Chapter 7 Web based information system and navigation

Keshav Raj Joshi

- Assignment -2
  - Describe details on:
    1. The structure of the web
    2. Link Analysis

Introduction

- An information system that utilizes Web technologies to deliver information and services to users or other information systems /applications.
- Key features of web-based Information System
  - **Cross platform compatibility** - can use different OS such as Windows, Linux or Mac to run the web applications from browsers
  - **More Manageable** - easier to maintain and update as usually it can all be done on the server
  - **Multiple concurrent users**
  - **Reduced cost** - lower requirements of support and maintenance on the end user system

Searching the web

- Search engine is a software program that searches for sites based on the words entered by user.
- Search engine is the huge database of internet resources that helps to locate information on the World Wide Web
- A search engine consists of two main things: a database of information, and algorithms that compute which results to return and rank for a given query.
- In the case of web search engines like Google, the database consists of trillions of web pages, and the algorithms look at hundreds of factors to deliver the most relevant results.

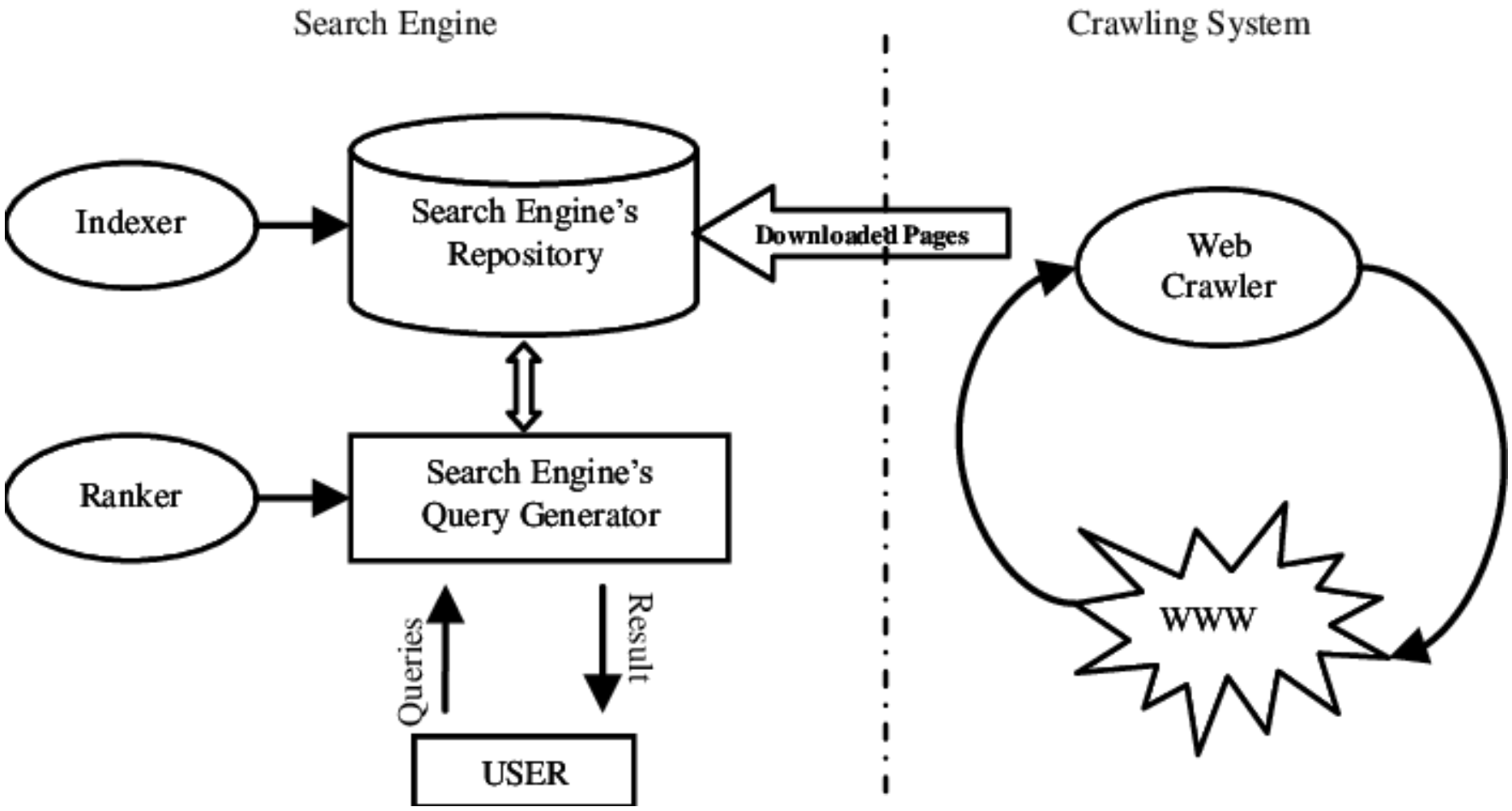
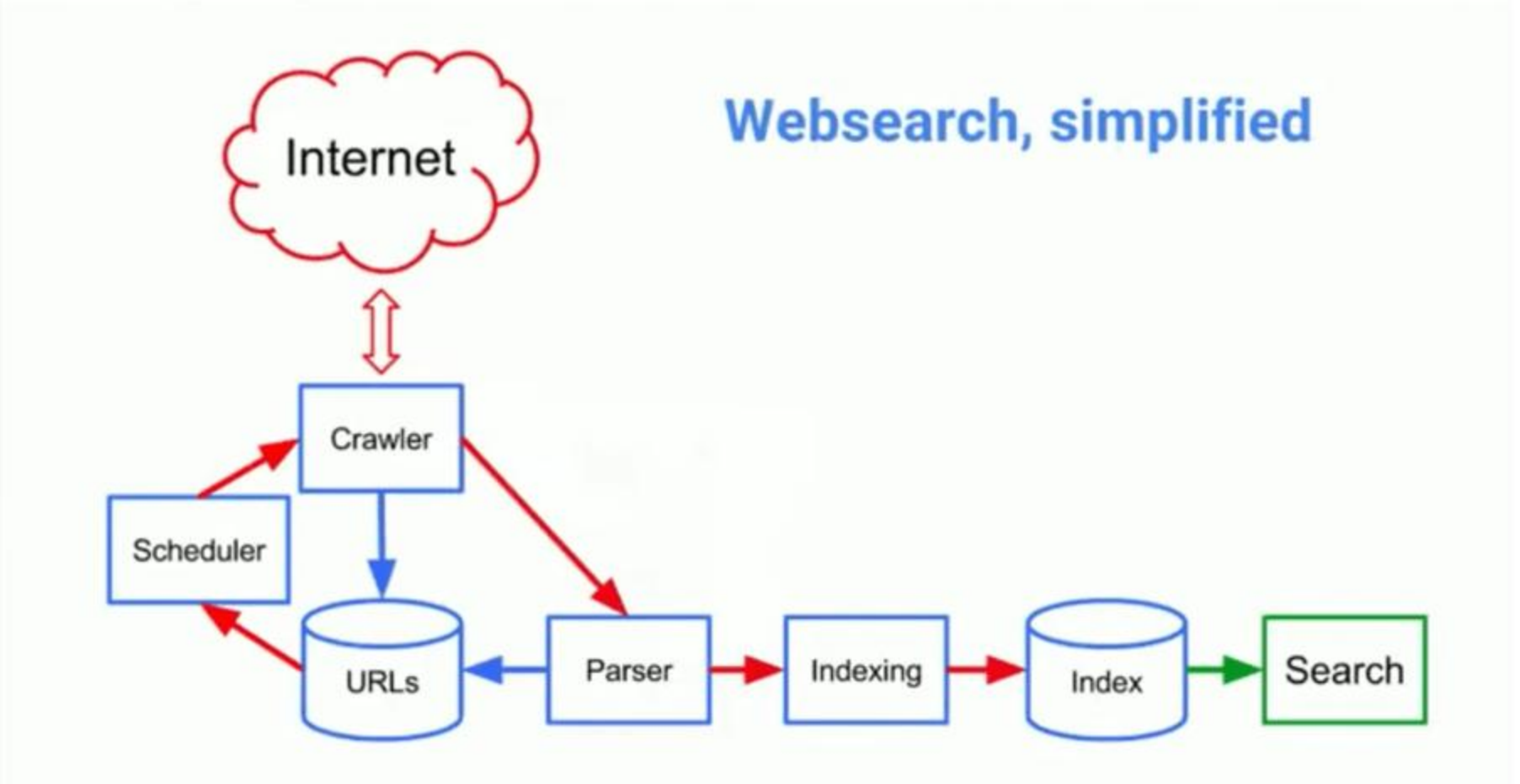


Fig. Functional-block-diagram-of-a-Search-Engine

How do search engines work?

- Search engines work by taking a list of known URLs, which then go to the scheduler. The scheduler decides when to crawl each URL. Crawled pages then go to the parser where vital information is extracted and indexed. Parsed links go to the scheduler, which prioritizes their crawling and re-crawling.



