Web Based Information System
and
Navigation

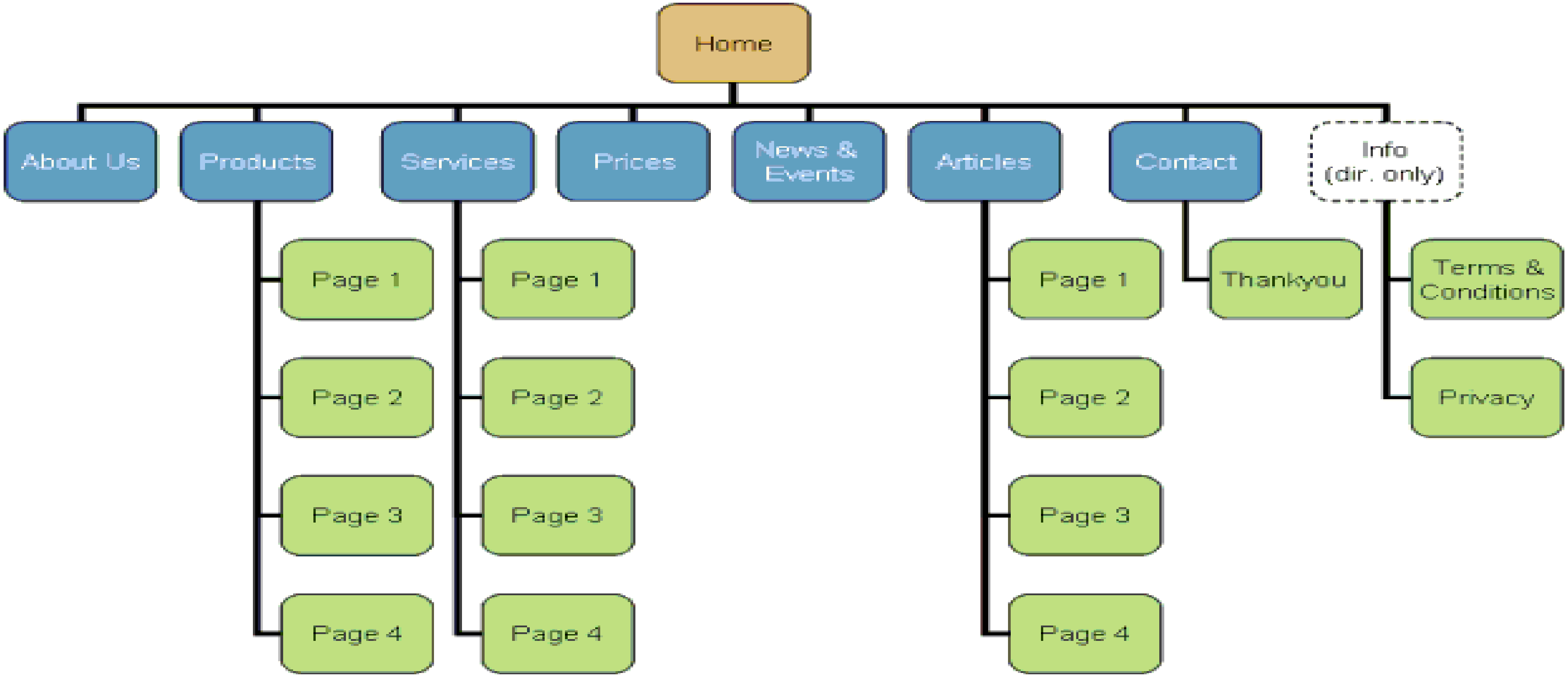
Structure of the Website

- Website structure is the process of defining the look and feel and the navigation of CM *websites*.
- A site structure is normally the role of an information architect, the reality is that everybody from designers to website owners find themselves working on it.
- The website structure consists of three components: Layout Templates, URL patterns, and Linkage Structure.

Layout Template

- Most web pages consist of HTML elements like table, menu, button, image, and input box.
- The layout of a web page describes what HTML elements are included in the page, as well as how these elements are visually distributed in page rendering.
- In a website, pages are generated based on distinguishable templates according to their functions.
- Visually similar pages usually have same function. In this way, user can easily identify a page’s function at a glance.

