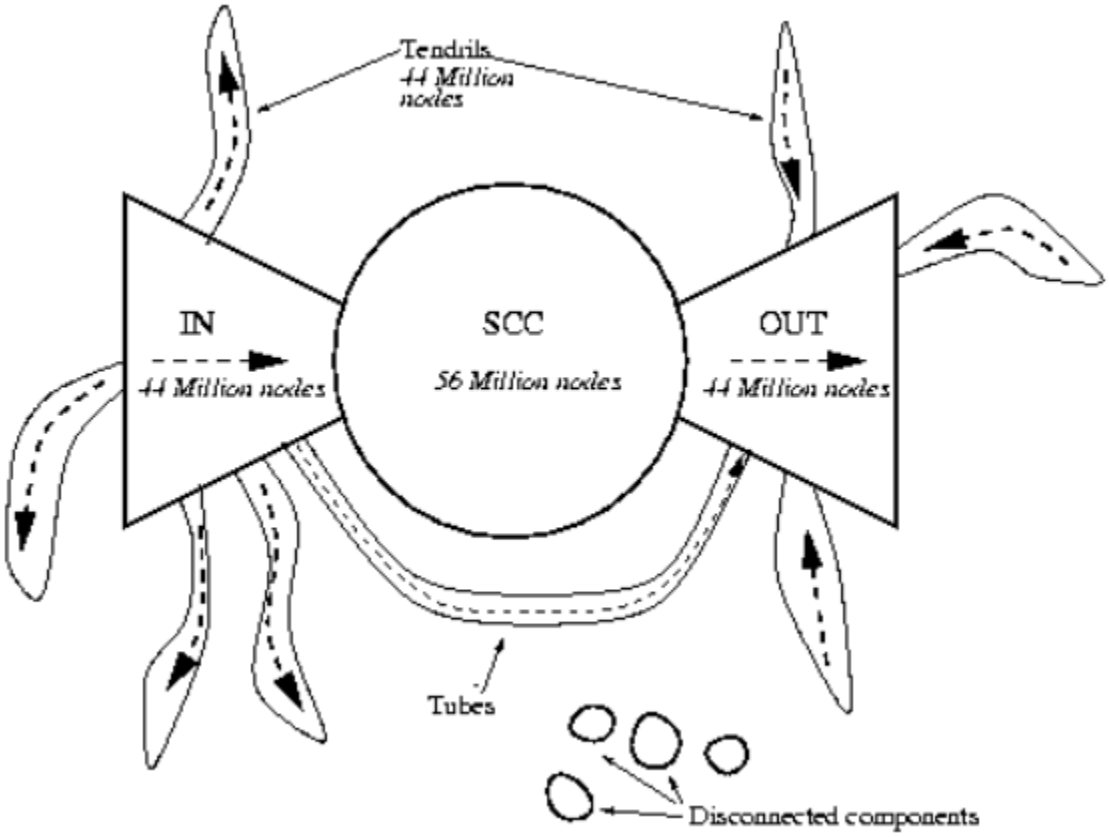
Web Based Information System and Navigation

Structure of Web

Bow-Tie Structure of the Web

- A structure of the web can be defined by the directed graph.
- Each node represents the page of the web.
- The directed arrow represents the reachability of one node from another node.
- The Bow-Tie structure of the web consists of following components:

1. Strongly Connected Components (SCC)
2. IN Component
3. OUT Component
4. Tendrils
5. Disconnected Components



Strongly Connected Components (SCC):

- A strongly connected component in a directed graph is a subset of the nodes such that:

a) Every node in the subset has a path to every other node.

b) The subset is not the part of some larger set where every node can reach every other nodes.

- A web consists of a giant strongly connected component.
- The number of web sites and search engines have links leading to directory type pages from which one can reach home pages of almost all of the web sites and also these websites link back to search engines.
- This shows the fact of mutual reachability forming the giant SCC.