- **Scheduling**
  - The scheduler assesses the relative importance of new and known URLs. It then decides when to crawl new URLs and how often to re-crawl known URLs.
- **Crawling**
  - The crawler is a computer program that downloads web pages. Search engines discover new content by regularly re-crawling known pages where new links often get added over time.
  - When a search engine like Google re-crawls that page, it downloads the content of the page with the recently-added links.
  - The crawler then passes the downloaded web page to the ***parser.***
- **Parsing**
  - The parser extracts *links* from the page, along with other *key information*. It then sends extracted URLs to the *scheduler* and extracted data for *indexing.*
- **Indexing**
  - Indexing is where parsed information from crawled pages gets added to a database called a search index. Think of this as a digital library of information about trillions of web pages
- **Ranking**
  - Providing the pieces of content that will best answer a searcher's query, which means that results are ordered by most relevant to least relevant.

- If you're not showing up anywhere in the search results, there are a few possible reasons why:
  - Your site is brand new and hasn't been crawled yet.
  - Your site isn't linked to from any external websites.
  - Your site's navigation makes it hard for a robot to crawl it effectively.
  - Your site contains some basic code called crawler directives that is blocking search engines.
  - Your site has been penalized by Google for spammy tactics.

Types of Search Engines

1. Crawler-based search engines

- use automated software programs to survey and categorize web pages. The programs used by the search engines to access your web pages are called ‘spiders’, ‘crawlers’, ‘robots’ or ‘bots’.
- A spider will find a web page, download it and analyze the information presented on the web page. The web page will then be added to the search engine’s database. Examples are Google, Bing

2. Human Powered Directories

- Depends on human based activities for category listings.
- Site owner submits a short description of the site to the directory along with category it is to be listed.
- Submitted site is then manually reviewed and added in the appropriate category or rejected for listing.
- Keywords entered in a search box will be matched with the description of the sites. This means the changes made to the content of a web pages are not taken into consideration as it is only the description that matters.
- Examples are Yahoo! Directory and DMOZ

3. Hybrid Search Engines

- Hybrid Search Engines use both crawler based and manual indexing for listing the sites in search results.
- Most of the crawler based search engines like Google basically uses crawlers as a primary mechanism and human powered directories as secondary mechanism.
- For example, Google may take the description of a webpage from human powered directories and show in the search results.

4. Meta Search Engines

- Meta search engines take the results from all the other search engines results, and combine them into one large listing.
- Examples of Meta search engines include Metacrawler, Dogpile etc.

Navigating the web

- **Navigation**: Getting the user from A to B (Where you want them to)
- If visitors can't figure out where to find what they want, they'll leave
- Create clear, hierarchical website navigation that helps your visitors find what they want instantly
- **Navigation affects traffic:** how high you'll rank, how much traffic you'll get from search
- **Navigation affects conversions:** how easy the site is to use, what percentage of visitors convert into leads and customers
- Why Is Navigation Important on a Website?
  - Without website navigation, your visitors can't figure out how to find your blog, your email signup page, your product listings, services, pricing, contact information, or help docs.

- **Role of navigation**

For Users:

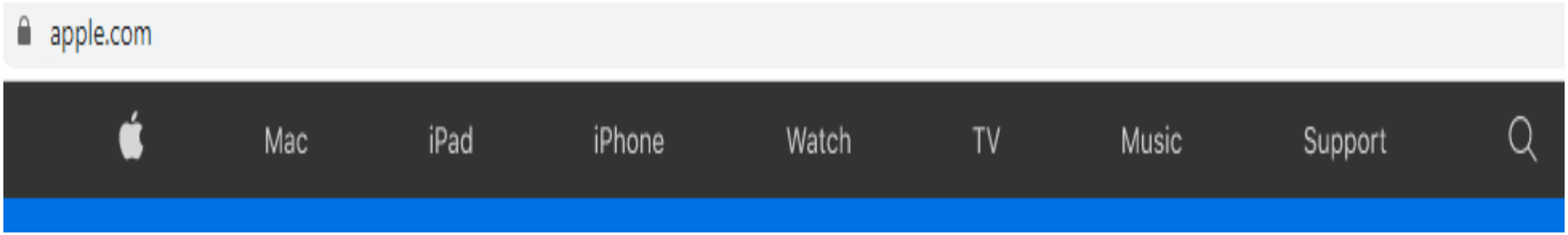
- Find stuff they want
- Get an overview of what's on the site
- See where they are
- See where they can go

For Site owners:

- Drive people to action points
- Cross sell services or highlight additional information
- Show what is/isn't available
- Be found on Google

