UNIT-V
WEB MINING

Mining the World-Wide Web

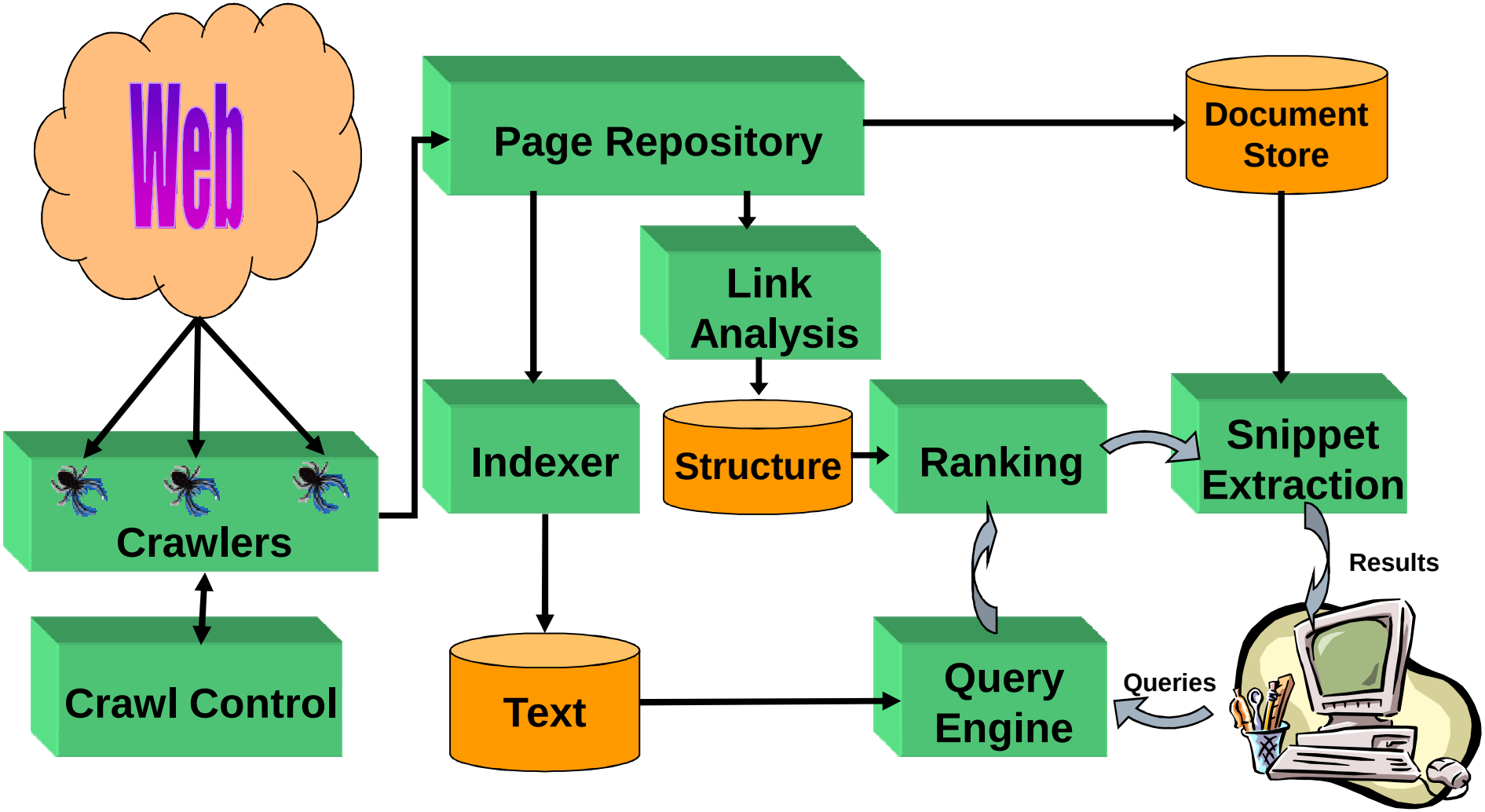
What is Web Mining?

Discovering useful information from the World-Wide Web and its usage patterns.

