Chapter -7

Web based information system and navigation

Information System *(CT 751)* **BCT IV/II**

By: Shayak Raj Giri

Outline

- **Web based information system and navigation**
  - The structure of the web
  - Link Analysis
  - Searching the web
  - Navigating the web
  - Web uses mining
  - Collaborative filtering
  - Recommender systems
  - Collective intelligence

Before you begin

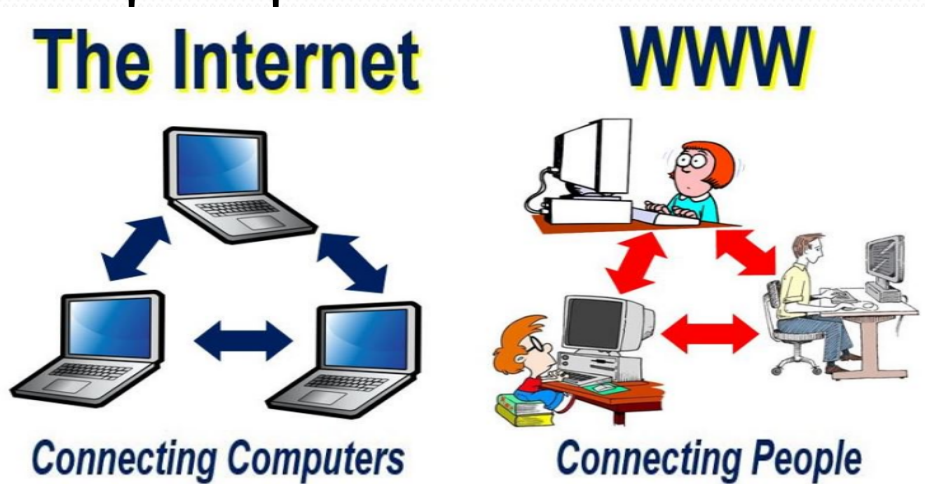
- How many people use the web?
- What is the size of the web?
- How many web sites are there?
- How many searches per day?
- How do web pages change?
- What is the graph structure of the web?
- How could the structure arise?
- What can we do with link analysis?

Web vs. Internet

Web vs. Internet

- The World Wide Web, or simply Web, is a way of accessing information over the Internet.
- It is an information-sharing model that is built on top of the Internet.
- The Web uses the HTTP protocol.
- The Web utilizes browsers to access Web documents called Web pages that are linked to each other via hyperlinks.
- The Internet is a massive network of networks, infrastructure, uses Internet Protocol.
- The **Internet** itself is a global,

  interconnected network of computing devices

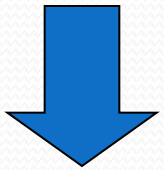


Web vs. Internet

- The World Wide Web (WWW) is one of many services that run on the Internet.
- Services that run on the Internet that are not the WWW include:
  - Email (especially those using the IMAP, POP or SMTP protocols.
  - DNS - Domain Name System - the ability for computers to reference each other by a name rather than a numeric address.
  - FTP - File Transfer Protocol - transferring of files across the Internet.
  - NTP - Network Time Protocol - a way for computers to synchronize their internal clocks over the Internet.

WWW (Web)… is

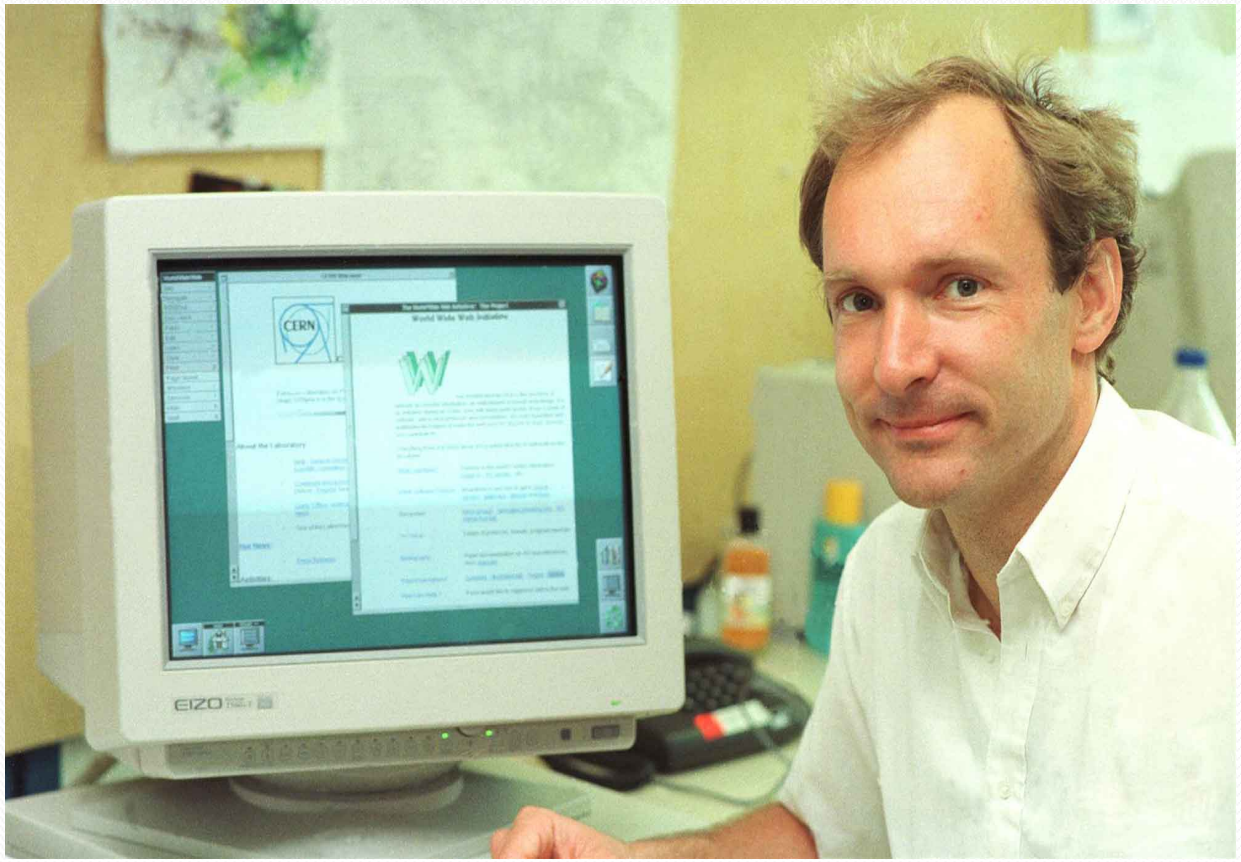
A large-scale, on-line repository of information that users can search using interactive application program called a browser



Interactive program that permits a user to view multimedia information as a Web document, *including* hyperlinks *to other Web documents*

Back to the web

- Created by Tim Burners-Lee
- A research project in 1989-1991 at CERN
- An application of the Internet.
- Two basic features:
  - Make documents on your computer publically accessible
  - Easily access these documents using a browser



The Web

- Basic components that make up the Web
  - **Web Documents - HTML (HyperText Markup Language)**
  - **URLs and Hyperlinks**
  - **HTTP (Hypertext Transfer Protocol)**
  - **Web Servers**
  - **Web Browser**

- The term "**hyperlink**" was coined by Ted Nelson and his assistant Calvin Curtin at the start of *Project Xanadu*

- A Global network of web documents
  - The World Wide Web (WWW)

- **Who defines the Web standards?**
  - **The Web standards are defined by the World Wide Web Consortium (W3C)**

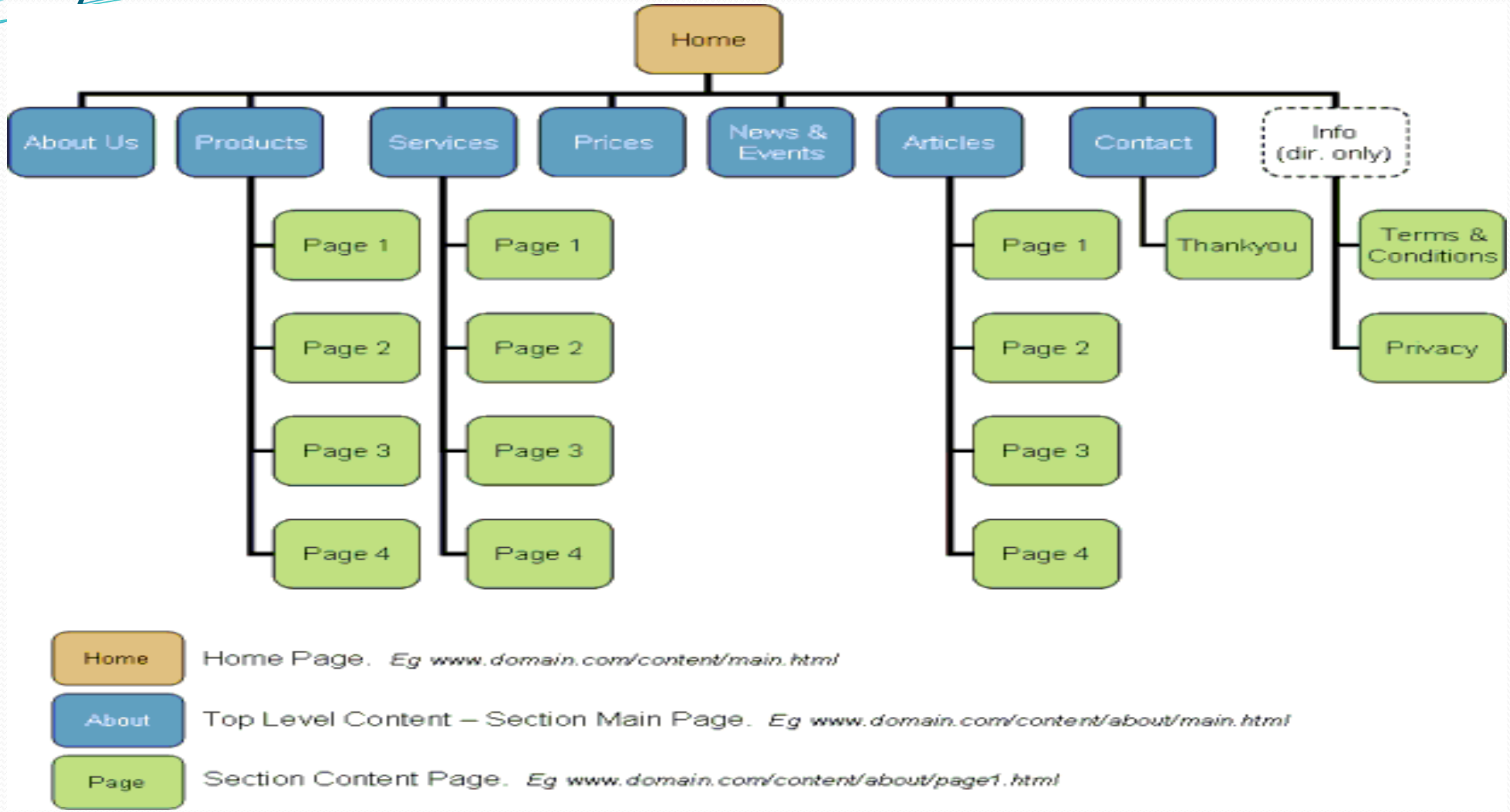
Structure of the Website

- The website structure normally consists of three components:
  - Layout Templates
  - URL patterns
  - Link Structure.