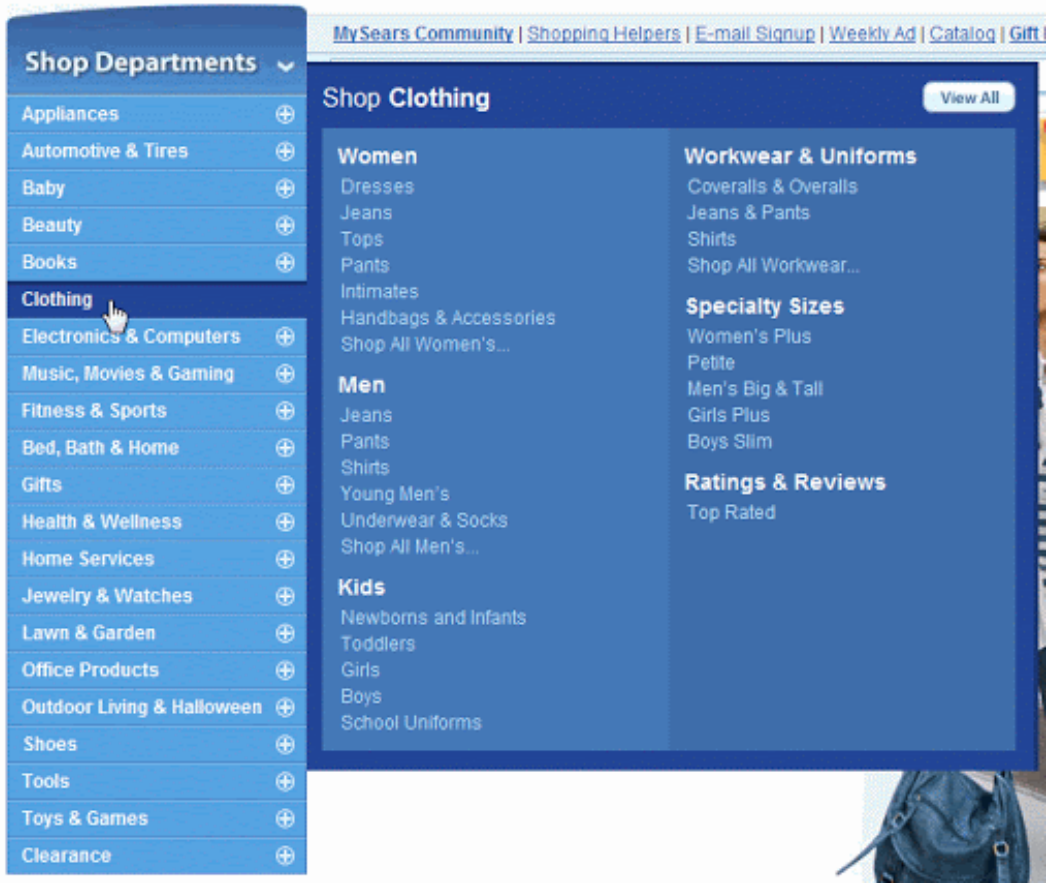
Horizontal Navigation

- Conserves most amount of screen space
- Makes for using drop downs for secondary navigation
- Location most familiar for users, we read left to right
- Does have a limit in terms of length of navigation item