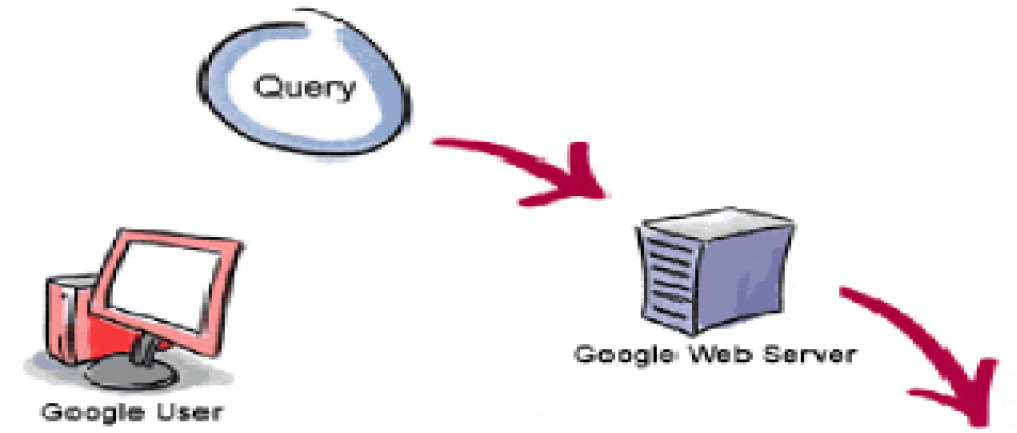
Web search engines

- **Index-based:** search the Web, index Web pages, and build and store huge keyword-based indices
- Help to locate sets of Web pages containing certain keywords
- Deficiencies
  - A topic of any breadth may easily contain hundreds & thousands of documents
  - Many documents that are highly relevant to a topic may not contain keywords defining them

Search Engine Architecture



Working of a search engine:



1. The web server sends the query to the index servers. The content inside the index servers is similar to the index in the back of a book--it tells which pages contain the words that match any particular query term.

2. The query travels to the doc servers, which actually retrieve the stored documents. Snippets are generated to describe each search result.

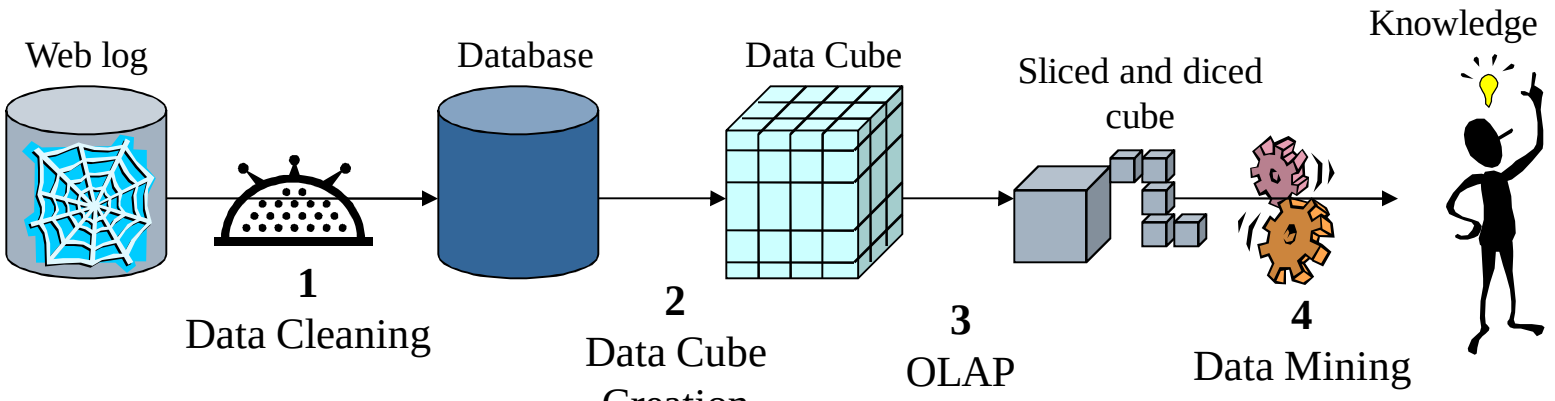
3. The search results are returned to the user in a fraction of a second.