Layout of Site



- Home
  - About Us
  - Products
    - Page 1
    - Page 2
    - Page 3
    - Page 4
  - Services
    - Page 1
    - Page 2
    - Page 3
    - Page 4
  - Prices
  - News & Events
  - Articles
    - Page 1
    - Page 2
    - Page 3
    - Page 4
  - Contact
    - Thankyou
  - Info (dir. only)
    - Terms & Conditions
    - Privacy

Home	Home Page. *Eg www.domain.com/content/main.html*
About	Top Level Content – Section Main Page. *Eg www.domain.com/content/about/main.html*
Page	Section Content Page. *Eg www.domain.com/content/about/page1.html*

URL Pattern

- A URL pattern is a generalization of a group of URLs sharing similar syntactic format.
- Some example URL patterns discovered, again, from the ASP.NET Forums.
  – List-of-thread pages
    - ^http://forums\.asp\.net/\d+\.aspx$
    - ^http://forums\.asp\.net/\d+\.aspx\?PageIndex=\d+&forumoptions=\d+:\d+:\d+::$
  – List-of-post pages
    - ^http://forums\.asp\.net/t/\d+\.aspx$
    - ^http://forums\.asp\.net/t/\d+\.aspx\?PageIndex=\d+$
    - ^http://forums\.asp\.net/p/\d+/\d+\.aspx$
    - ^http://forums\.asp\.net/ThreadNavigation\.aspx\?PostID=\d+&NavType=(Previous|Next)$
  – User profile pages
    - ^http://forums\.asp\.net/user/Profile\.aspx\?UserID=\d+$
    - ^http://forums\.asp\.net/members/[^/?]*$
- It is noticed that one layout templates can have more than one related URL pattern. For example, a bookseller website usually designs one template to show a list of books, and provides different query parameters to generate such a list.
- Various query parameters in this scenario will lead to different URL patterns, but the search results are shown with the same template.
- Another common case is duplicate pages, i.e., pages with the same content (and very likely the same layout) but different URLs.

Link Structure

- Based on the layout templates and URL patterns, we can construct a directed graph to represent the website organization structure.
- Each layout template is considered as a node in a graph, and two nodes are linked if there are hyperlinks between the pages belonging to the two nodes.
- The link direction is the same as the related hyperlinks. And each link is characterized with the URL pattern of the corresponding hyperlink URLs.
- It should be noticed that there could be multiple links from one node to another if the corresponding hyperlinks have more than one URL pattern.

Web Structure

- WEB STRUCTURE is a pioneer in "fusion engineering"; fusing design sensitivity with cost consciousness to develop the most cost effective structures in which the traditional separation between architectural design and structures is erased in a seamless harmony of design intent.
- In the completed work, architecture and structure resonate to create a single entity; an essential symbiotic interaction, a fusion that presents itself in works of beauty.

Site Structure

- When confronted with a new and complex information system, users build mental model.
- They use these models to assess relations among topics and to guess where to find things they haven't seen before.
- The success of the organization of web site will be determined largely by how well site's information architecture matches users' expectations.
- A logical, consistently named site organization allows users to make successful predictions about where to find things.
- Consistent methods of organizing and displaying information permit users to extend their knowledge from familiar pages to unfamiliar ones.
- If you mislead users with a structure that is neither logical nor predictable, or constantly uses different or ambiguous terms to describe site features, users will be frustrated by the difficulties of getting around and understanding what you have to offer.
- Don't want user's mental model of web site to look like figure1.