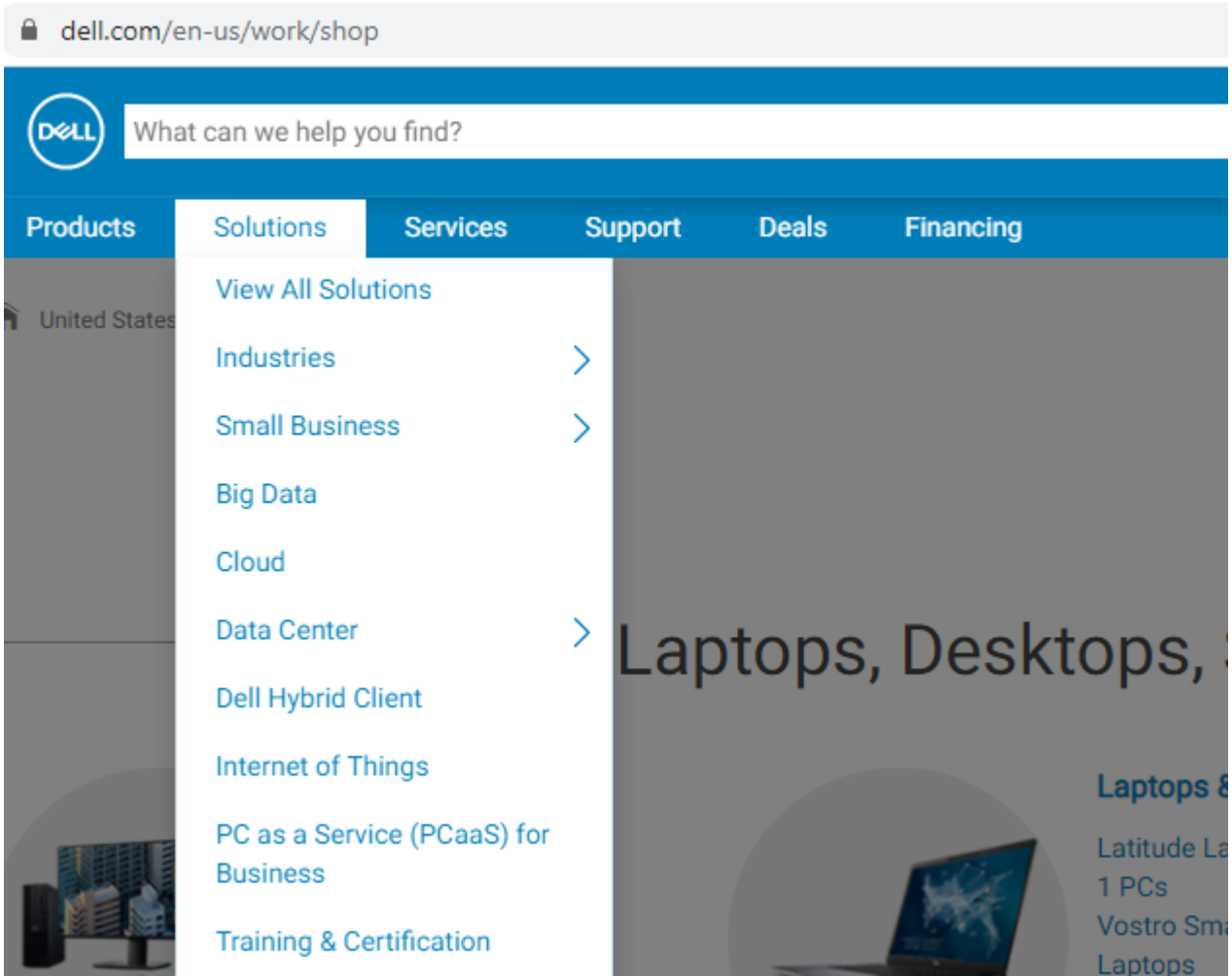
Vertical Navigation

- Good for sites with a lot of main navigation items
- Supports multiple types of secondary navigation
- Can act as a page design element
- Takes up more space than horizontal navigation



Secondary Navigation

- Common location for displaying B-level pages
- Moves out of the way when not active to conserve screen space



The golden rules of web navigation

- Don’t make the user think
- Don’t test user patience
- Focus their attention
- Organize your content
- Use natural descriptions
- Don’t assume
- Avoid dropdown menus

Types of Web Navigation

1. Global Website Navigation
   - With global website navigation, the menu and the links are identical across all pages of the website.
   - 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.
   - Header menus are displayed on every page
2. Hierarchical Website Navigation
   - menus change depending on the context of each page.
   - The structure of the website navigation is built from general to specific.

e.g. if you visit the top page of a newspaper, you will typically see links to the top news categories in the header menu. If the menu were global, it would remain the same after clicking to a different category. If it is hierarchical, it reveals new links that lead to subcategories of the category page we visit. For example if you click on Science page, you see links to different sub-sections of science research and articles

3. Local Website Navigation

- Local website navigation refers to internal links that are included in the content itself. Usually, the user is given options at the same level of a hierarchy or one level deeper, or links to navigate to other relevant pages.
- example is magazine websites, which often use links to help readers explore the deeper context of a certain article.

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

- 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.

2. 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.