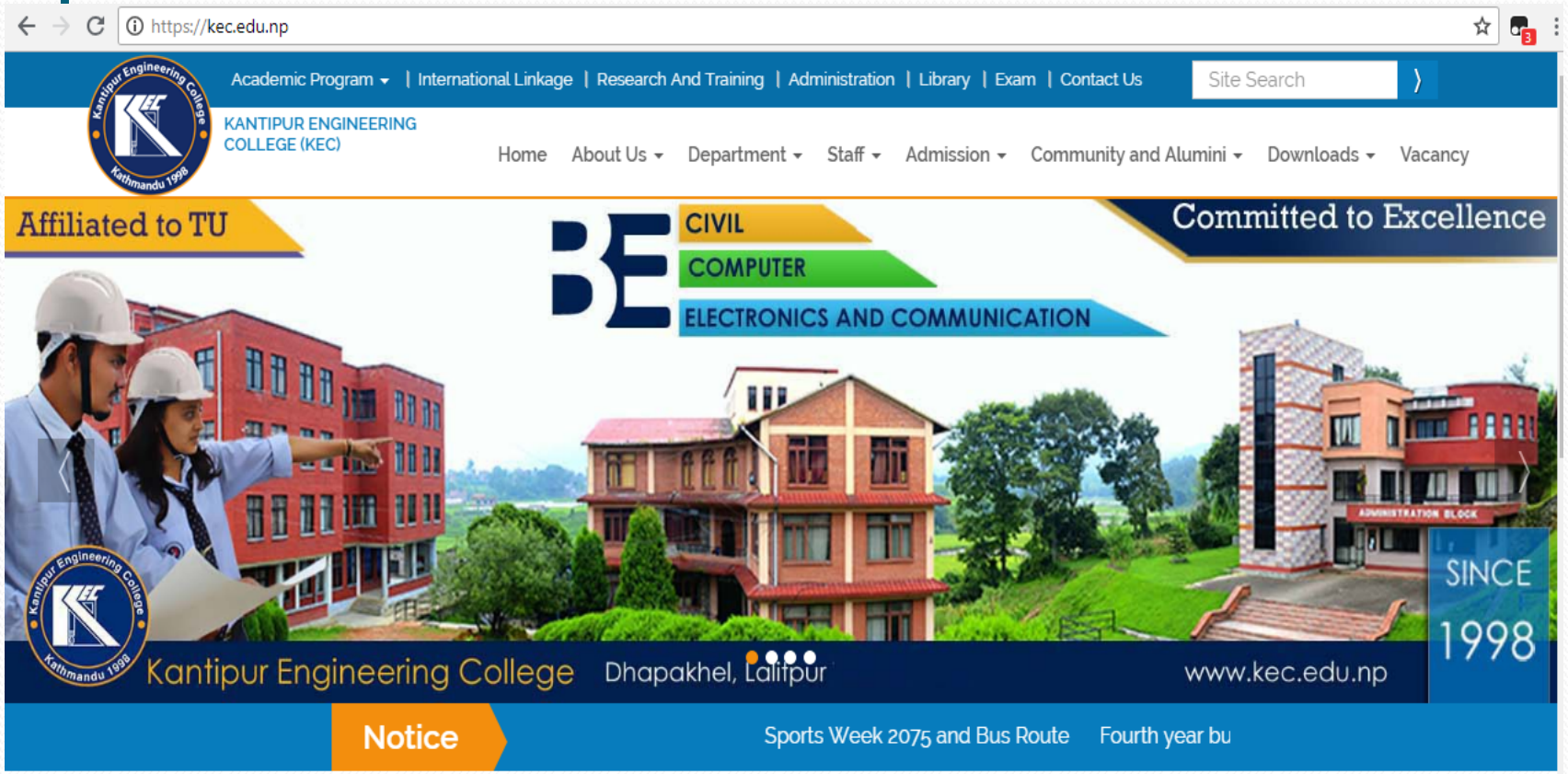
Layout Template

- 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



Example:


https://kec.edu.np
Academic Program | International Linkage | Research And Training | Administration | Library | Exam | Contact Us
Site Search
KANTIPUR ENGINEERING COLLEGE (KEC)
Home About Us Department Staff Admission Community and Alumni Downloads Vacancy
Affiliated to TU
BE
CIVIL
COMPUTER
ELECTRONICS AND COMMUNICATION
Committed to Excellence
SINCE 1998
Kantipur Engineering College Dhapakhel, Lalitpur www.kec.edu.np
Notice
Sports Week 2075 and Bus Route Fourth year bu
Welcome to KEC - Kantipur Engineering College
KEC – Kantipur Engineering College, under the affiliation of Tribhuvan University, has been imparting engineering education since 1998 with the objective to produce qualified and proficient engineers capable of facing the engineering challenges of this modern era, where science and technology is dominating every fields of our life. At the time, engineering subjects were revolutionary—today, they're more important than ever. In these years, we put enormous efforts and spent a lot of resources to pursue excellence in engineering education. At present, KEC is one of the largest and among the few colleges having its own physical infrastructure. To date, we have produced...

URL Pattern

- A URL pattern is a generalization of a group of URLs sharing similar syntactic format.

Valid URL Patterns	Examples	Explanation
Any **substring** of a URL that includes the host/path separating slash	`http://www.google.com/`	Any page on `www.google.com` using the HTTP protocol.
	`www.google.com/`	Any page on `www.google.com` using any supported protocol.
	`google.com/`	Any page in the `google.com` domain.
Any **suffix** of a string. You specify the suffix with the `$` at the end of the string.	`home.html$`	All pages ending with `home.html`.
	`.pdf$`	All pages with the extension `.pdf`.
Any **prefix** of a string. You specify the prefix with the ^ at the beginning of the string. A prefix can be used in combination with the suffix for exact string matches. For example, `^candy cane$` matches the exact string for "candy cane."	`^http://`	Any page using the HTTP protocol.
	`^https://`	Any page using the HTTPS protocol.
	`^http://www.google.com/page.html$`	Only the specified page.

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 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.

  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

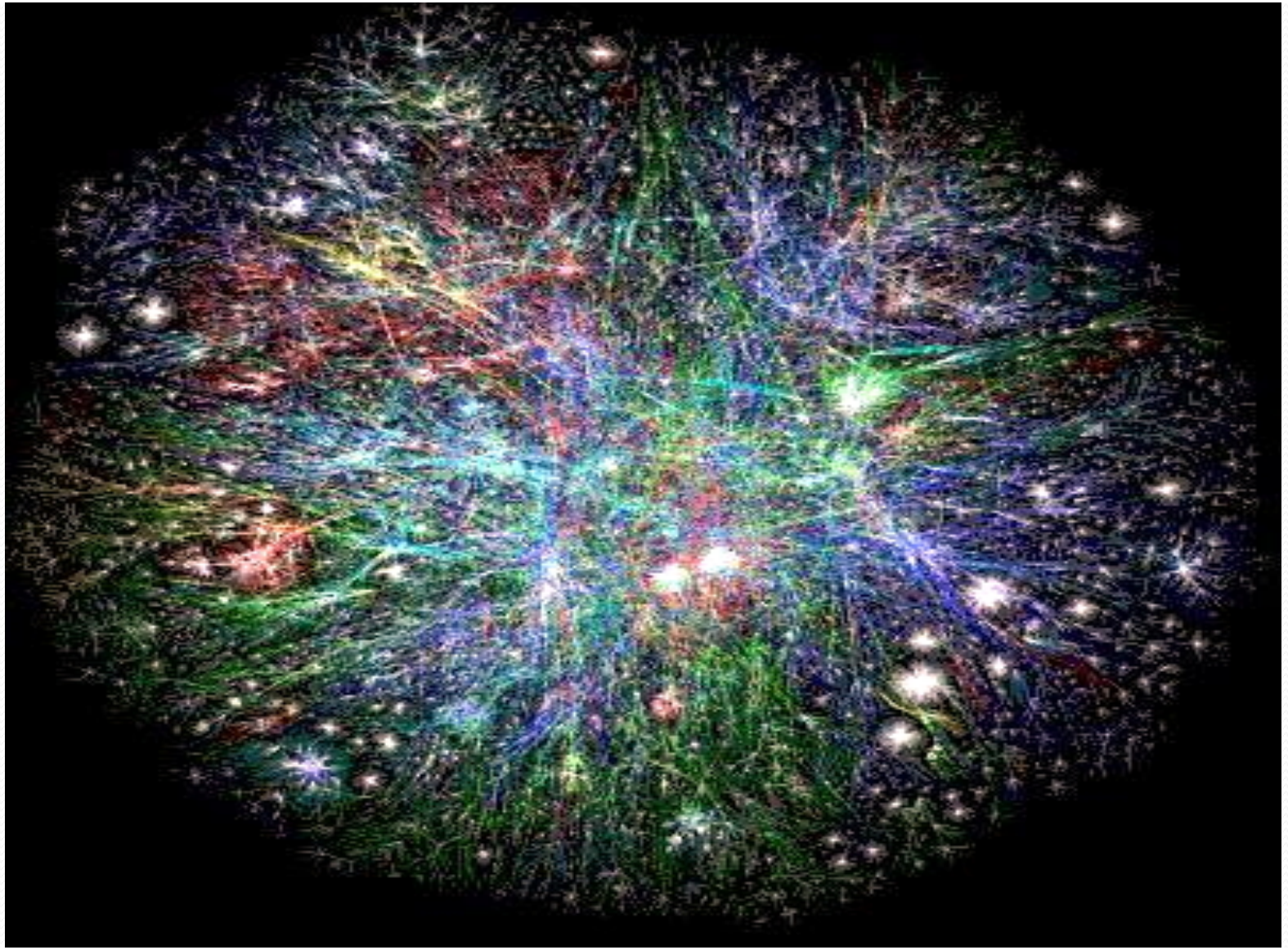
- The head section begins with the <head> start tag and terminates with the</head> end tag.
- Immediately following this comes the <body> start tag and just before the html end tag comes the </body> end tag.
- There is only *one set* of <html>...</html> tags,
- *one set* of <head>...</head> tags and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

How do Web Pages Change

- Most pages do not change much.
- 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.

A Map of The World Wide Web
(http://www.vlib.us/web/worldwideweb3d.html)

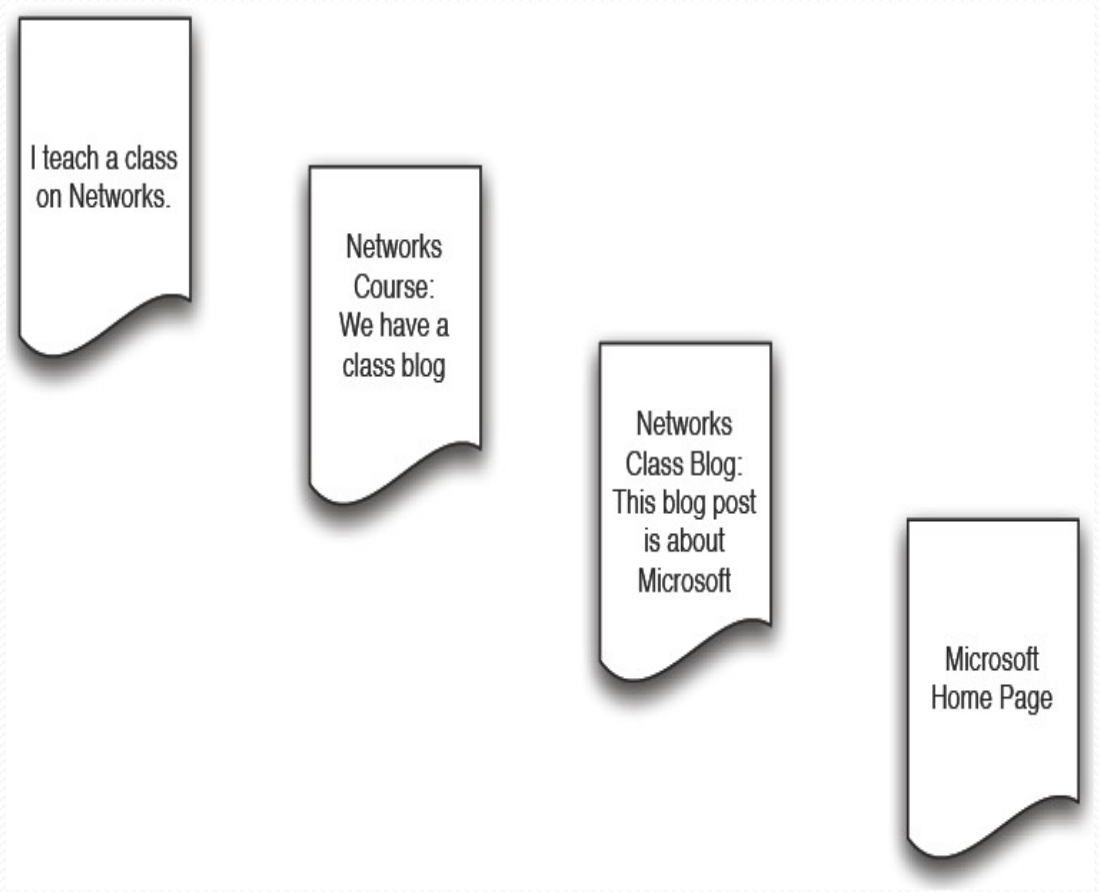
Hyperlinked graph/ network of web pages



The Indexed Web contains **at least 4.75 billion pages** (20 July, 2015)
Source- http://www.worldwidewebsize.com

