sources, etc. Applications of collaborative filtering typically involve very large data sets.

Collaborative filtering methods have been applied to many different kinds of data including: sensing and monitoring data, such as in mineral exploration, environmental sensing over large areas or multiple sensors; financial data, such as financial service institutions that integrate many financial sources; or in electronic commerce and web applications where the focus is on user data, etc.

In the newer, narrower sense, collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating).

Uses for CF: User Tasks

- ✓ What tasks users may wish to accomplish

- Help find new items I might like
- Advise a particular item

- Help find a user(or some users)
- Help group find something new

Uses for CF: System Tasks

- What CF systems support
- Recommended Items

  Eg: Amazon.com
- Predict for given item
- Constrained recommendations
- Recommended from set of items

Particle Issues: Ratings

1) Explicit Ratings

Users rate themselves for an item examples are:

- ✓ Asking a user to rate an item on a sliding scale.
- ✓ Asking a user to search.
- ✓ Asking a user to rank a collection of items from favorite to least favorite.
- ✓ Presenting two items to a user and asking him/her to choose the better one of them.
- ✓ Asking a user to create a list of items that he/she likes.

2) Implicit Rating

Observations of user behavior examples are:

- ✓ Analyzing item/user viewing times
- ✓ Keeping a record of the items that a user purchases online.
- ✓ Obtaining a list of items that a user has listened to or watched on his/her computer.

3) Analyzing the user's social network and discovering similar likes and dislikes

4) Can be collected with less or no cost to user

- ✓ Rating Scales -Scalar Ratings
- ✓ Numerical Scales- 1-5, 1-7, etc.
- ✓ Binary Ratings
- ✓ Agree/Disagree, Good/Bad etc.

- Unary Ratings
  Good, Purchase, etc.
- Absence of rating indicates no information

7.7. RECOMMENDER SYSTEMS

Recommender systems or **recommendation systems** (sometimes replacing "system" with a synonym such as platform or engine) are a subclass of information filtering system that seek to predict the 'rating' or 'preference' that a user would give to an item.

Recommender systems have become extremely common in recent years, and are applied in a variety of applications. The most popular ones are probably movies, music, news, books, research articles, search queries, social tags, and products in general. However, there are also recommender systems for experts, collaborators, jokes, restaurants, financial services, life insurance, persons (online dating), and Twitter followers .

The recommender system compares the collected data to similar and dissimilar data collected from others and calculates a list of recommended items for the user.

Why Recommender System?

- Enhances user experience
- Assist users in finding information
- Reduce search and navigation time
- Increase Productivity

Types of Recommender System

1. Content Based RS
2. Collaborative RS
3. Hybrid RS

1. Content Based RS

Content-based filtering methods are based on a description of the item and a profile of the user's preference. In a content-based recommender system, keywords are used to describe the items and a user profile is built to indicate the type of item this user likes.

In other words, these algorithms try to recommend items that are similar to those that a user liked in the past (or is examining in the present). In particular, various candidate items are compared with items previously rated by the user and the best-matching items are recommended. This approach has its roots in information retrieval and information filtering research.

LIMITATION OF CONTENT BASED RS

1. Not all contents is well represented by keywords e.g. images
2. Items represented by same set of features are indistinguishable
3. Overspecialization: Unrated items not shown
4. Users with thousands of purchase is a problem
5. New users: no history available
6. Shouldn't show items that are too different or too similar

2. Collaborative Recommender System

Collaborative filtering methods are based on collecting and analyzing a large amount of information on users' behaviors, activities or preferences and predicting what users will like based on their similarity to other users.

A key advantage of the collaborative filtering approach is that it does not rely on machine analyzable content and therefore it is capable of accurately recommending complex items such as movies without requiring an "understanding" of the item itself. Many algorithms have been used in measuring user similarity or item similarity in recommender systems.

Use other users' recommendation ratings to judge item's utility

- ✓ Key is to find users/user group whose interest match with current user
- ✓ Vector Space model widely used (directions of vectors are user specified ratings)
- ✓ More users , more rating: better results
- ✓ Can account for items dissimilar to ones seen in the past too
- ✓ Example:Movielens.org

LIMITATION OF COLLABORATIVE RS

1. Different users might use different scales. Possible solution: weighted
2. Rating, i.e. deviations from average rating
3. Finding similar users/user group isn't very easy
4. New User: No preferences available
5. New Item: No rating available
6. Demographic filtering is required
7. Multi-criteria ratings is required

3. Hybrid Recommender Systems

Recent research has demonstrated that a hybrid approach, combining collaborative filtering and content-based filtering could be more effective in some cases. Hybrid approaches can be implemented in several ways: by making content-based and collaborative-based predictions separately and then combining them; by adding content-based capabilities to a collaborative-based approach (and vice versa); or by unifying the approaches into one model.

Netflix is a good example of the use of hybrid recommender systems. They make recommendations by comparing the watching and searching habits of similar users (i.e. collaborative filtering) as well as by offering movies that share characteristics with films that a user has rated highly (content-based filtering).

POPULAR RS TECHNIQUES IN E-COMMERCE

1. Browsing
2. Similar item/s
3. Email
4. Text Comments
5. Average Rating
6. Top –N-results
7. Ordered search results

RELEVANCE TO INFORMATION ARCHITECTURE

1. Increased Find ability

2. Reduced searching efforts

3. Improved Organizational systems

4. Enhanced Browsing

5. Provide more useful “local navigation” options

6. “Targeted Advertising” a much better substitute to common advertisements that are often irrelevant

7.8. COLLECTIVE INTELLIGENCE

Collective intelligence is shared or group intelligence that emerges from the collaboration, collective efforts, and competition of many individuals and appears in consensus decision making. Information is constantly being added and updated by a large online community, e.g. Wikipedia

FEATURES

- Constantly updated information
- Continual user feedback
- Vast range of easily accessible content
- Minimum cost for extensive amounts of data-hosted
- Self-policing communities
- Greater accuracy from many sources

CONSTRAINTS

- Negative networking effects
- Concern over intellectual property online
- Vast number of users necessary to achieve reliability

EXAMPLES

- YAHOO
- WIKIPEDIA

- GOOGLE
- BLOGGING