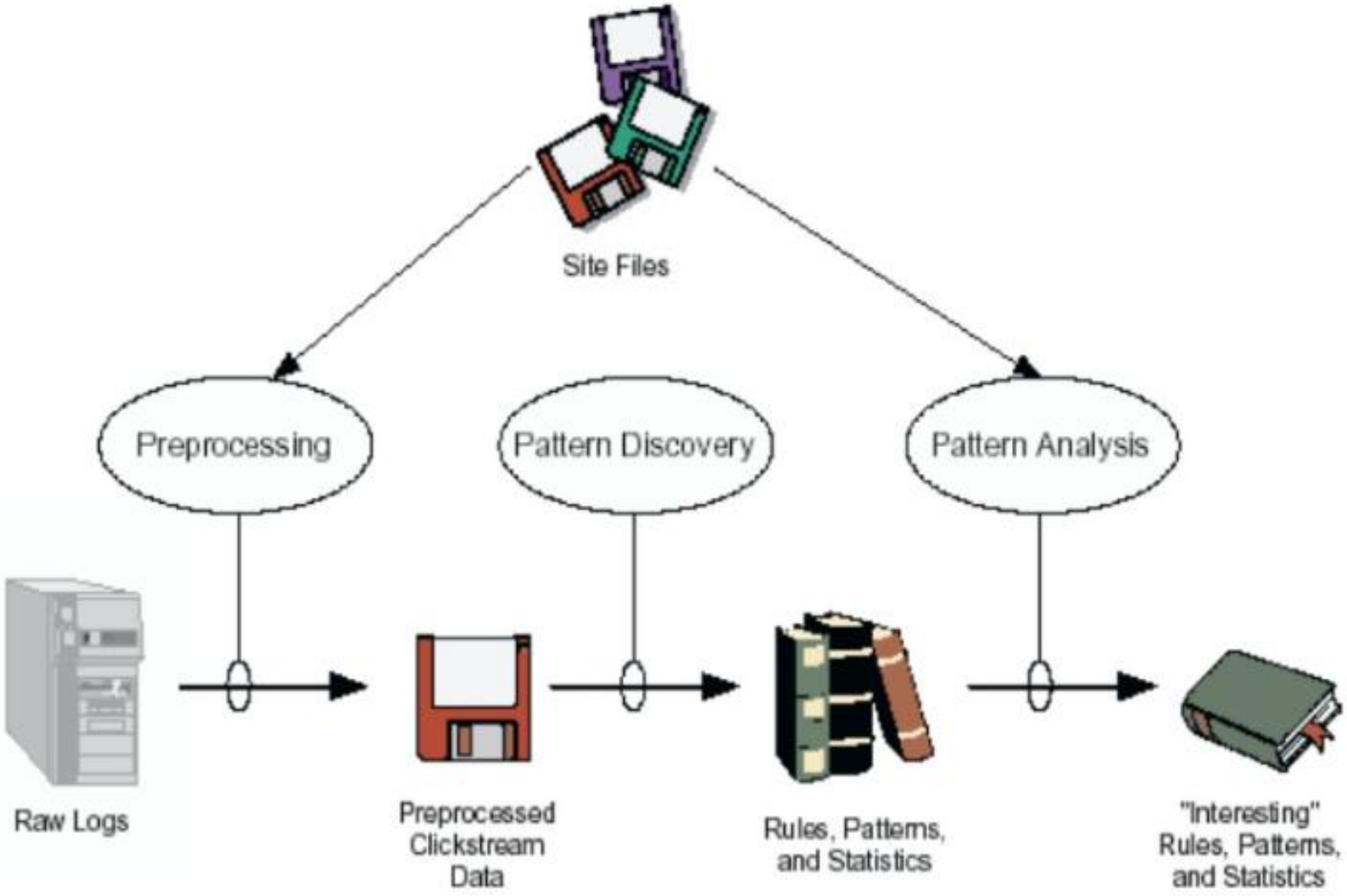
3. Web Usage Mining:

- 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.
- **Goal**
  - To analyze the behavioral patterns and profiles of users interacting with a Web site.

Web usage data sources

- Server access logs
- Server referrer logs (clicking a link from another site)
- Client side cookies
- User profiles
- Search engine logs
- Database logs
- Agent log (client browser info)
- Error log

Web usage mining technique



Preprocessing

- Conversion of the raw data into the data abstraction (users, sessions, episodes, page views) necessary for applying data mining algorithms
  - **Data Cleaning**: remove outliers and/or irrelative data
  - **User Identification**: associate page references with different users
  - **Session Identification**: divide all pages accessed by a user into sessions
  - **Path Completion**: add important page access records that are missing in the access log due to browser and proxy server caching
  - **Formatting**: format the sessions according to the type of data mining to be accomplished.

Pattern discovery

- Key component of web usage mining which converges techniques from data mining, machine learning, statistics and pattern recognition.
- **Statistical Analysis:** frequency analysis, mean, median, etc.
  - Improve system performance
  - Provide support for marketing decisions
  - Simplify site modification task
- **Clustering:**
  - Clustering of users help to discover groups of users with similar navigation patterns → provide personalized Web content
  - Clustering of pages help to discover groups of pages having related content → search engine

- **Classification**: the technique to map a data item into one of several predefined classes
  - Develop profile of users belonging to a particular class or category
- **Association Rules:** discover correlations among pages accessed together by a client
  - Help the restructure of Web site
  - Page perfecting
  - Develop e-commerce marketing strategies

Pattern analysis

- Final stage of WUM which involves validation and interpretation of the mined pattern
  - **Validation**: to eliminate the irrelative rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process
  - **Interpretation**: the output of mining algorithms is mainly in mathematic form and not suitable for direct human interpretations

Recommender systems

- 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 aim to predict users' interests and recommend product items that quite likely are interesting for them.
- Data required for recommender systems:
  - Explicit data:
    - Customer ratings
    - Feedback
    - Demographics/ physiographic
  - Implicit data:
    - Purchase history
    - Click or browse history
  - Product information:
    - Product taxonomy
    - Product description /Product attributes

Why do we need recommender systems?