IN Component:

- It consists of all the nodes that can reach the giant SCC but can not be reached from it.
- It is the nodes that are upstream of giant SCC.

OUT Component:

- It consists of all the nodes that can be reached from the giant SCC but can not reach giant SCC.
- It is the nodes that are downstream of giant SCC.

Tendrils:

- It consists of:
1. Nodes reachable from IN component that can not reach the giant SCC.
2. Nodes that can reach OUT but can not be reached from giant SCC.

Disconnected Component:

- It consists of the nodes that would not have any path to the giant SCC even if we completely ignored the direction of the edges.

Link Analysis

Web Link Analysis

- Link is the portion of a web page which refers to other pages.
- Link analysis means the process of analyzing the links present in the web pages.
- Link analysis is very important for the search engines to display their results.

- It helps to analyze whether the links present are active or dead.
- Link analysis helps the analyst to determine whether the search engine is able to find and index the website.
- Link analysis is used by the search engines to compute a composite score for a web page on any given query.

Page Rank

- Page rank is the composite score given by the search engines to the web pages to find and index them when user searches for the query.
- It is the algorithm that is used by the search engines to rank the websites in their results.
- It works by counting the number and quality of links to a page to determine a rough estimate of how important the website is.
- Page rank is a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set.

- Page rank is computed as follows:
1. In a network with n nodes, we assign all nodes the same initial Page Rank, set to be 1/n.
2. We choose a number of steps k.
3. We then perform a sequence of k updates to the Page Rank values, using the following rule for each update:
a) Basic Page Rank Update Rule: Each page divides its current Page Rank equally across its out-going links, and passes these equal shares to the pages it points to. (If a page has no out-going links, it passes all its current Page Rank to itself.) Each page updates its new Page Rank to be the sum of the shares it receives.

<u>Searching the Web</u>

Search Engine:

- Search engine is the huge database of internet resources that helps to locate information on the World Wide Web.
- Users can search for any information in a search engine by passing query in the form of keywords or phrase.
- The query is then searches in its database and the results are displayed back to the users.

Working of Search Engine:

1. The user enters the keyword to search for the required information through a query in the search engine user interface.

2. The search engine looks for the keyword in the index for predefined database instead of going directly to the web to search for the keyword.

3. It then uses software to search for the information in the database. This software component is known as web crawler.

4. Once web crawler finds the pages, the search engine then shows the relevant web pages as a result. These retrieved web pages generally include title of page, size of text portion, first several sentences etc.

5. User can click on any of the search results to open it to get the relevant information.

Architecture of Search Engine:

- Search Engine consists of following components:
1. Content Collection and Refinement
2. Search Core
3. User and Application Interface

Web Uses Mining

Web Mining:

- Web mining is the data mining technique that is used to discover patterns from the World Wide Web.
- It is the process of gathering information by mining (extracting something useful) the web.
- It is divided into three types:
1. Web Content Mining
2. Web Usage Mining
3. Web Structure Mining

	Web mining			
	Web Content Mining		Web Structure Mining	Web Usage Mining
	IR view	Db View		
View of Data	-Unstructured -Structured	-Semi Structured -Web Site as DB	-Link Structure	-Interactivity
Main Data	-Text documents -Hypertext documents	-Hypertext documents	-Link Structure	-Serves Logs -Browser Logs
Representation	-Bag of words, n-gram Terms, -phrases, Concepts or ontology -Relational	-Edge labeled Graph, -Relational	-Graph	-Relational Table -Graph
Method	-Machine learning -Statistical (including NLP)	-Proprietary algorithms -Association rules	-Proprietary algorithms	-Machine Learning -Statistical -Association rules
Application Categories	-Categorization -Clustering -Finding extract rules -Finding patterns in text	-Finding frequent sub structures -Web site schema discovery	-Categorization -Clustering	-Site Construction -adaptation and management -Marketing -User Modeling

---

Web Content Mining:

- Web content mining is the process of mining useful information and knowledge from the contents of the web pages and web documents.
- As the web contents are mostly text, images, audio and video files, NLP techniques are mostly used for mining.