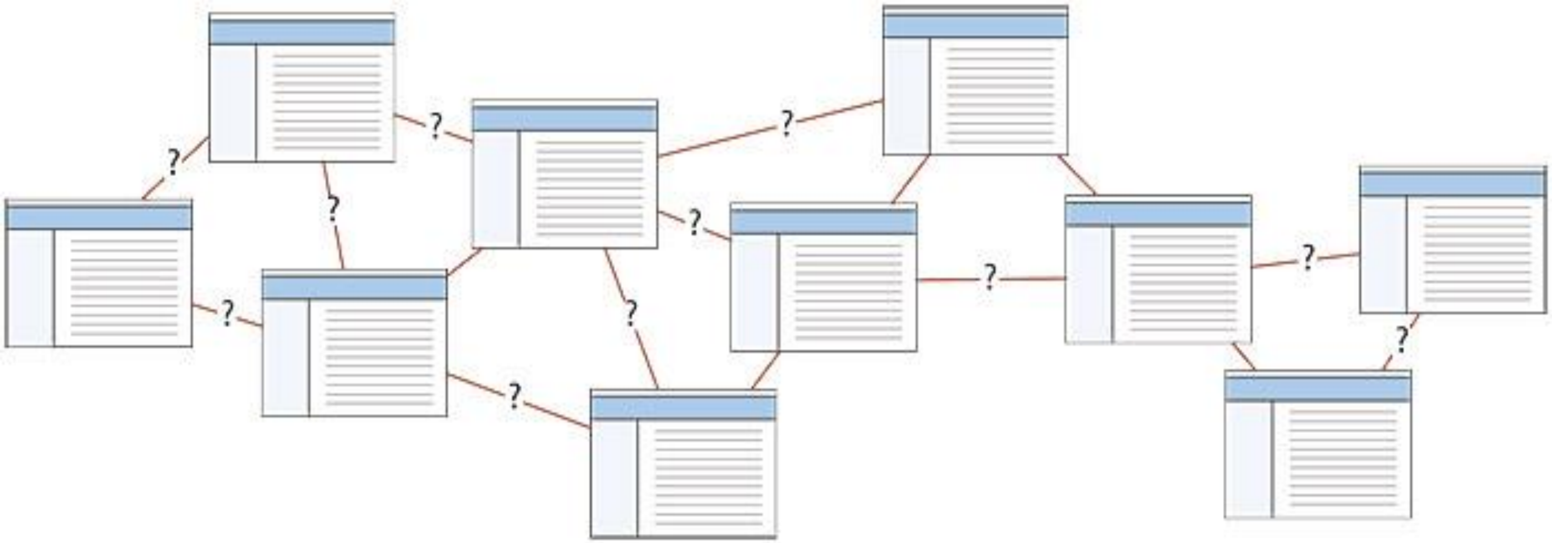


Figure 1 — Don’t make a confusing web of links. Designers aren’t the only ones who make models of sites. Users try to imagine the site structure as well, and a successful information architecture will help the user build a firm and predictable mental model of site.

Web Page Structure

- A web page constructed using HTML has a basic and essential structure. The page always begins with the **start tag** of the html element and always terminates with the **end tag** of the html element as follows:

  Example 1

  ```
  <html>
  ...web page...
  </html>
  ```

- All other element tags are 'nested' within the start and end html tags. The web page is then further subdivided into two main sections which are the 'head' and the 'body'.

Web Page Structure

- The head section begins with the <head> start tag and terminates with the</head> end tag.
- Immediately following this comes the <body> start tag and just before the html end tag comes the </body> end tag.
- There is only *one set* of <html>…</html> tags,
- *one set* of <head>…</head> tags and
- *one set* of <body>…</body> tags.

How do Web Pages Change

- Most pages do not change much.
- Larger pages change more often.
- Commercial pages change more often.
- Past change to a web page is a good indicator of future change.
- About 30% of pages are very similar to other pages, and being a near-duplicate is fairly stable.

Link Analysis

- Link analysis is a data-analysis technique used to evaluate connections between nodes. Relationships may be identified among various types of nodes (objects), including organization, people and transactions.
- The analysis of hyperlinks and the graph structure of the Web has been instrumental in the development of web search.
- Link analysis for web search has intellectual antecedents in the field of citation analysis, aspects of which overlap with an area known as Bibliometrics.

Link Analysis

- Link analysis is an important part of site assessment, either your own or competitor’s.
- ***Outbound links*** are links on your site which refer to other sites, they go beyond the borders of your site.
- ***Internal links*** are links which point to another page of your site, i.e. they refer to some place within your site.

Website Navigation

Website navigation is important to the success of website visitor's experience to website.

The website's navigation system is like a road map to all the different areas and information contained within the website.

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 links with the text of your web pages, linking to other pages within the website.

Website Navigation Use

- To be consistent throughout the website. The website visitors will learn, through repetition, how to get around the website.
- The main navigation links kept together. This makes it easier for the visitor to get to the main areas of the website.
- Reduced clutter by grouping links into sections. If the list of website navigation links are grouped into sections and each section has only 5-7 links, this will make it easier to read the navigation scheme.
- Minimal clicking to get to where the visitor wants to get to. If the number of clicks to the web page the visitor wishes to visit is minimal, this leads to a better experience.

Website Navigation Use

- Some visitors can become confused or impatient when clicking a bunch of links to get to where they want to be.
- In large websites, this can be difficult to reduce. Using breadcrumbs is one way to help the visitor see where they are within the website and the path back up the navigation path they took.
- Creating the website navigation system at the planning stage of the website will effect the overall design of the web page layout and help develop the overall plan for the website.