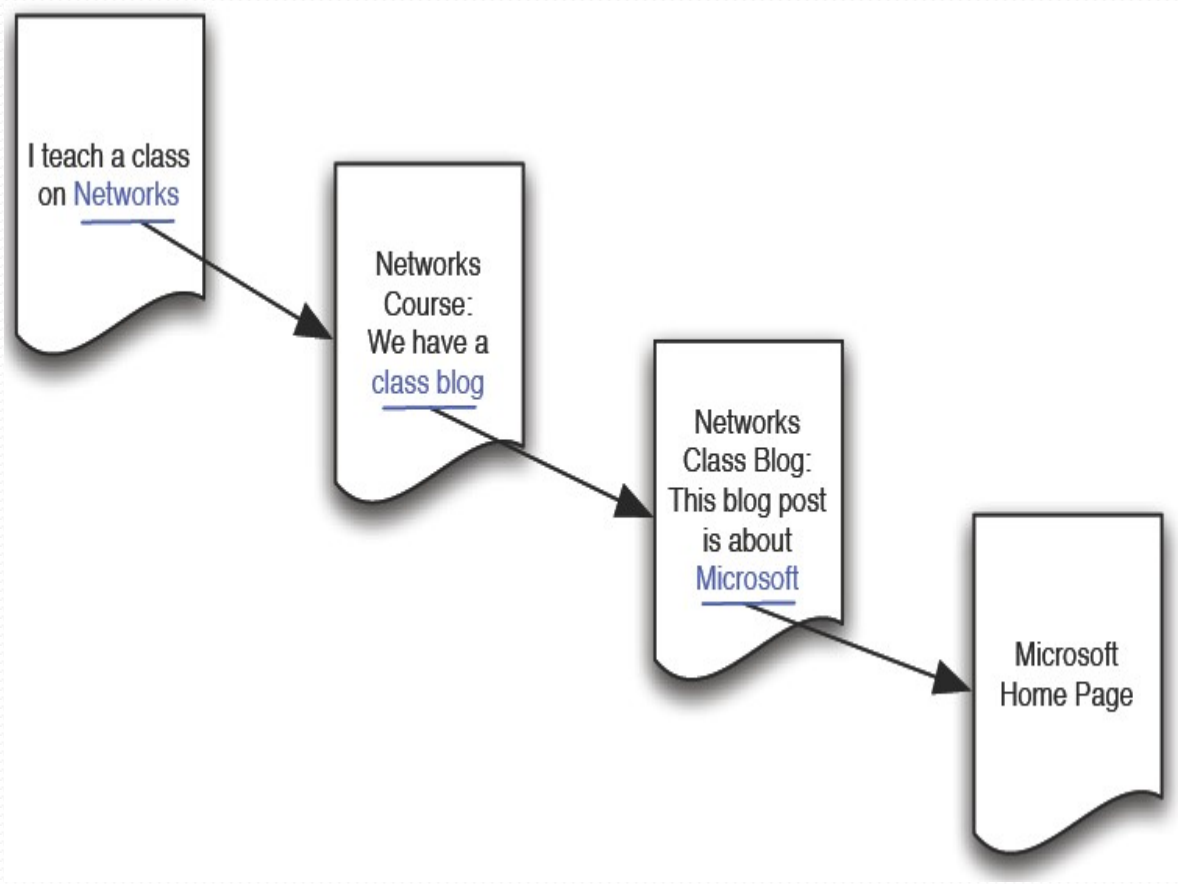
WWW as pages and browsers

- Ex:
  - home page of a college instructor who teaches a class on networks; the home page of the networks class he teaches; the blog for the class, with a post about Microsoft listed at the top; and the corporate home page for Microsoft
  - pages as part of a single coherent system (WWW)
  - pages files on four separate computers, controlled by several organizations, and publically accessible through Web browsers



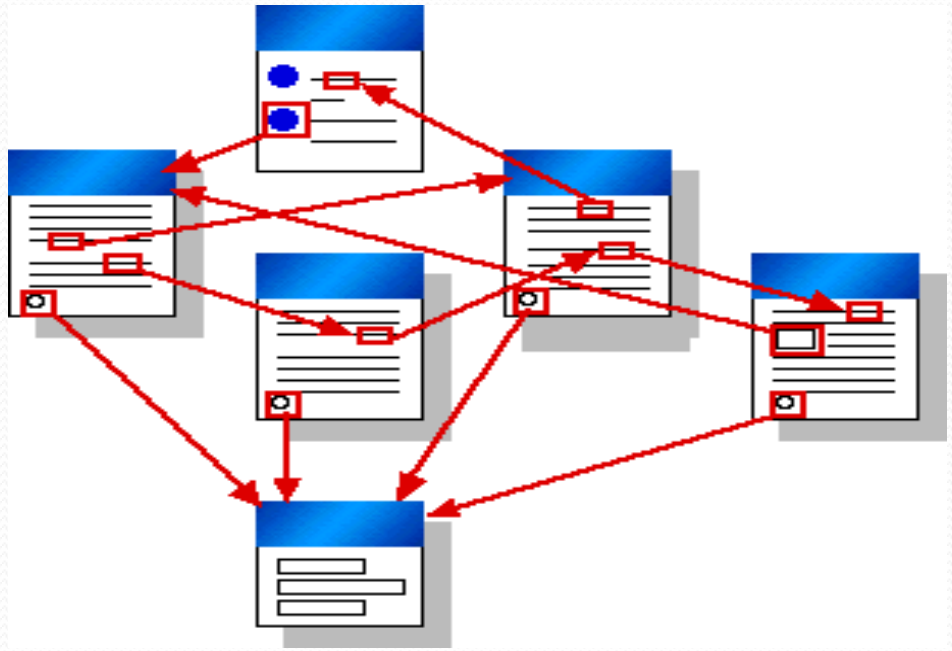
WWW as Information Network

- Hypertext: annotate any portion of a Web page with a virtual link to another Web page



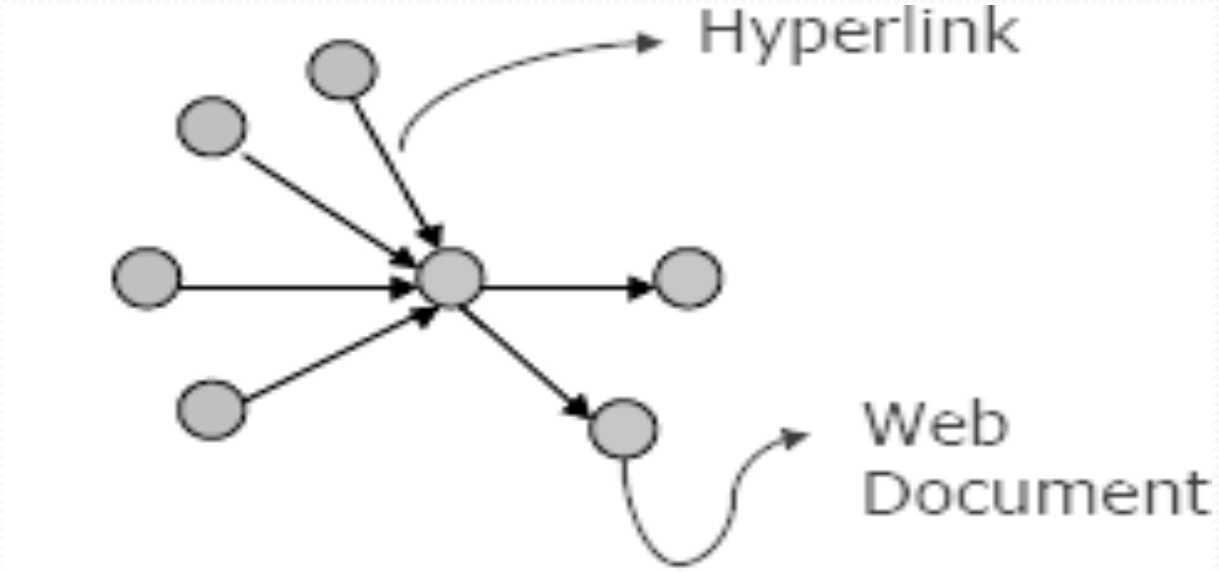
Hypertext

- Replace traditional linear structure of text with a networkstructure
  - any portion of a text linking to any other part
- Web brought hypertext to a global audience
- Web is the largest information network today
- Ted Nelson coined the term hypertext,
  - is the concept behind WWW links



The Web as a Directed Graph

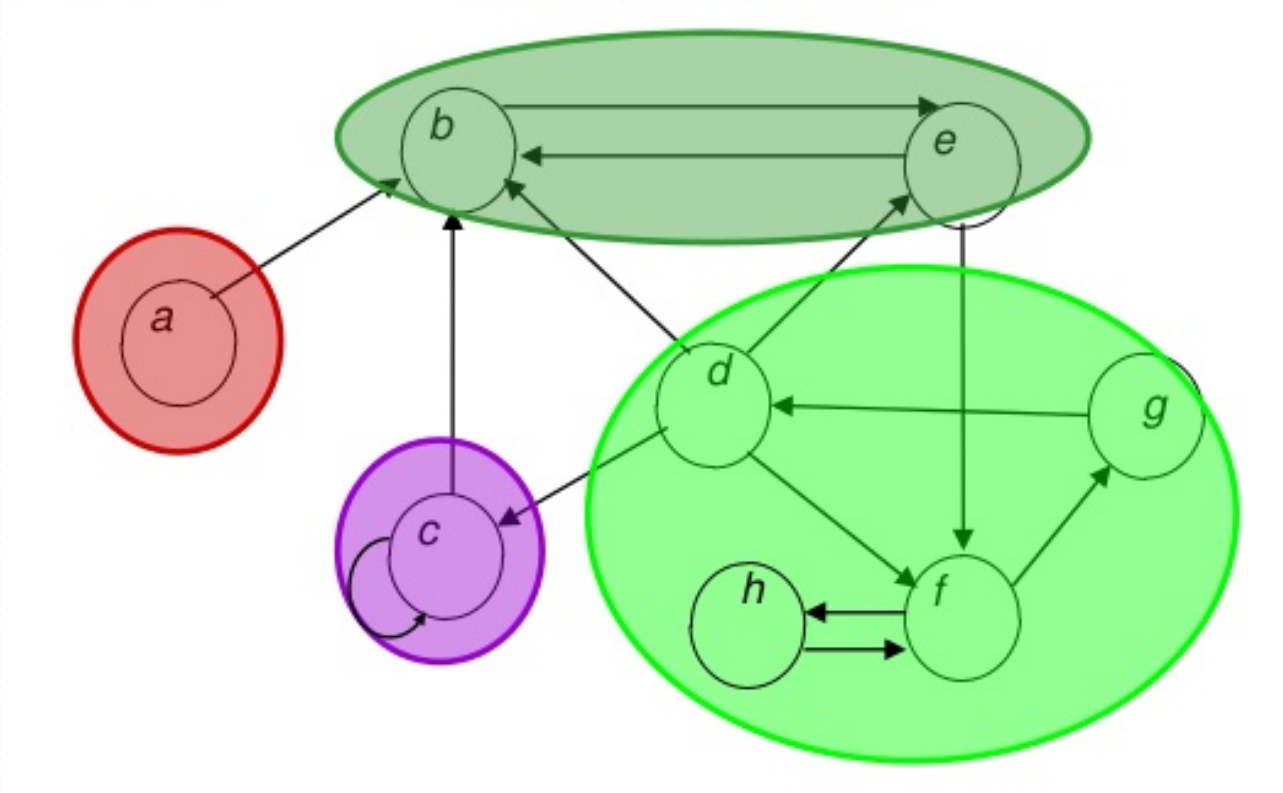
- **Nodes:** web pages
- **Directed edges:** navigational links