Mining the World-Wide Web

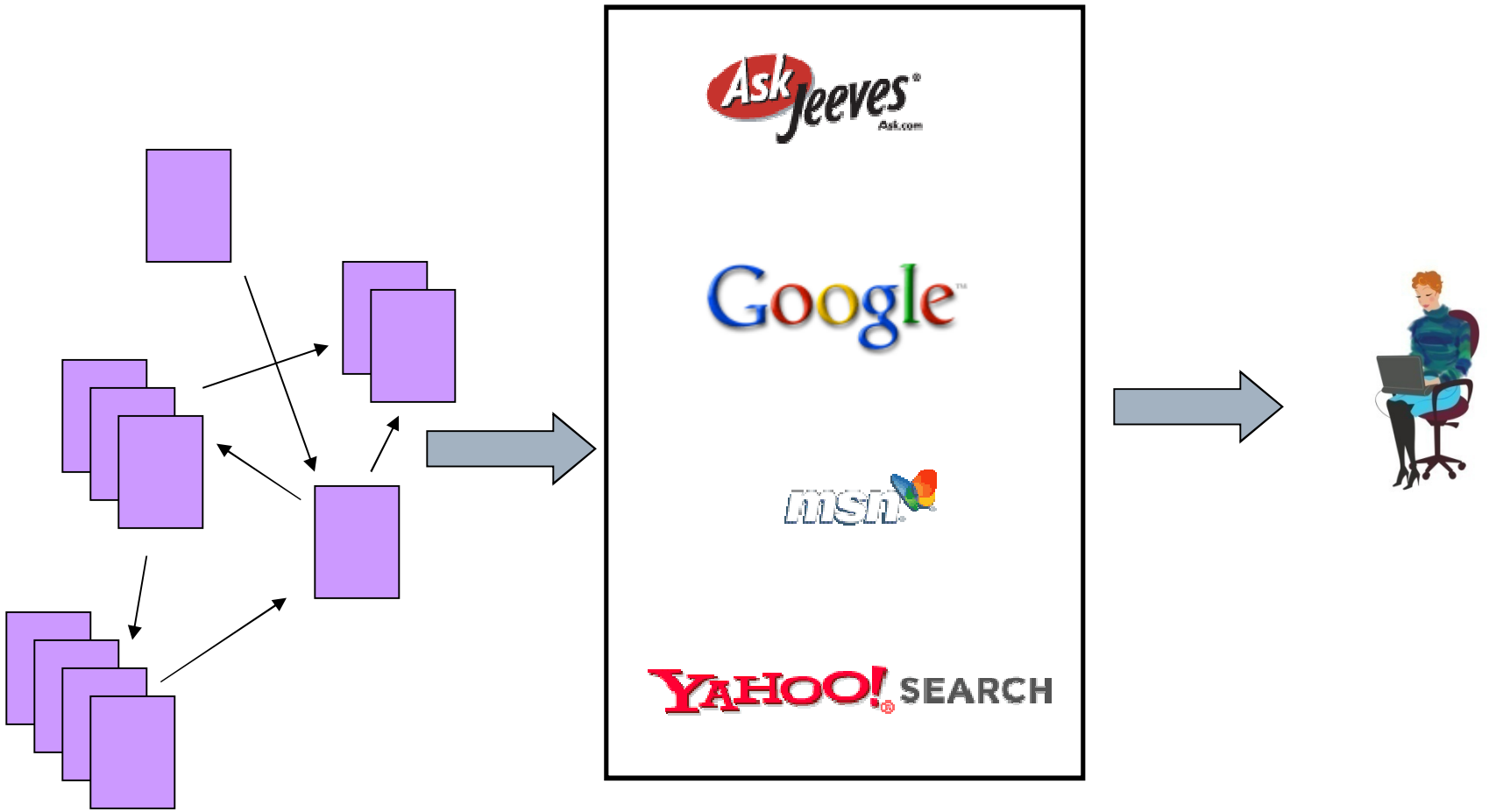
- The WWW is huge, widely distributed, global information service centre for
  - **Information services**: news, advertisements, consumer information, financial management, education, government, e-commerce, etc.
  - **Hyper-link information**
  - **Access and usage information**
- **WWW provides rich sources for data mining**
- Challenges
  - Too huge for effective data warehousing and data mining
  - Too complex and heterogeneous: no standards and structure

Mining the World-Wide Web

- Design of a Web Log Miner
  - Web log is filtered to generate a relational database
  - A data cube is generated form database
  - OLAP is used to drill-down and roll-up in the cube
  - OLAM is used for mining interesting knowledge



Searching the Web



The Web

Content aggregators

Content consumers

Benefits of Web Mining

- Finding relevant information
- Discovering new knowledge from the Web
- Personalized Web page creation
- Learning about individual users

cont…

- We use either Browser or search service to find specific information on the web.
- Specifying keyword query we get response from a web search engine in list of pages on rank basis.