Web Mining

- Web mining can be broadly defined as discovery and analysis of useful information from the World Wide Web.
- Based on the different emphasis and different ways to obtain information, web mining can be divided into two major parts: ***Web Contents Mining*** and ***Web Usage Mining***.
- **Web Contents Mining** can be described as the automatic search and retrieval of information and resources available from millions of websites and on-line databases though search engines / web spiders.
- **Web Usage Mining** can be described as the discovery and analysis of user access patterns, through the mining of log files and associated data from a particular Web site.

Web Usage Mining

- The automatic discovery of patterns in clickstreams and associated data collected or generated as a result of user interactions with one or more Web sites.

Goals

- To analyze the behavioral patterns and profiles of users interacting with a Web site.
- The discovered patterns are usually represented as collections of pages, objects, or resources that are frequently accessed by groups of users with common interests.

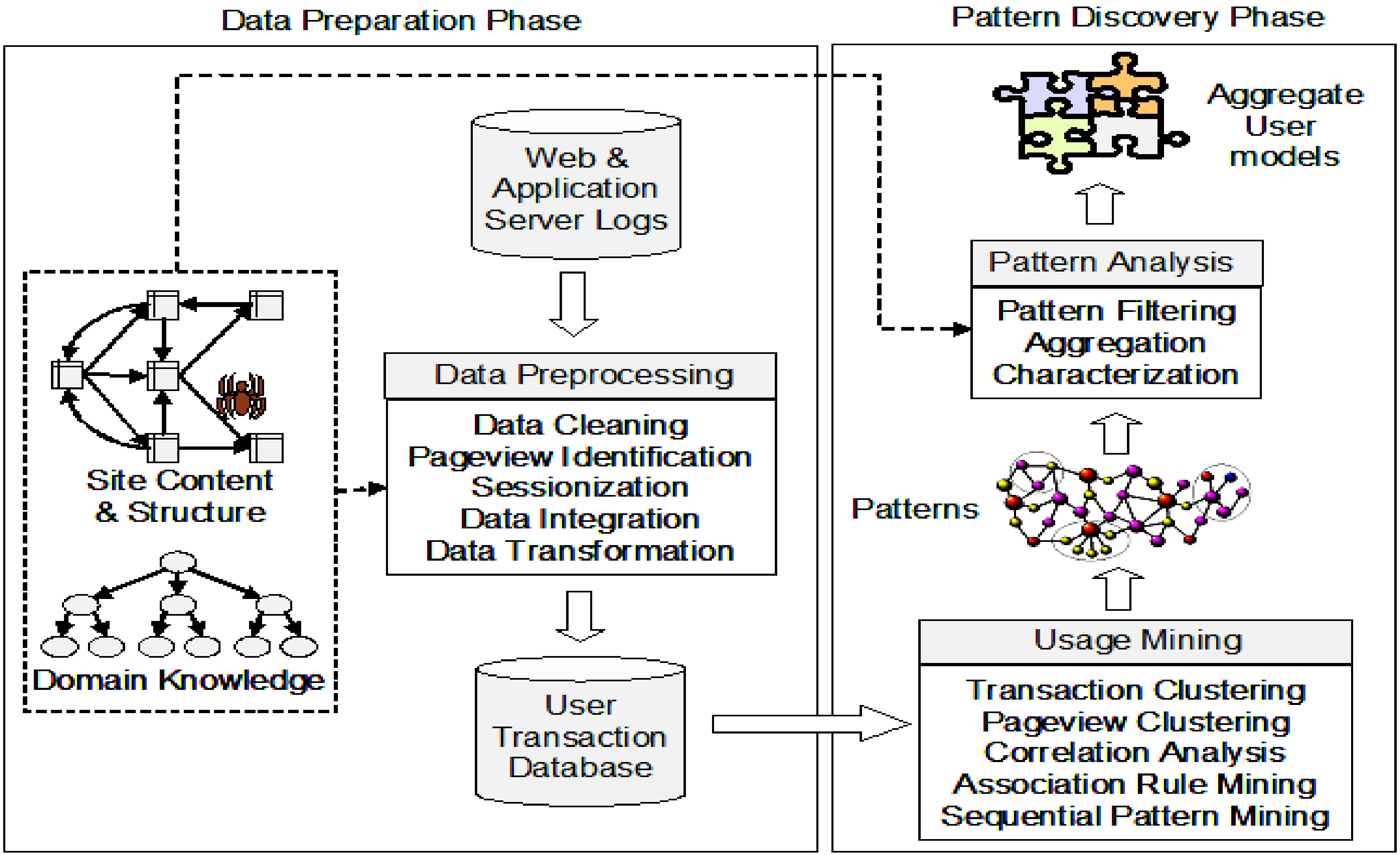
How to perform Web Usage Mining?