Web Graph Structure

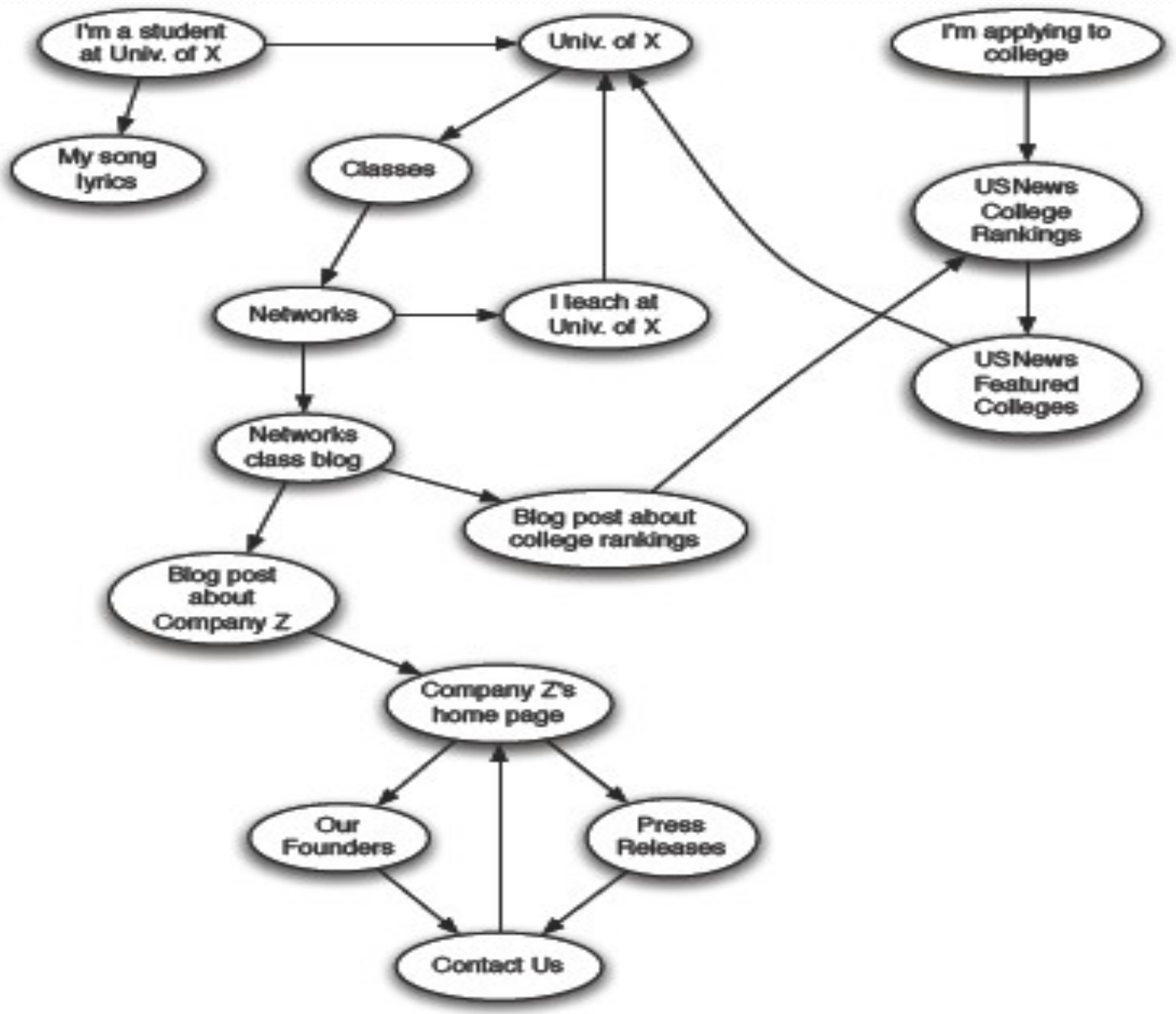
Paths and Strong Connectivity

- Path from node A to node B: sequence of nodes beginning with A and ending with B, where each consecutive pair of nodes is connected by a directed edge (forward direction).



Strongly Connected Directed Graphs

- A directed graph is strongly connected, if there is a path from every node to every other node.
- Three options:
  - Pair of nodes for which each can reach the other("Univ. of X" and "US News College Rankings)
  - Pairs for which one can reach the other but not vice-versa ("US News College Rankings" and "Company Z's hone page")
  - Pairs for which neither can reach the other ("I'm a student as Univ. of X" and I'm applying to college")

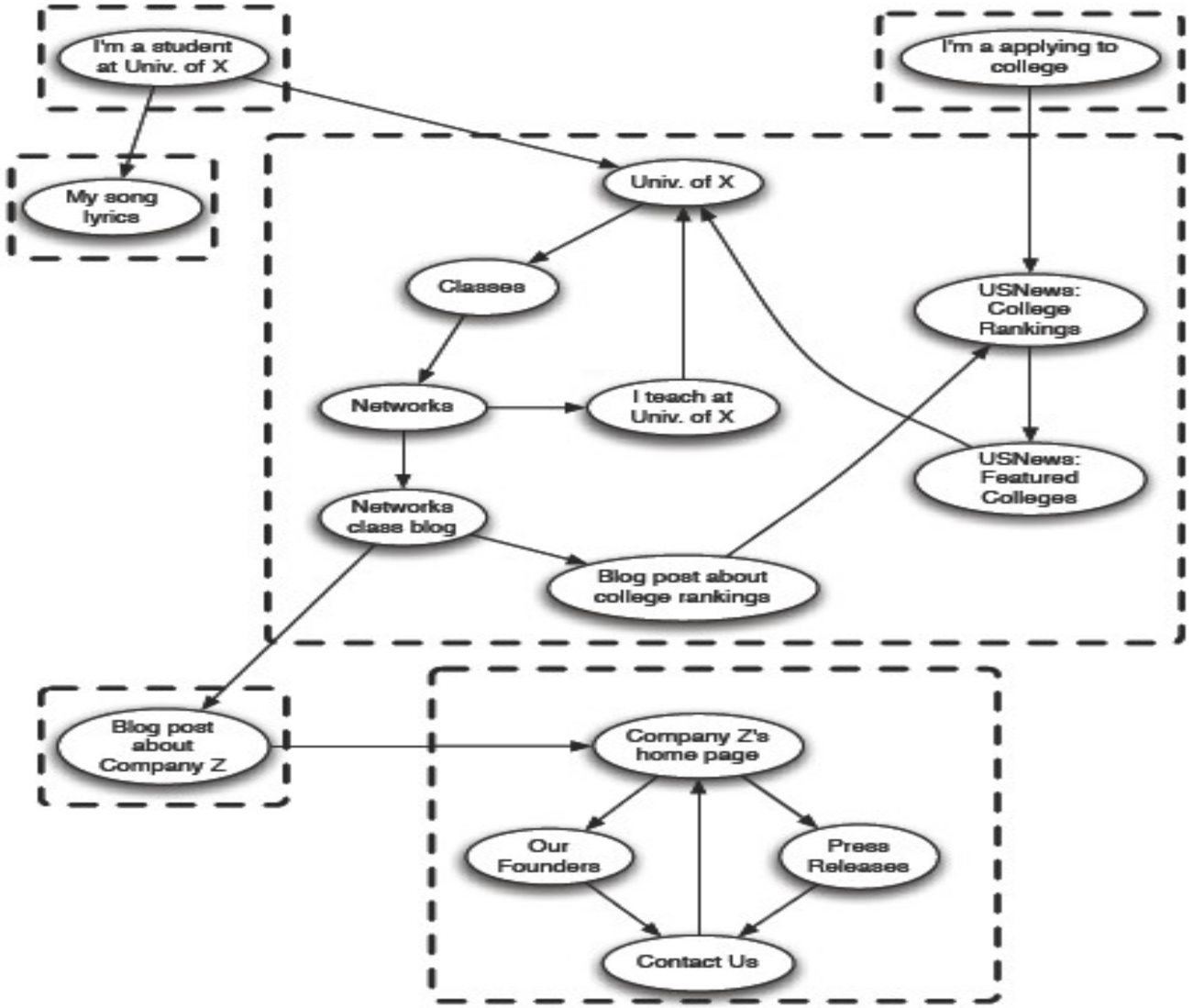


Strongly Connected Component (SCC) Example

- SCC in a directed graph is a subset of the nodes such that:

  (i) every node in the subset has a path to every other; and

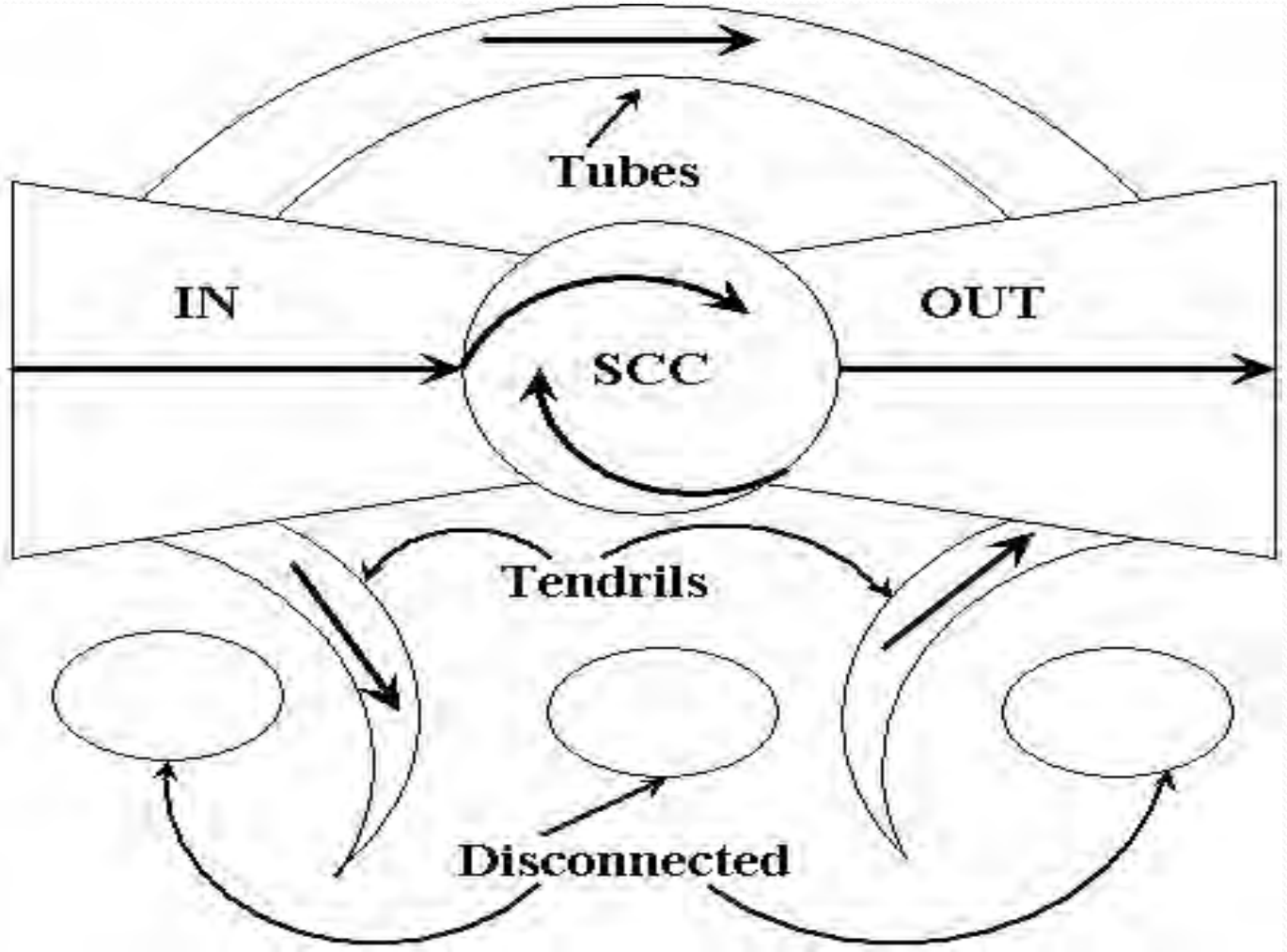
  (ii) the subset is not part of some larger set with the property that every node can reach every other



The Bow-Tie Structure of the Web

- Andrei Broder et al., 1999
  - A global map of the Web, using SCC as the basic building blocks (component graph).
  - ***Dividing Web into a few large pieces and show how they fit together.***
- Data
  - navigational "backbone" indexed by AltaVista (1999)
  - Pioneering research verified by others (newer studies with navigational "backbone" indexed by Google, Wikipedia, etc.)