Cont...

- There are three major approaches while assessing information stored on the web:
- **Keyword-based search or topic-**
- Directory browsing with **search engine** such as **Google or Yahoo.**
- **Querying deep web sources-**

  Where information such as amazon.com and realtor.com

Introduction

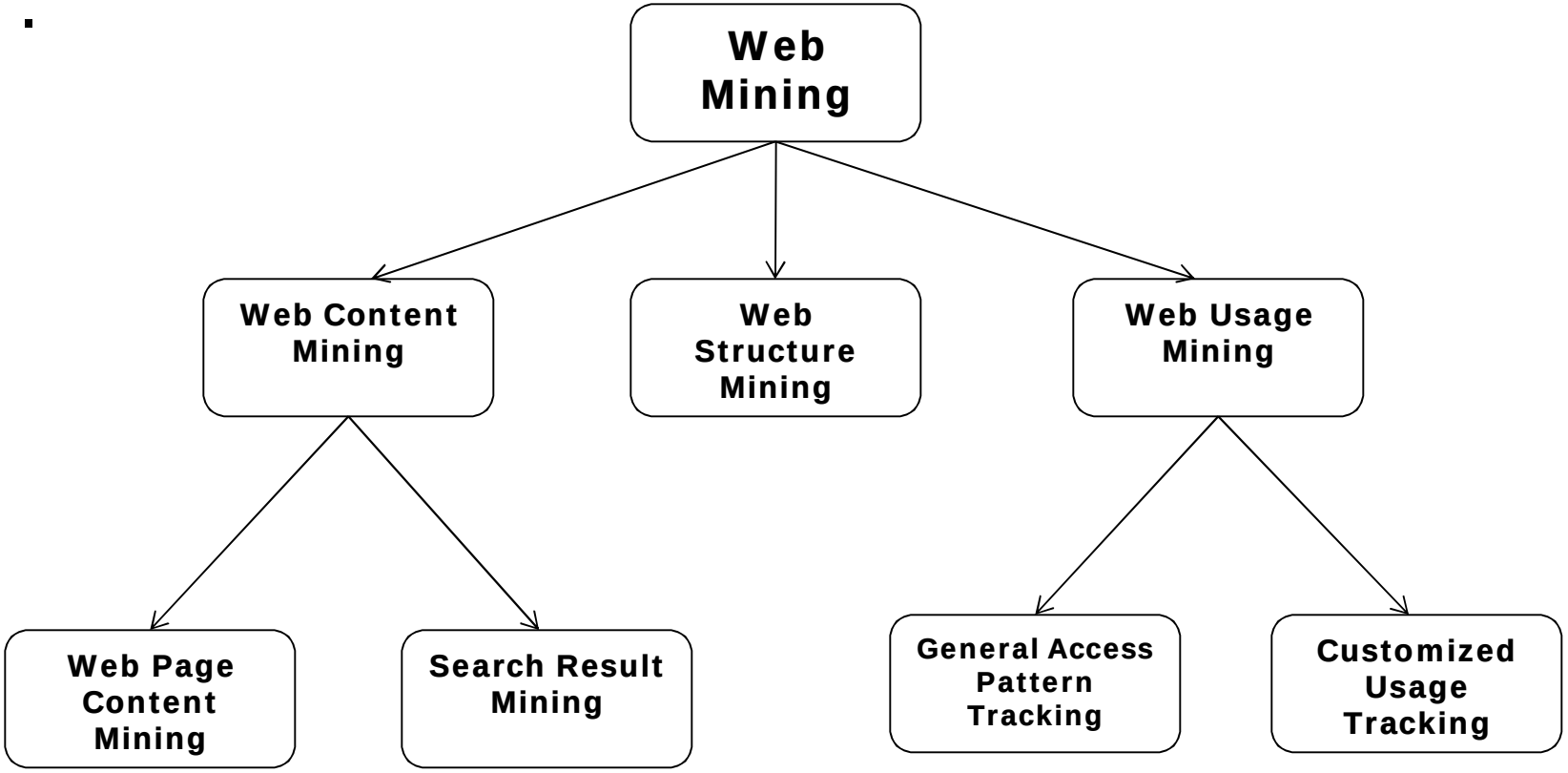
- Web mining techniques provides a set of techniques that can be used to solve the problems of the customers like...
- What the customer do and want.
- Mass customizing the information to the intended customers or
- Personalizing in to individual users
- Problems related to effective web site design and management
- Problem related to marketing etc.