- Companies using recommender systems focus on increasing sales as a result of very personalized offers and an enhanced customer experience
- typically speed up searches and make it easier for users to access content they're interested in
  - Two-thirds of movies watched by Netflix customers are recommended movies
  - 38% of click-through rates on Google News are recommended links
  - 35% of sales at Amazon arise from recommended products

Recommendation Engine – Examples

Facebook–"People You May Know"

Netflix–"Other Movies You May Enjoy"

LinkedIn–"Jobs You May Be Interested In"

Amazon–"Customer who bought this item also bought …"

YouTube–"Recommended Videos"

Google–"Search results adjusted"

Pinterest–"Recommended Images"

Types of Recommender System:

1. Collaborative Filtering
2. Content based Filtering

1. Collaborative filtering

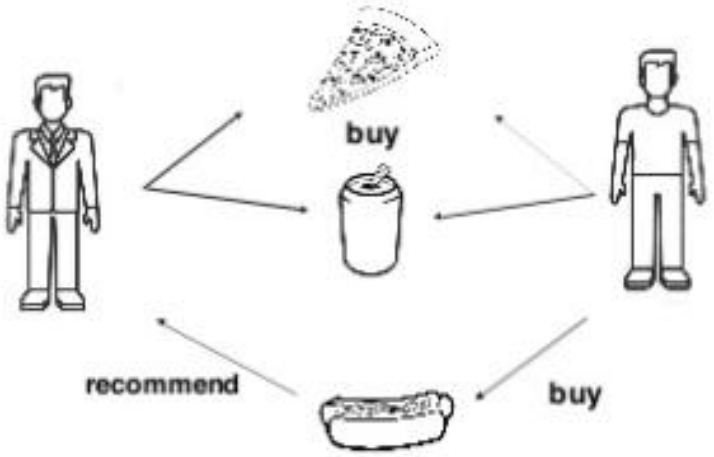
- Collaborative filtering systems are methods that are based solely on the past interactions recorded between **users and items** in order to produce new recommendations. These interactions are stored in the so-called “user-item interactions matrix”
- It does not require understanding of the content 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.

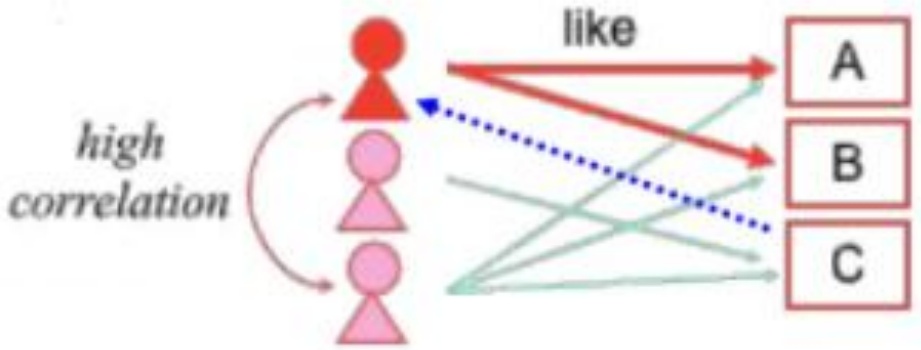
Main approaches

A. User based collaborative filtering

- Use user-item rating matrix
- Make user-to-user correlations
- Find highly correlated users
- Recommend items preferred by those users

- User-based 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.



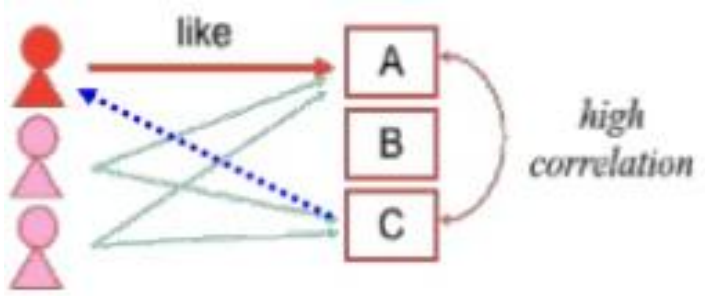


- **Advantage**
  - No knowledge about item features needed
- **Problems**
  - New user cold start problem (The system requires a huge amount of existing data on a user so as to make accurate recommendations. This problem is termed as cold start.)
  - New item cold start problem: items with few rating cannot easily be recommended
  - Sparsity problem: If there are many items to be recommended, user/rating matrix is sparse and it hard to find the users who have rated the same item.(All the users do not rate the items. So, even the most popular items may have few ratings)
  - Popularity Bias: Tend to recommend only popular items.
  - 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.