The Bow-Tie Structure of the Web



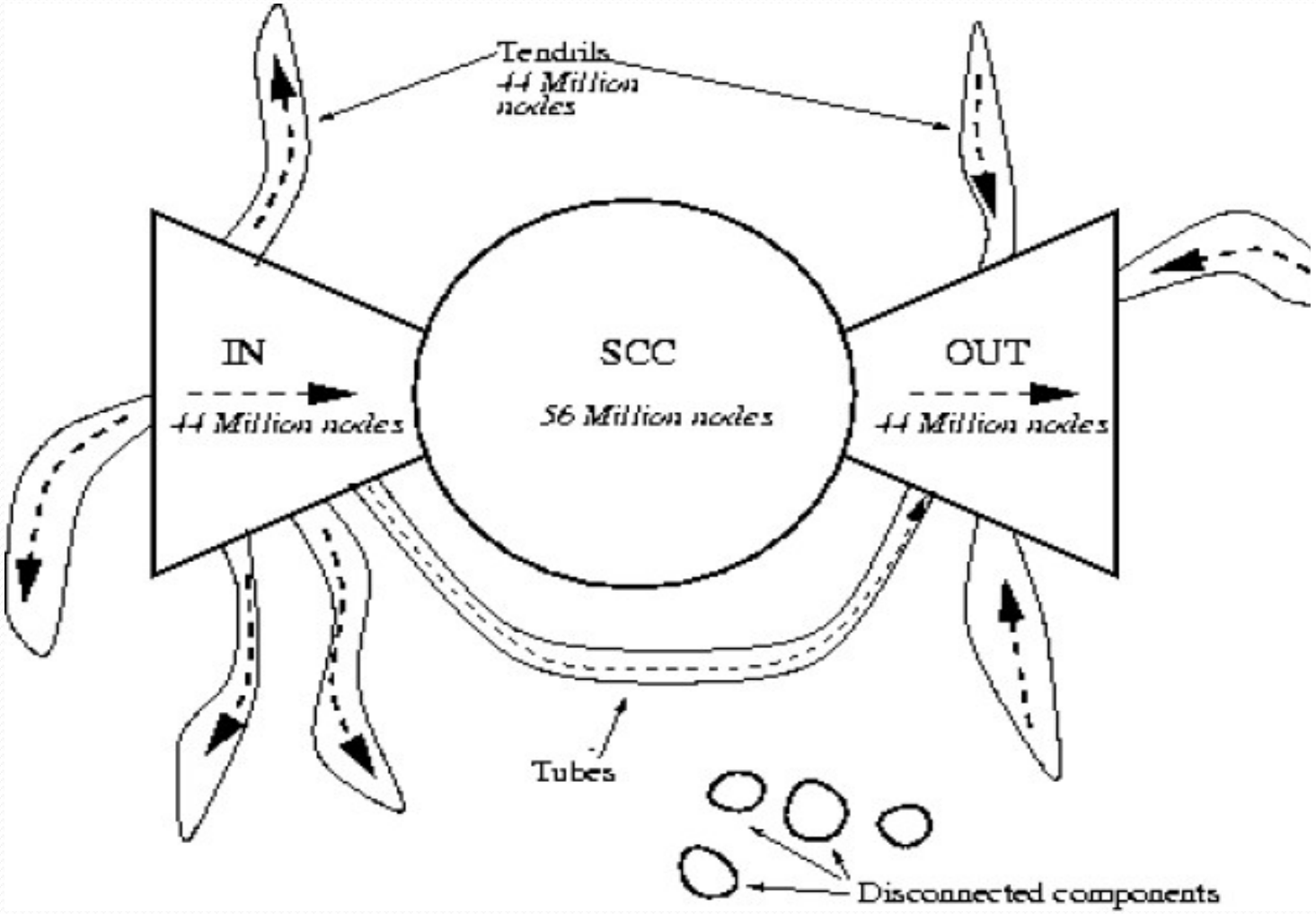
SCC – the Strongly Connected Component

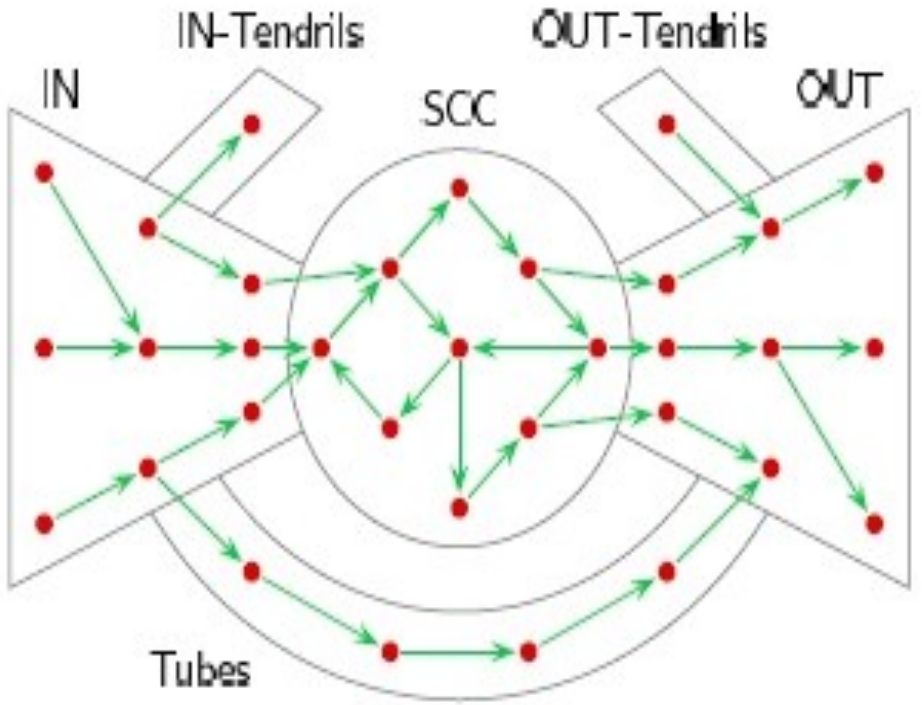
(27.5%)

IN (21.5%)
OUT (21.5%)
Tubes and Tendrils (21.5%)
Disconnected (8%)

See ref. book: *An Introduction to Search Engines and Web Navigation* (M. Levene)

The Bow-Tie Structure of the Web





The Bow-Tie Structure of the Web

- The Web contains a single giant SCC
  - giant SCC contains a significant fraction of all pages (also most important pages: major commercial, governmental, and non-profit organizations)
- Position all the remaining SCC in relation to the giant SCC, by classifying nodes by their ability to reach and be reached from the giant SCC
  - IN: nodes that can reach the giant SCC but cannot be reached from it
    - Pages not "discovered" by members of the giant SCC
  - OUT: nodes that can be reached from the giant SCC but cannot reach it
    - Pages receiving links from the giant SCC, but not linking back

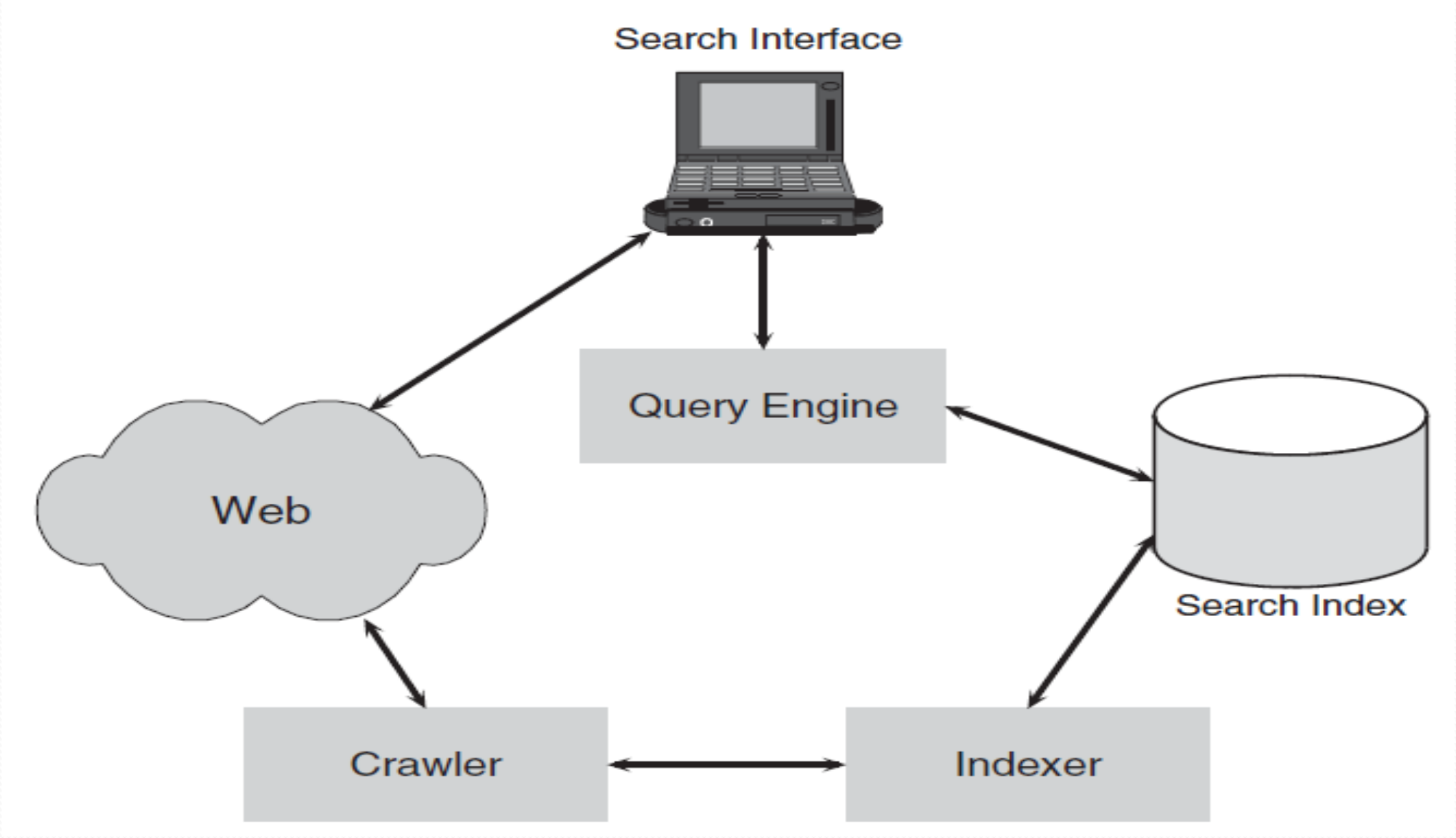
The Bow-Tie Structure of the Web

- There are pages that belong to none of IN, OUT, or the giant SCC
  - neither reach the giant SCC nor be reached from it
- Categories of such pages:
  - Tendrils: (a) nodes reachable from IN that cannot reach the giant SCC, or (b) nodes that can reach OUT but cannot be reached from the giant SCC
  - Tube: nodes satisfying both (a) and (b) above
  - Disconnected: otherwise (nodes that would not have a path to the giant SCC even if we ignored directions of the edges)

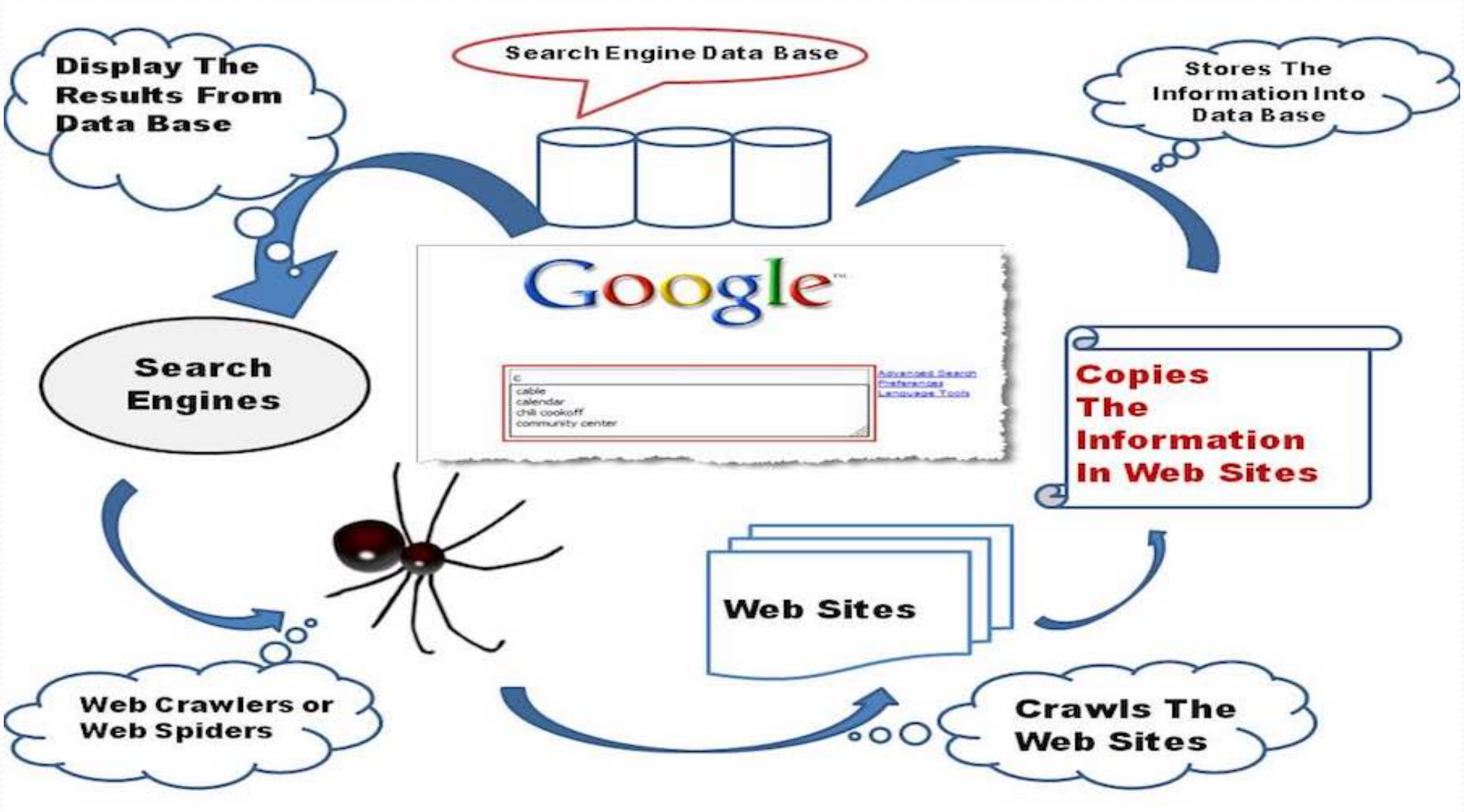
Finding Information on the Web

- Search Engines
- Navigation / Browsing – exploratory search
  - Follow Hyperlinks
  - Web directories
  - Web portals
- Information in Databases
  - Hidden Web / Deep Web
- Recommendation systems