Web Usage Mining:

- Web usage mining is the process of extracting patterns and information from server logs to gain insight on user activity including where the users are from, how many clicked what item on the site and the types of activities being done on the site.
- It provides basic insights on how the users are using the web.
- It helps to discover the web usage patterns from the web data to understand and serve the needs of web based applications.

Web Structure Mining:

- Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site.
- It helps to extract the patterns from the hyperlinks in the web.
- It helps to analyze the document structure to describe the structure of the web site.
- Web structure mining can be used for page ranking of the web sites for search engines.

Recommender System

- A recommender system is an information filtering system that seeks to predict the preference that a user would give to an item.
- It provides the recommendation to the user based on their previous historical data.
- It is assumed to be the alternative to search algorithms as they help the users discover items they might not have found by themselves.
- The recommender system produces a list of recommendation in two ways as follows:
1. Collaborative Filtering
2. Content based Filtering

Collaborative Filtering:

- Collaborative filtering is based on collecting and analyzing a large amount of information on users' behavior, activities or preferences and predicting what users will like based on their similarity to other users.
- It does not depend up on machine analyzable content, so is able to produce accurate recommendation for complex items too.
- It does not require understanding of the content of an item.
- It is based on the assumption that people who agreed in the past will also agree in the future and that they will like similar kinds of items as they liked in the past.
- The data on users' behavior can be collection explicitly (asking user to search, asking a user to rank items and so on) or implicitly (observing the items that a user views in an online store, analyzing viewing time of an item, keeping record of items that a user purchases online, analyzing social network of user and so on.)
- The collected data is compared to the similar and dissimilar data collected from others and calculates a list of recommended items for the user.

Problems of Collaborative Filtering:

1. Cold Start:
- The system requires a huge amount of existing data on a user so as to make accurate recommendations.
- This problem is termed as cold start.

2. Scalability:
- In the real world system, there are millions of users and products.
- So, to calculate recommendations, a large computational power should be possessed by the system.

3. Sparsity:
- All the users do not rate the items.
- So, even the most popular items may have few ratings.

Real World Examples : Collaborative Filtering

1. Last.fm (It recommends music based on a comparison of the listening habits of similar users.)
2. Facebook, MySpace, LinkedIn (They use collaborative filtering to recommend new friends, groups and other social connections by examining the network of connections between a user and their friends.)

Content Based Filtering:

- Content based filtering is based on a description of the item and a profile of user's preferences.
- Keywords are used to describe an item and a user profile is built to indicate the type of item this user likes.
- It recommends items that are similar to those that a user liked in the past.
- Item presentation algorithm is used to abstract the features of the items in the system.
- User profile are created by focusing on model of user's preference and history of user interaction with the recommender system.
- The system consists of item profile and content based profile of users based on the weighted vector of item features.
- The weights denote the importance of each feature to the user.
- It uses machine learning techniques like Bayesian classifier, decision tree and ANN to estimate the probability that the user is going to like the item.

Problems with Content Based Filtering:

1. It is effective to recommend same type of items as the user is using. For eg: recommending news articles based on browsing of news.

Real World Examples : Content Based Filtering

1. Pandora Radio (It plays music with similar characteristics to that of a song provided by the user as the initial seed)
2. Rotten Tomatoes (Movie recommendation system)

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.
- It is an emergent property between expert and ways of processing information.
- The main goal of collective intelligence is mutual recognition and enrichment of individuals rather than the cult of hypostatized communities.
- In case of computer science, collective intelligence is the capacity of networking information system to enhance the collective pool of social knowledge by simultaneously expanding the extent of human interactions.
- It contributes to the shift of knowledge and power from the individual to the group.
- c factor (general collective intelligence factor) indicates a group's ability to perform a wide range of tasks.