B. Item based collaborative filtering

- Use user-item ratings matrix
- Make item-to-item correlations
- Find items that are highly correlated
- Recommend items with highest correlation

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



- **Advantages**
  - No knowledge about item features needed
  - Better scalability, because correlations between limited number of items instead of very large number of users
  - Reduced sparsity problem
- **Problems**
  - New user cold start problem
  - New item cold start problem

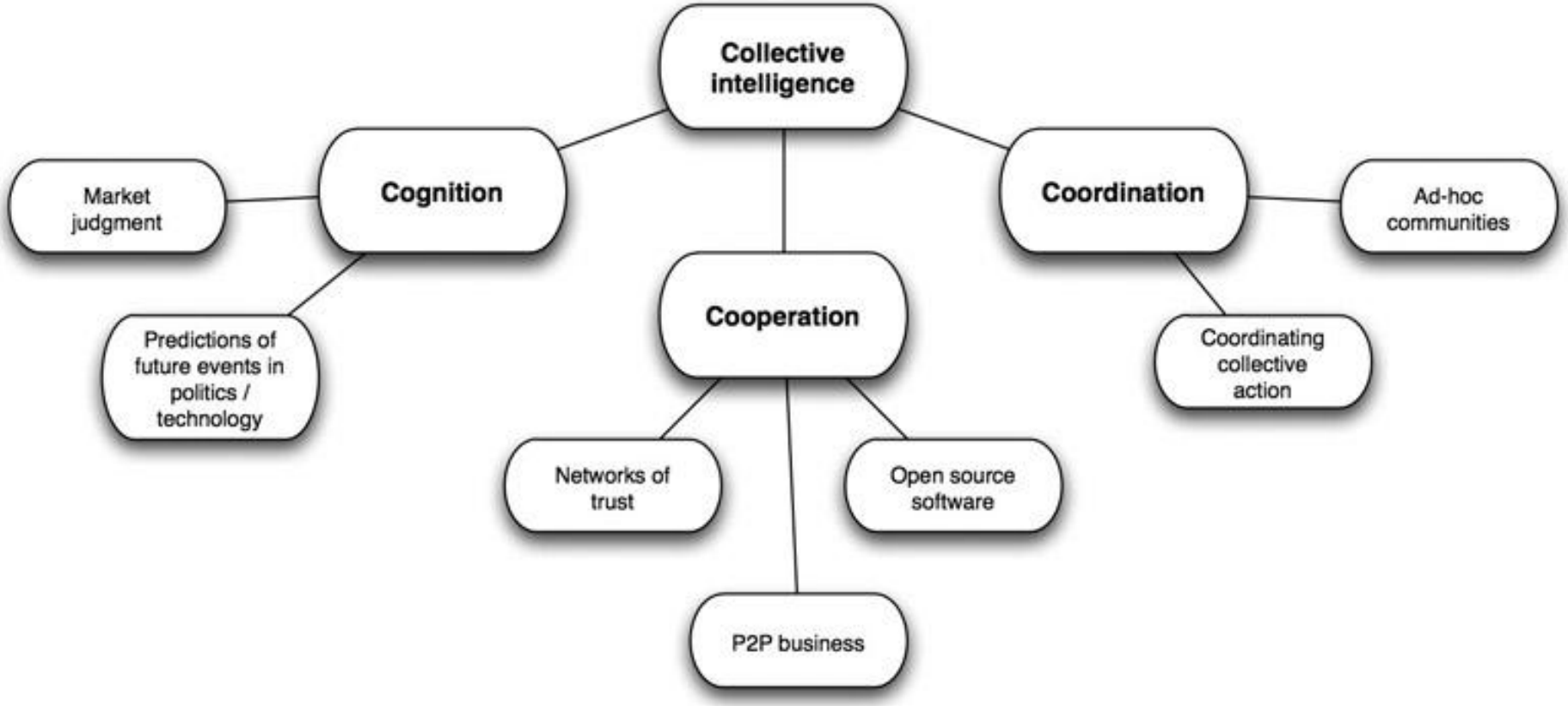
2. Content filtering

- Unlike collaborative methods that only rely on the user-item interactions, content based approaches use additional information about users and/or items.
- 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.
- e.g. in movies recommender system, additional information, for example, the age, the sex, the job or any other personal information for users as well as the category, the main actors, the duration or other characteristics for the movies are taken.

- **Advantages**
  - No need for data on other users
  - No cold start and sparsity
  - Able to recommend users with unique taste
  - Able to recommend new and unpopular items
  - Can provide explanation for recommendation
- **Limitations**
  - Data should be in structured format
  - Unable to use quality judgements from other users.

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.


Collective intelligence
Cognition
Market judgment
Predictions of future events in politics / technology
Cooperation
Networks of trust
Open source software
P2P business
Coordination
Ad-hoc communities
Coordinating collective action

Thank You!