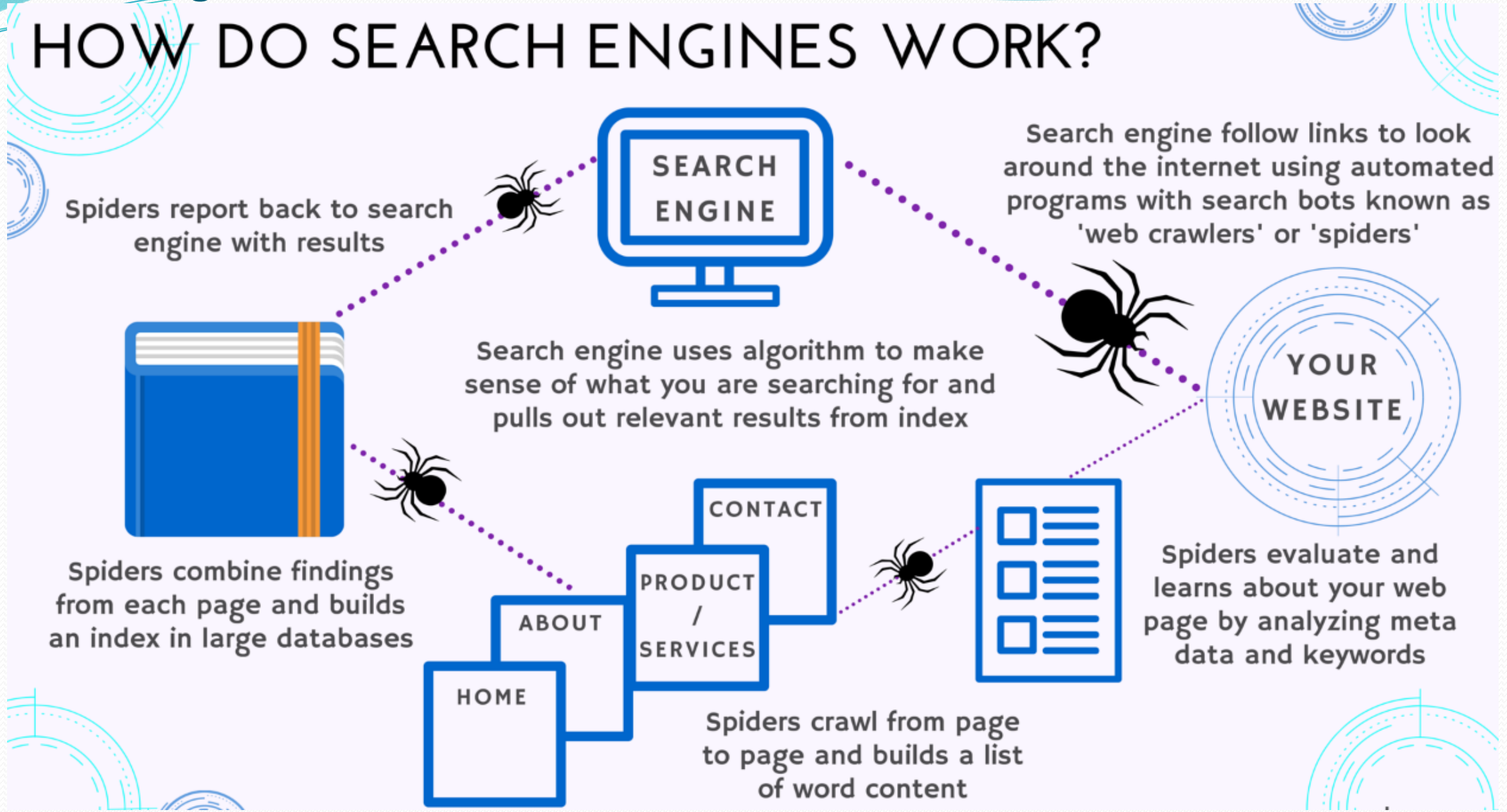
Searching the web/How Search Engines Work?



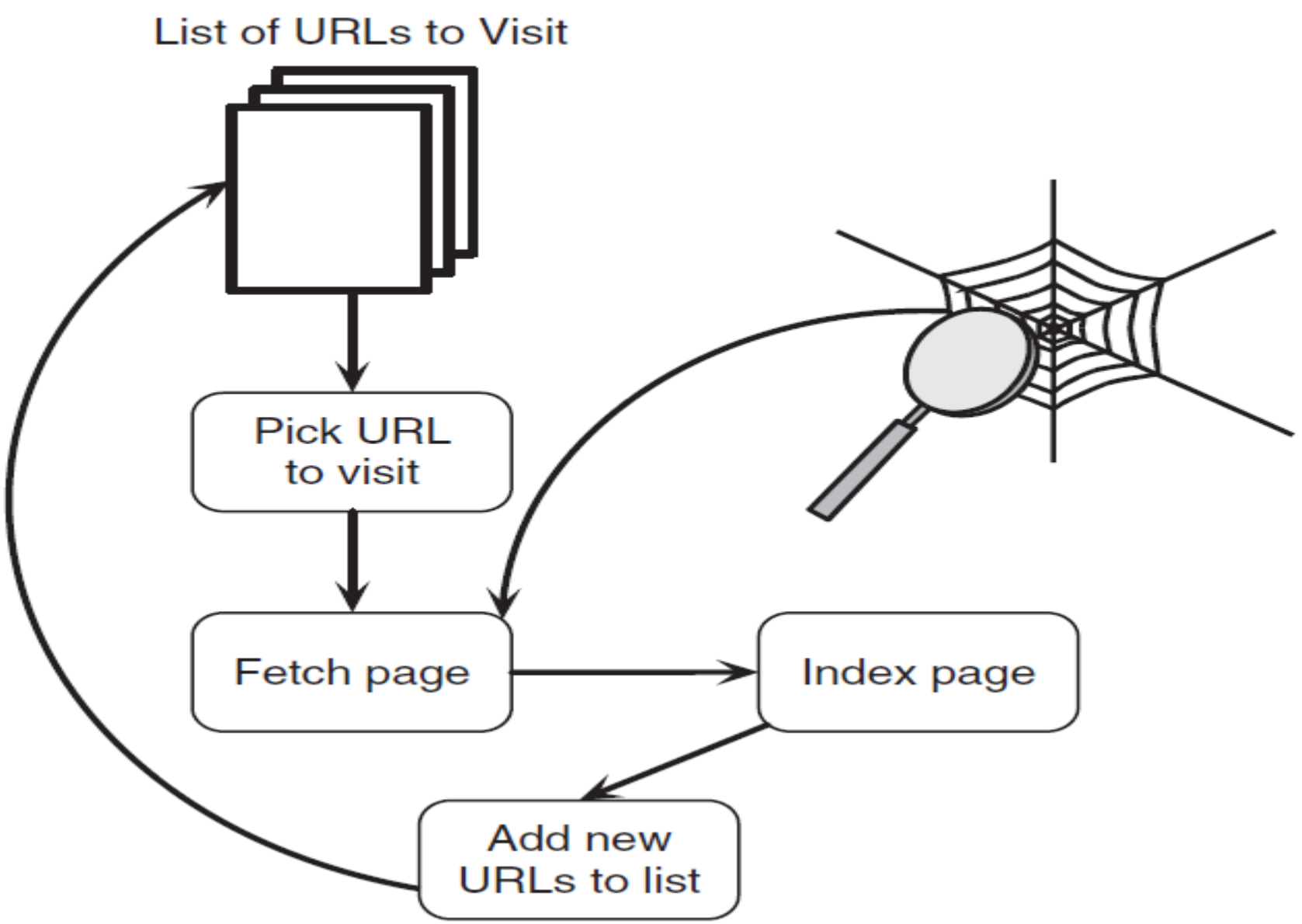
Searching the web



Searching the web



Crawling the Web



Crawling the Web

- Crawling refers to the ability of a search engine **to traverse the billions of interlinked pages on the world wide web.**
- Crawling is the process by which search engines discover updated content on the web, such as new sites or pages, changes to existing sites, and links.
- When a web crawler visits a page, it collects every link on the page and adds them to its list of next pages to visit. It goes to the next page in its list, collects the links on *that* page, and repeats.
- Web crawlers also revisit past pages once in a while to see if any changes happened.
- Google's spiders could read several thousands of pages per second.
-

Indexing Web Pages

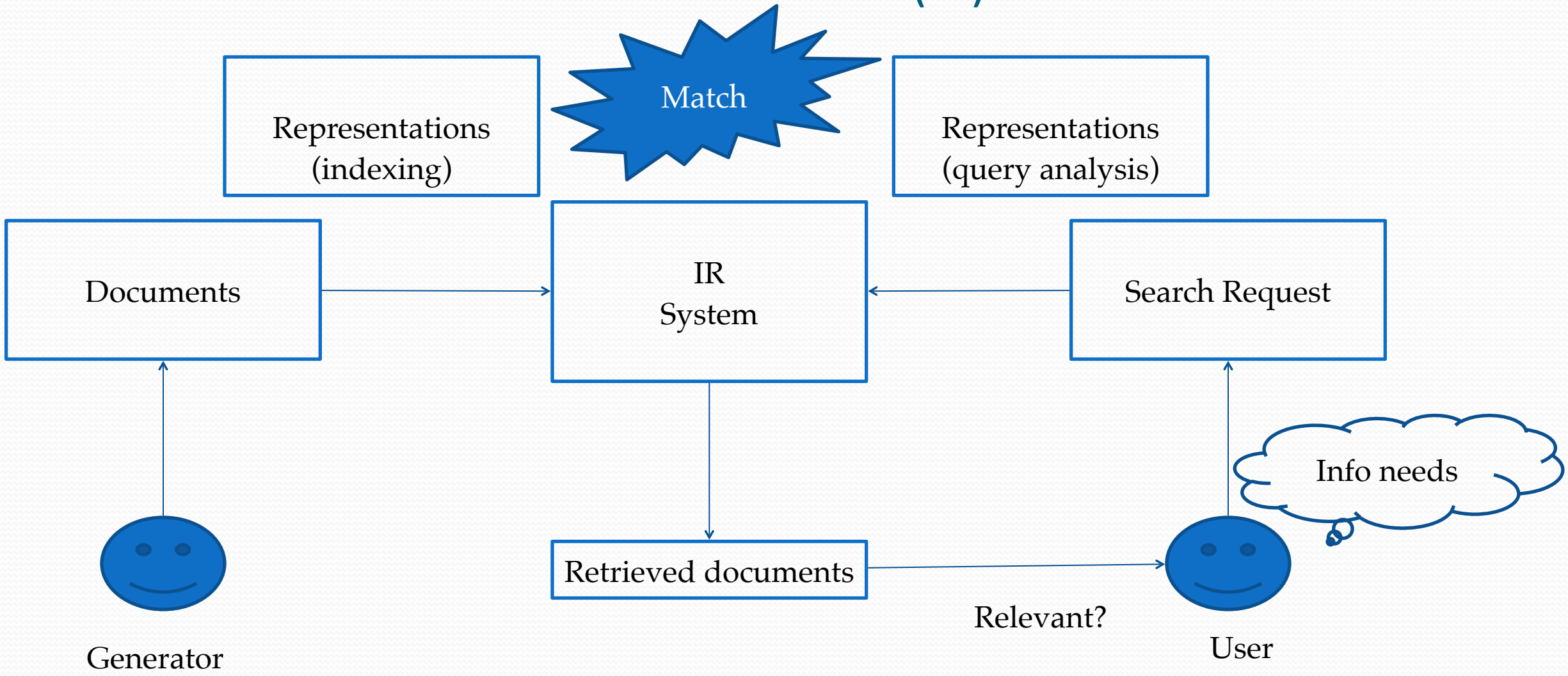
- Once a spider has crawled a web page, the copy that is made is returned to the search engine and stored in a data center(cached).
  For two purposes:
  - to return results **related** to a search engine user's query
  - to rank those results **in order** of importance and relevancy
- After completing the crawling, it compiles a massive index of all the words it sees and their location on each page.
- **It is essentially a database of billions of web pages.**
- Indexing is the process of organizing the masses of data and pages so they can be searched quickly for relevant results to your search query.