Other related techniques of web mining.....

- Information retrieval (IR)
- Natural language processing (NLP)
- In data mining terms there are three operations of interests.

1. **Clustering** (e.g., finding natural groupings of users, pages etc.)
2. **Associations** (e.g., which URLs tends to be requested together.) and
3. **Sequential analysis** (e.g., the order in which URLs tend to be accessed.

Web Mining Tasks



Web Mining Category

- Mining techniques in the Web is commonly categorized into three areas of interest...

1. Web Content Mining
2. Web Structure Mining
3. Web usage Mining

Cont...

- **Web Content Mining**- Application of data mining techniques to unstructured or semi structured text. Typically HTML documents.
- **Web Structure Mining** – Use of hyperlink structure of the web as an additional information source.
- **Web Usage Mining** – Analysis of user interaction with a web server.

  Usage patterns

  - Number of visitors
  - Popularity e.g., products, movies, music

WEB CONTENT MINING

Web Content Mining

- Web content mining consists of several types of data such as...
- Textual
- Image
- Audio
- Video
- Metadata, as well as
- Hyperlinks.

Cont...

- Recent research on mining multi-types of data is termed as multimedia data mining.
- The textual parts of web content data consists of unstructured data such as free text, semi-structure data such as HTML documents and more structured data such as data in the tables or database-generated HTML.
- Most of the web content data is unstructured, free text data.

Cont...

- As a result, the techniques of text mining can be directly employed for web content mining in such cases.

Web Content Mining

- It describes the discovery of useful information from the web content. Information
- Web contains many kinds of data such as..
- Government information are gradually being placed on the web in recent years.
- Many commercial institutes are transforming their business and services electronically.

Cont...

- Existence of Digital Libraries that are also accessible from web.
- We can not ignore another type of web content-
- The existence of web applications so that the users could access the applications through web interfaces.
- Many applications are being migrated to the web and
- Many types of applications are emerging in the web environment itself.

Web Mining

- **Content:** text & multimedia mining
- **Structure:** link analysis, graph mining
- **Usage:** log analysis, query mining
- Relate all of the above

–Web characterization

– Particular applications

WEB STRUCTURE MINING

Web Structure Mining

- **Web Structure Mining is concerned with discovering the model underlying the link structure of web.**
- It is used to study the topology of the **hyperlinks** with or without the description of the links.
- This model is **used to categorized web pages**.
- It is useful to generate information such as the **similarity and relationship between different web sites**.

Cont...

- Web mining is also **used to discover authority sites** for the subjects and overview (or hub) sites for the subjects that point to many authorities.
- It is seen that Web content mining attempts to explore the structure within a document (intra-document structure).
- Web structure mining studies the structure of documents within the web itself (inter-document structure).

Cont...

- Some algorithms to model web topology such as...

1. HITS
2. PAGE RANK
3. CLEVER

- These models are applied to calculate the quality of rank or relevancy of each web page.

Techniques used in modeling topology

- **Page rank-**
  - In this importance of the document is measured by counting citations or back links to a given document.
  - This gives some approximation of a document's importance or quality.

Cont...

Social Network-

- It is another way of studying the web link structure.
- Web structure mining utilizes the hyperlinks structure of the web to apply social network analysis.
- Social network studies ways to measure the relative standing or importance of individuals in the network.

WEB USAGE MINING

Web usage mining

- **Web usage mining** deals with studying the data generated by the **web surfer's session or behavior**.
- **Web content and structure** mining utilize the **real or primary data** on the web.
- Where as **web usage** mines the **secondary data derived from the interactions** of the users with the web.

Cont...

- The secondary data includes the data from the-
  - Web server access logs
  - Proxy server logs
  - Browser logs
  - User profiles
  - Registration data
  - User sessions or transactions
  - Cookies

Cont...

- User queries
- Bookmark data
- Mouse clicks and scrolls &
- Any other data which are the results of these interactions.

Cont...

- This data can be accumulated by the web server.
- Analysis of the web access logs of different web sites can facilitate an understanding of the user behavior and the web structure .

Size of the Web

- Number of pages
  - Technically, infinite
  - Much duplication (30-40%)
  - Best estimate of “unique” static HTML pages comes from search engine claims
    - Until last year, Google claimed 8 billion(?), Yahoo claimed 20 billion
    - Google recently announced that their index contains 1 trillion pages
      - How to explain the discrepancy?