- Web usage mining is achieved first by reporting visitors traffic information based on Web server log files and other source of traffic data.
- Web server log files were used initially by the webmasters and system administrators for the purposes of “how much traffic they are getting, how many requests fail, and what kind of errors are being generated”, etc. **However, Web server log files can also record and trace the visitors’ on-line behaviors.**
- Web log file is one way to collect Web traffic data. The other way is to “sniff” TCP/IP packets as they cross the network, and to “plug in” to each Web server.

Data in Web Usage Mining

- Data in Web Usage Mining:
  - Web server logs
  - Site contents
  - Data about the visitors, gathered from external channels
  - Further application data
- Not all these data are always available.
- When they are, they must be integrated.
- A large part of Web usage mining is about processing usage/ clickstreams data.
  - After that various data mining algorithm can be applied.

Web usage mining process



Data Preparation Phase

Web & Application Server Logs

Site Content & Structure

Domain Knowledge

Data Preprocessing

Data Cleaning
Pageview Identification
Sessionization
Data Integration
Data Transformation

User Transaction Database

Pattern Discovery Phase

Usage Mining

Transaction Clustering
Pageview Clustering
Correlation Analysis
Association Rule Mining
Sequential Pattern Mining

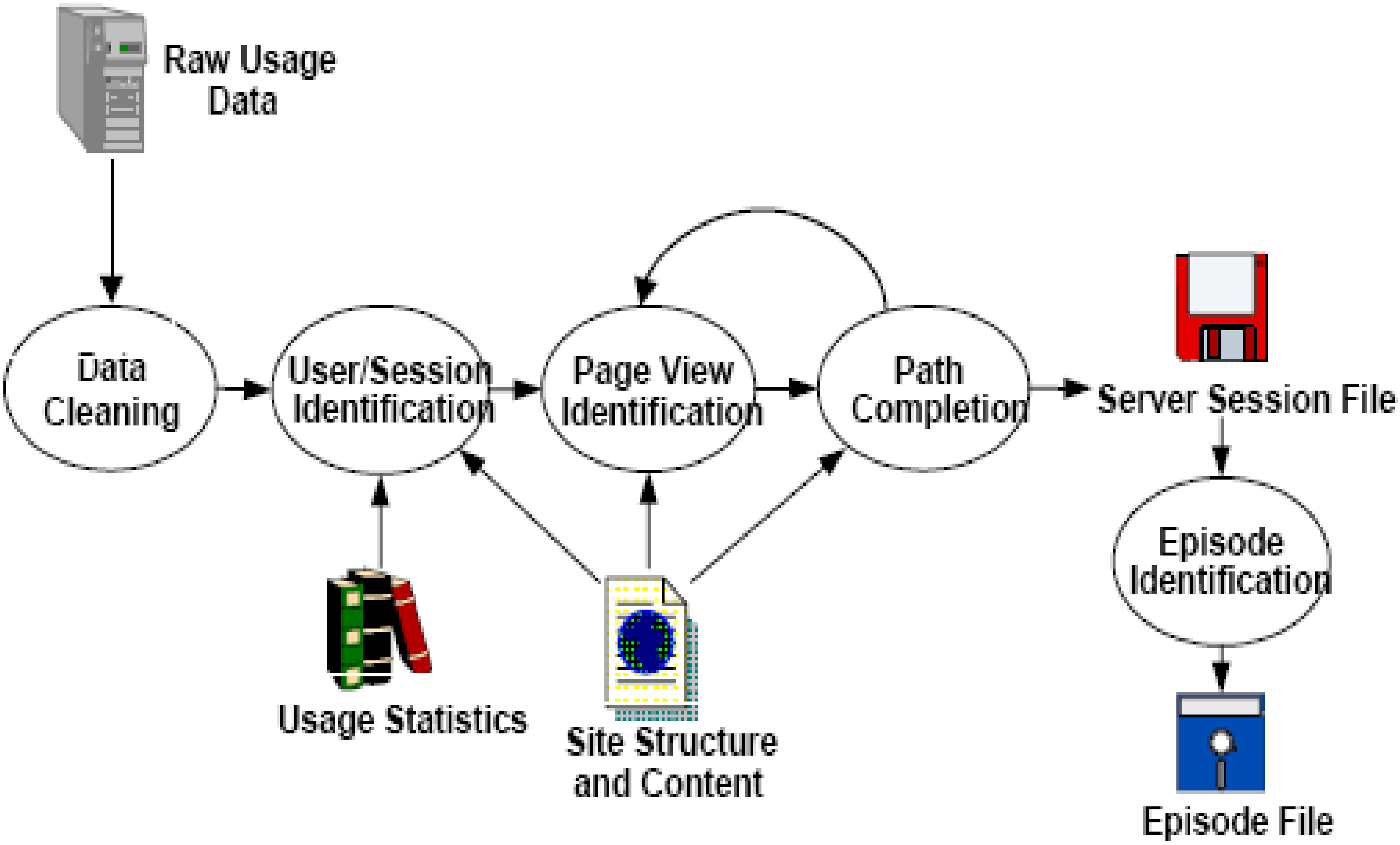
Patterns

Pattern Analysis

Pattern Filtering
Aggregation
Characterization

Aggregate User models

Pre-processing of web usage data



Data Cleaning

- Data cleaning
  - remove irrelevant references and fields in server logs
  - remove references due to spider navigation
  - remove erroneous references
  - add missing references due to caching (done after sessionization)