Indexing Web Pages

- Index the contents of each crawled web page
- Document preprocessing
  - Strip tags
  - Segment/tokenize
  - Remove stop words
  - Stemming
    - Reduce every word to its root form
- Metadata
  - <meta> element
    - Keywords, content, author, etc.

Retrieval and Ranking

General Process of Information Retrieval (IR)



Relevance

- The issue of “relevance” has haunted the IR community for decades
- Relevance is a relative concept and depends not only on the query but also on the user and the context in which a query is issued.
- How do we measure **relevance** of a web page (document) to a query?

Ranking Search Results by Relevance

- Content relevance
  - Vector space model
  - Language model
- Link analysis
  - Page rank algorithm - Google

Vector Space Model

- Every document and every query is represented by a vector or n-tuple
  - Each element represents a particular term
- Vectors may be
  - Binary: 1 or 0, presence or absence of the term
  - Eg,
    - Terms = <t1, t2, t3, t4, t5, t6, t7, t8, t9>
    - docA = <0, 1, 1, 1, 1, 0, 0, 1, 0>
    - docB = <1, 0, 1, 0, 1, 1, 1, 1, 0>
    - Query1 = <0, 0, 1, 1, 0, 0, 0, 0, 0>
  - Non-binary: values represent weights assigned to terms
    - Eg, docF = <3, 0, 0, 9, 1, 4, 4, 5, 0>

Vector Space Model

- The Vector Space Model (VSM) is a way of representing documents through the words that they contain
- It is a standard technique in Information Retrieval
- The VSM allows decisions to be made about which documents are similar to each other and to keyword queries

How it works

- Each document is broken down into a word frequency table.
- The tables are called vectors and can be stored as arrays.
- A vocabulary is built from all the words in all documents in the system.
- Each document is represented as a vector based against the vocabulary.

Example

- Document A
  - "a dog and a cat."

a	dog	and	cat
2	1	1	1

- Document B
  - "a frog."

a	frog
1	1

- The vocabulary contains all words used

  a, dog, and, cat, frog

- The vocabulary needs to be sorted

  a, and, cat, dog, frog

Example, continued

- Document A: “a dog and a cat.”

a	and	cat	dog	frog
2	1	1	1	0

  - Vector: (2,1,1,1,0)
- Document B: “a frog.”

a	and	cat	dog	frog
1	0	0	0	1

  - Vector: (1,0,0,0,1)
- **Queries** can be represented as vectors in the same way as documents:
  - Dog = (0,0,0,1,0)

Similarity measures

- There are many different ways to measure how similar two documents are, or how similar a document is to a query.
- The cosine measure is a very common similarity measure.
- Using a similarity measure, a set of documents can be compared to a query and the most similar document returned.

The cosine measure

- For two vectors d and d′ the cosine similarity between d and d′ is given by:

$$\frac{d \times d'}{|d||d'|}$$

- Here d X d′ is the vector product of d and d′, calculated by multiplying corresponding frequencies together.
- The cosine measure calculates the angle between the vectors in a high-dimensional virtual space.

Example

- Let d = (2,1,1,1,0) and d′ = (0,0,0,1,0)
  - dXd′ = 2X0 + 1X0 + 1X0 + 1X1 + 0X0=1
  - $|d| = \sqrt{(2^2+1^2+1^2+1^2+0^2)} = \sqrt{7}=2.646$
  - $|d'| = \sqrt{(0^2+0^2+0^2+1^2+0^2)} = \sqrt{1}=1$
  - Similarity = 1 / (1 X 2.646) = 0.378
- Let d = (1,0,0,0,1) and d′ = (0,0,0,1,0)
  - Similarity =??

Ranking Results

- Calculate a similarity measure (eg, Cosine Similarity) between the query and document vectors.
- Documents more similar to the query are ranked higher in the results.

- Evaluation of Information Retrieval
  - Recall (How many relevant results retrieved?)
  - Precision (How many results are relevant?)

Some Information Retrieval Packages

- Lucene
  - Java based
- SMART
  - Vector space model
  - ftp://ftp.cs.cornell.edu/pub/smart
- Lemur
  - Language model
  - http://www-2.cs.cmu.edu/~lemur/

Other Considerations

- Phrase matching
- Synonyms
- Link text
- URL analysis
- Date last updated
- HTML structure weighting
- Spelling variations and errors
- ……

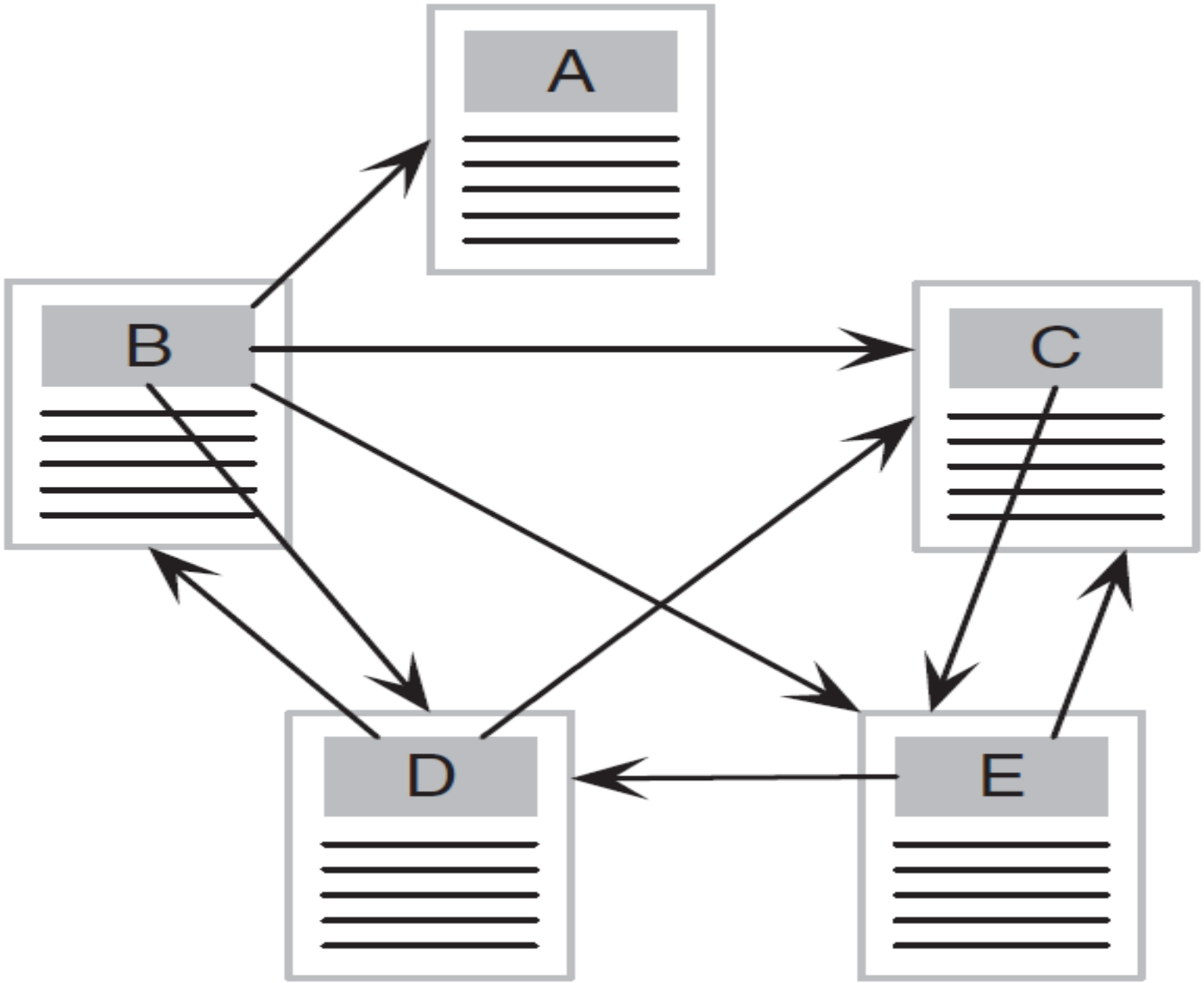
Link Analysis

- 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.
- Web pages have links to related and useful resources
- Meaning of a hyperlink from page A to B
  - a recommendation or endorsement of page B by the author of page A
- Google interprets a link from page A to page B as a vote by the author of page A for page B.

Google's PageRank Explanation



- Random surfing (with occasional teleportation)
- PageRank of a web page is the long-run probability that the surfer will visit that page
- Statistical modeling using Markov Chains

Google's PageRank Explanation

- PageRank (PR) is an algorithm used by Google Search to rank websites in their search engine results.
- PageRank was named after Larry Page, one of the founders of Google.
- PageRank is a way of measuring the importance of website pages.
  According to Google:
  - PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. ***The underlying assumption is that more important websites are likely to receive more links from other websites.***

Link Analysis

- Two types of links
- Referential Links
  - Navigational links
  - May be unrelated pages
- Informational links
  - If A has a link to B, then the content of B is likely to be similar to the content of A
  - This is called Topic locality
- Combining Link Analysis with Content Relevance

Navigating the Web

- Navigation potential of a web page
  - How good starting point for navigation?

- First, it should be **relevant** to the information seeking goals of the surfer
- Secondly, it should be **central** - its distance to other web pages should be minimal
- Thirdly, it should be well **connected** - able to reach a maximum of other web pages

- Some metrics for navigation potential
  - Potential gain, Compactness, Stratum

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 the 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.
- 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 sites 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

- Collection and analysis of Web access information for Web pages
- Discover patterns in web access/usage data
- Capture the identity or origin of Web users along with their browsing behaviour
- Web usage data collected from
  - Web/application server logs
  - Proxy server logs
  - User's browsers

Web Server logs

1	2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://dataminingresources.blogspot.com/
2	2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://maya.cs.depaul.edu/~classes/cs589/papers.html
3	2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200 318814 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1) http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey
4	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/
5	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/styles2.css - 200 1636 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html
6	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/header.gif - 200 6027 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)

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.