Identify sessions (Sessionization)

- In Web usage analysis, these data are the sessions of the site visitors: the activities performed by a user from the moment she enters the site until the moment she leaves it.
- Difficult to obtain reliable usage data due to proxy servers and dynamic IP addresses, missing references due to caching, and the inability of servers to distinguish among different visits.

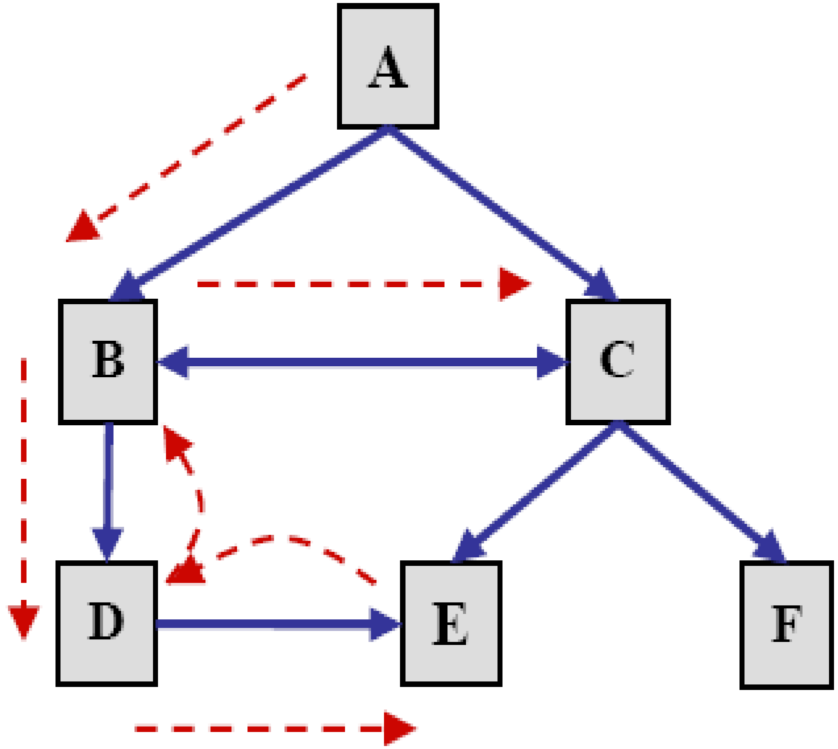
Page View

- A pageview is an aggregate representation of a collection of web objects contributing to the display on a user's browser resulting from a single user action (such as a click-through).
- Conceptually, each pageview can be viewed as a collection of web objects or resources representing a specific "user event" e.g., reading an article, viewing a product page, or adding a product to the shopping cart.

Path Completion

- Client- or proxy-side caching can often result in missing access references to those pages or objects that have been cached.
- For instance,
  - if a user returns to a page A during the same session, the second access to A will likely result in viewing the previously downloaded version of A that was cached on the client-side, and therefore, no request is made to the server.
  - This results in the second reference to A not being recorded on the server logs.

Missing references due to caching



User's actual navigation path:

A →B → D → E → D → B → C

What the server log shows:

URL	Referrer
A	--
B	A
D	B
E	D
C	B

Fig. 12.7. Missing references due to caching.

Contd….

- The problem of inferring missing user references due to caching.
- Effective path completion requires extensive knowledge of the link structure within the site
- Referrer information in server logs can also be used in disambiguating the inferred paths.
- Problem gets much more complicated in frame-based sites.

Collaborating Filtering

- Collaborative filtering(CF) is the process of filtering or evaluating items through the opinions of other people.
- CF technology brings together the opinions of large interconnected communities on the web, supporting filtering of substantial quantities of data.
- Collaborative filtering systems produce predictions or recommendations for a given user and one or more items. Items can consist of anything for which a human can provide a rating, such as art, books, CDs, journal articles, or vacation destinations.

Collaborating Filtering

Ratings in a collaborative filtering system can take on a variety of forms.

- Scalar ratings can consist of either numerical ratings, such as the 1-5 stars provided in ordinal ratings such as strongly agree, agree, neutral, disagree, strongly disagree.
- Binary ratings model choices between agree/disagree or good/bad.
- Unary ratings can indicate that a user has observed or purchased an item, or otherwise rated the item positively.

The absence of a rating indicates that we have no information relating the user to the item (perhaps they purchased the item somewhere else).

Collaborative Filtering

Match people with similar interests as a basis for recommendation.

1) Many people must participate to make it likely that a person with similar interests will be found.
2) There must be a simple way for people to express their interests.
3) There must be an efficient algorithm to match people with similar interests.

How does CF Work?

- Users rate items – user interests recorded.
- Ratings may be:
  - Explicit, e.g. buying or rating an item
  - Implicit, e.g. browsing time, no. of mouse clicks
- Nearest neighbour matching used to find people with similar interests
- Items that neighbours rate highly but that you have not rated are recommended to you
- User can then rate recommended items

Example of CF MxN Matrix with M users and N items (An empty cell is an unrated item)

Items / Users	Data Mining	Search Engines	Data Bases	XML
Alex	1		5	4
George	2	3	4	
Mark	4	5		2
Peter			4	5

Observations

- Can construct a vector for each user (where 0 implies an item is unrated)
  - E.g. for Alex: <1,0,5,4>
  - E.g. for Peter <0,0,4,5>
- On average, user vectors are sparse, since users rate (or buy) only a few items.
- Vector similarity or correlation can be used to find nearest neighbour.
  - E.g. Alex closest to Peter, then to George.

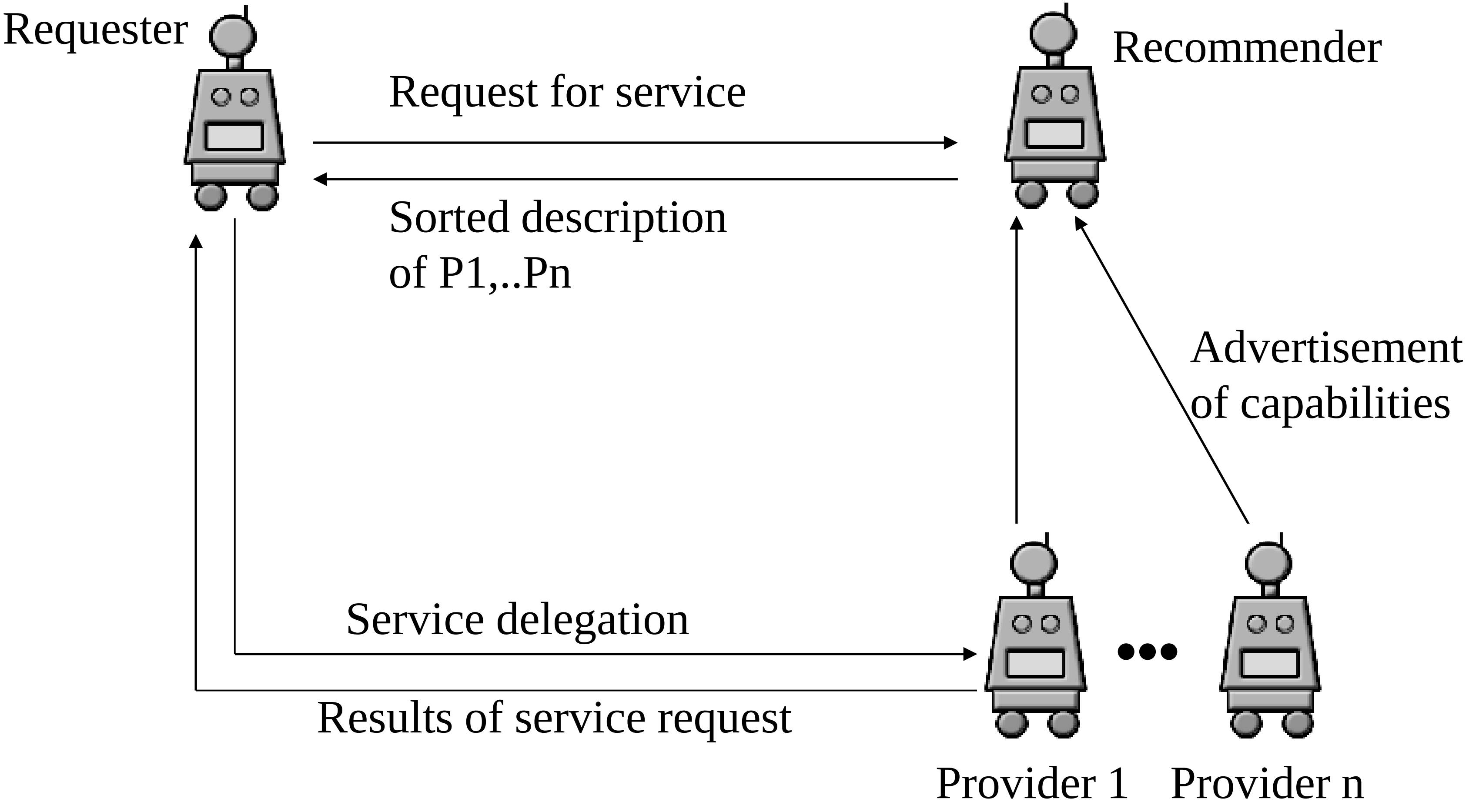
Case Study – Amazon.com

- Item-to-item collaborative filtering
  - Find similar items rather than similar customers.
- Record pairs of items bought by the same customer and their similarity.
  - This computation is done offline for all items.
- Use this information to recommend similar or popular books bought by others.
  - This computation is fast and done online.

Recommender systems

- Too much information: information overload – consumers have too many options
- A recommender system is a system which provides recommendations to a user
- Applications: Books, music CDs, movies. Even documents, services and other products such as software games

Recommender



Information needed

Information used for recommendations can come from different sources:

- browsing and searching data
- purchase data
- feedback explicitly provided by the users
- textual comments
- expert recommendations
- demographic data

Providing recommendations

Recommendations can take the following forms:

- Attribute-based recommendations: based on syntactic attributes of products (e.g. *science fiction* books)
- Item-to-item correlation (as in shopping basket recommendations)
- User-to-user correlation (finding users with similar tastes)
- Non-personalized recommendations (as in traditional stores, i.e. dish of the day, generic book recommendations etc.)

Recommender systems in e-commerce

- Turning browsers into customers: they can stimulate the users’ needs (need identification stage)
- Cross-selling: suggest additional products which may match the user’s interests or current shopping basket
- Personalization: personalized services, or the site can be personalized to the user’s liking – unique shopping experience
- Keeping customers informed
- Retaining customer loyalty

Collective Intelligence

- A shared or group intelligence that emerges from the collaboration and competition of many individuals.
- Groups of people and computers, connected by the Internet, collectively doing intelligent things. For example, Google technology harvests knowledge generated by millions of people creating and linking web pages and then uses this knowledge to answer queries in ways that often seem amazingly intelligent.
- In Wikipedia, thousands of people around the world have collectively created a very large and high quality intellectual product with almost no centralized control, and almost all as volunteers!

Examples

- One example of collective intelligence would be political parties and the way in which the take the views of the people to form policies, select their candidates and run election campaigns.
- Online multi-player games are another example of collective intelligence. Games such as Halo, Second Life and Call of Duty rely on gamers coming together as a community to form the game’s Identity.

Examples

- The online encyclopaedia Wikipedia is one of the best examples of collective intelligence. Anyone can add information to an exiting page or indeed create a new page of information; pages also hyperlink to other areas of the website that people have edited.
- Google is a prominent example of collective intelligence. The search engine is made up of millions of websites, which have been created by people all over the world.

Examples

- The social networking world is perhaps the most popular of collective intelligence. Friend post statuses which then act as newsfeed, which informs other friends of their thoughts. Friends can also recommend other friends, applications and pages to any person on their friend list.

Example

- If a person has a Amazon account they can buy or sell products to other people with accounts this is collective intelligence because the people are making up the website.
- The website also recommends items that may also interest you judging on what you have already looked at which is collective intelligence also.
- Things such as customer reviews can also be heavily influential when choosing a product. You are essentially basing your opinion off of the opinions of other members of the public.

Thank You