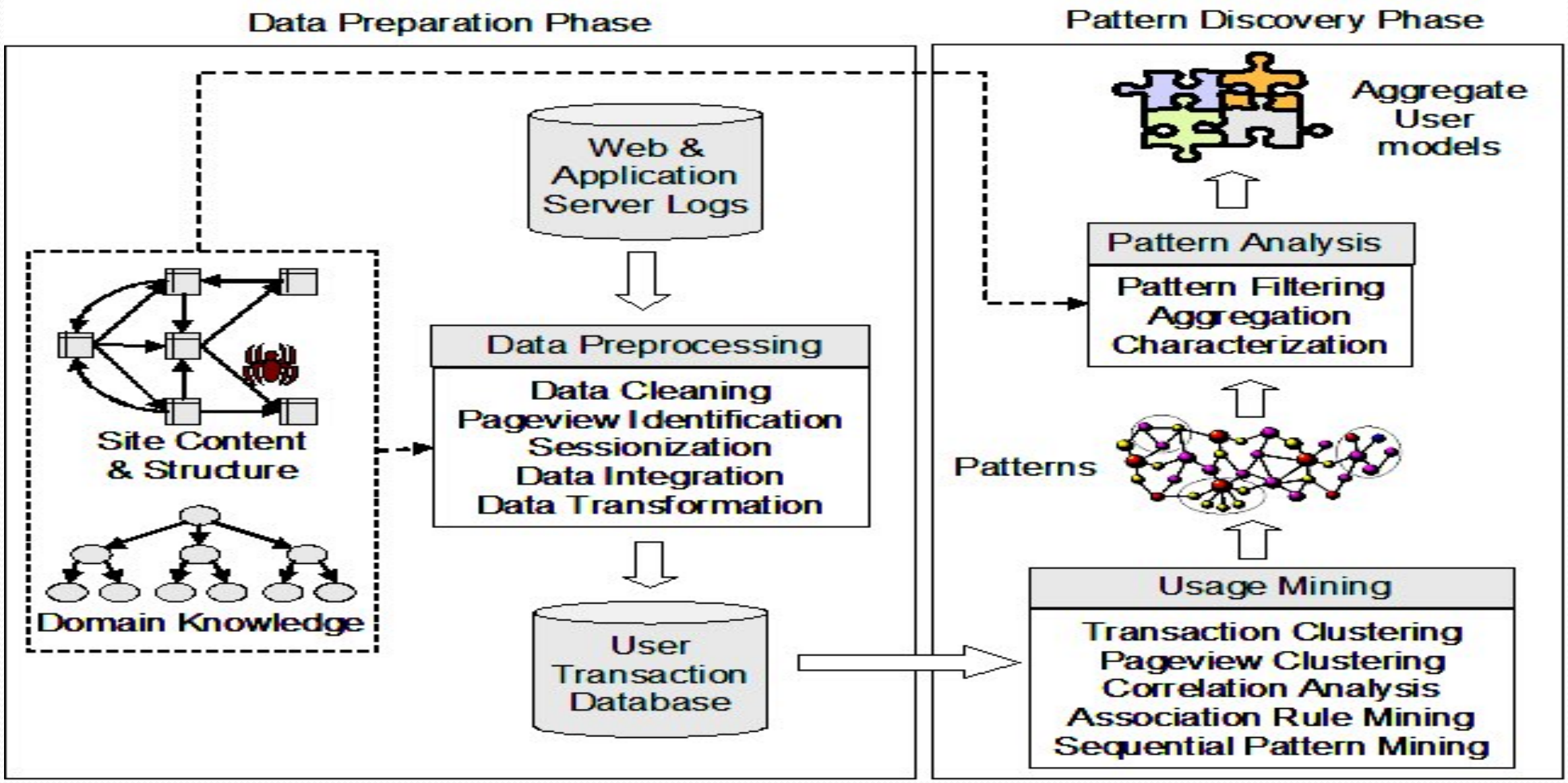
How to perform Web Usage Mining?

- 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.

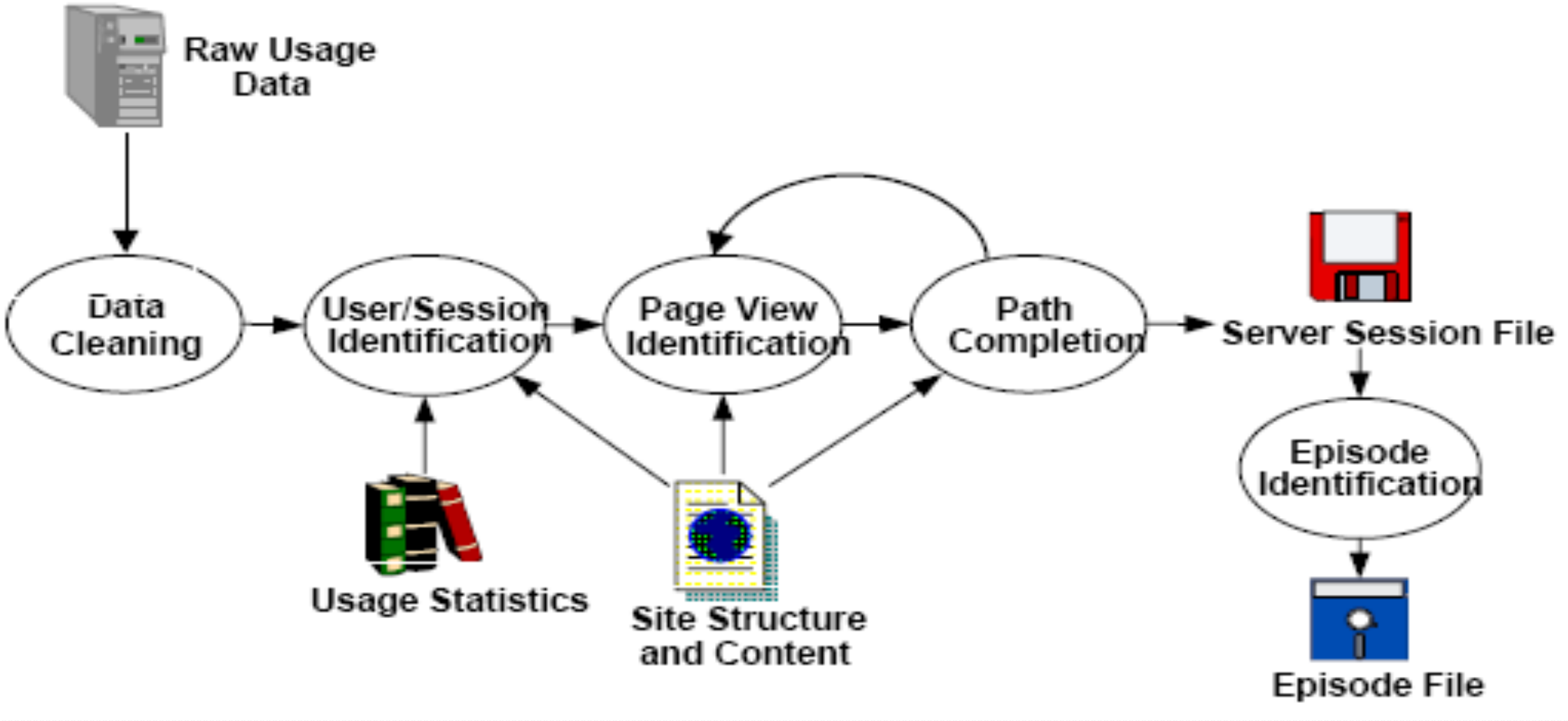
Data in Web Usage Mining

- 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



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)

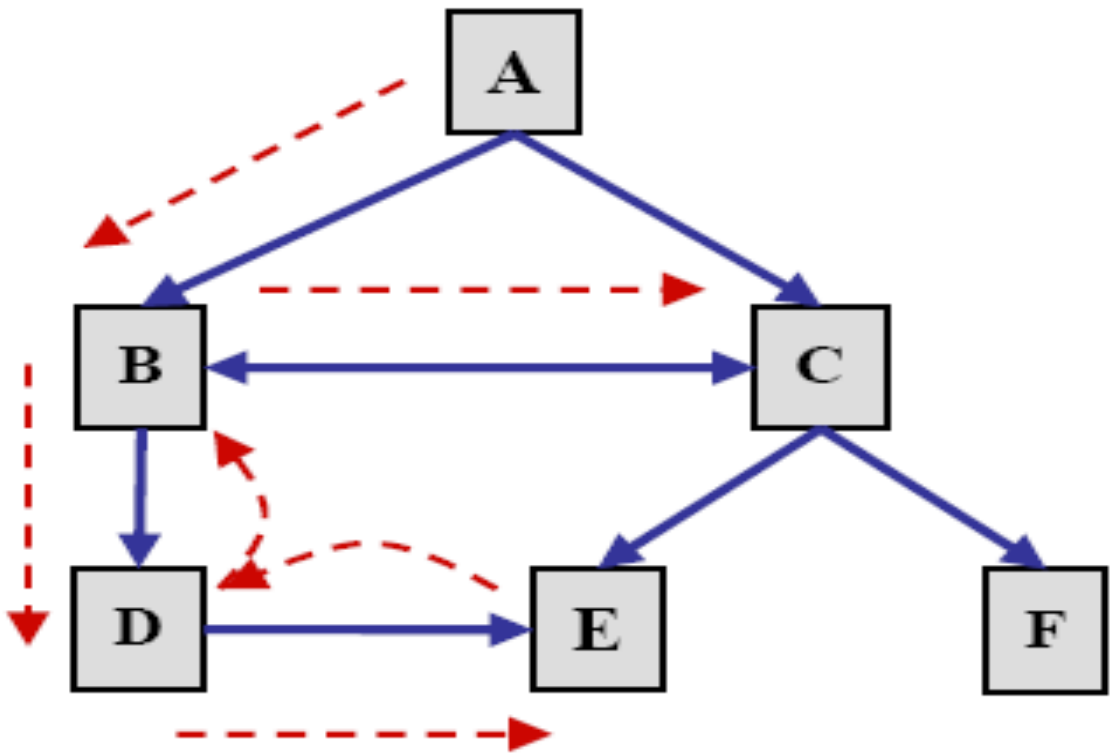
Identify sessions (Sessionization)

- 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.

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.

Identifying the Surfer

- Host name or IP address recorded in log file
- Cookies from user browser
- Registered User login
- Sessions
  - Time-oriented sessionizing
  - Navigation-oriented sessionizing
- Geographical analysis

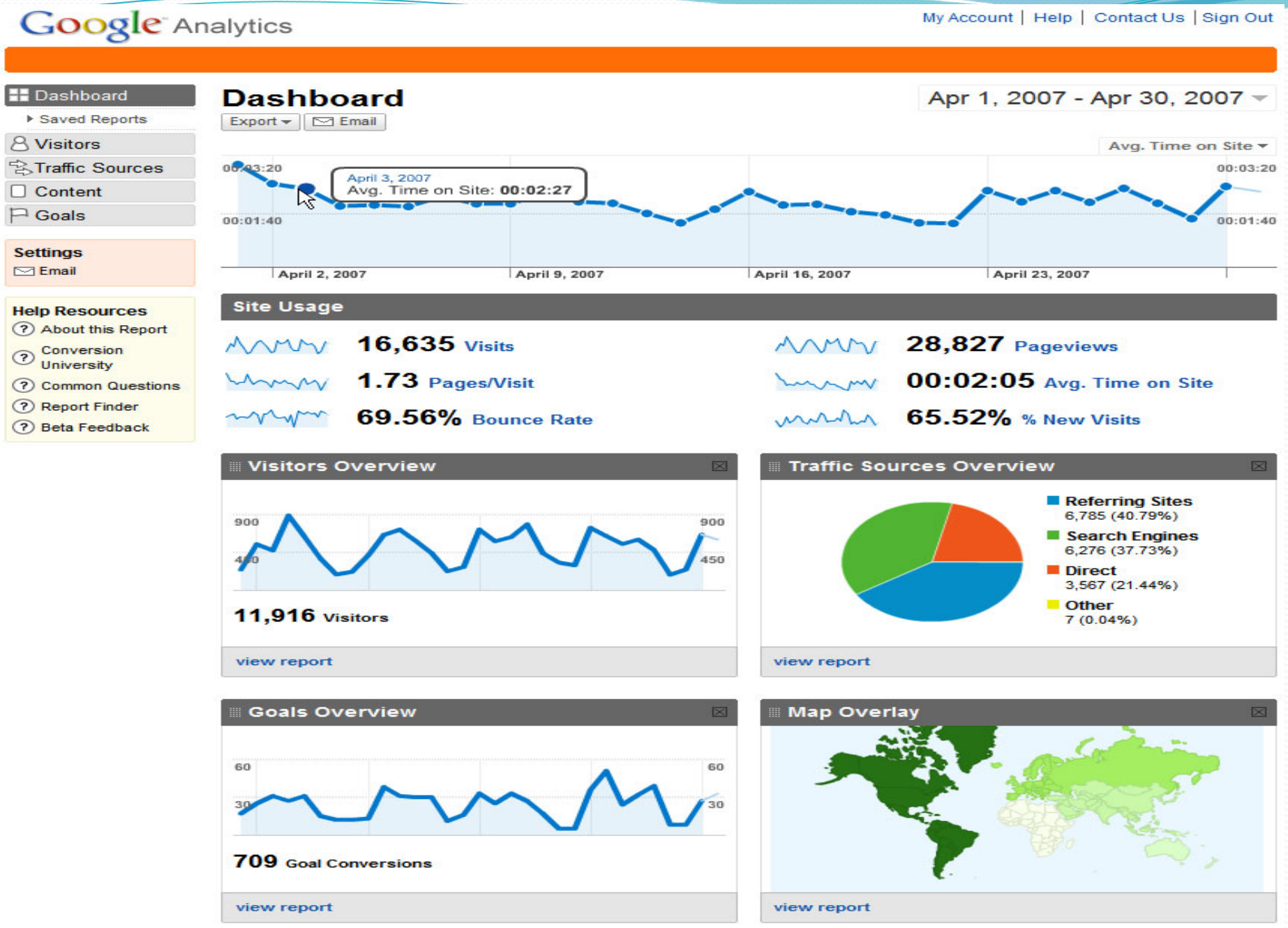
Web Analytics Measures

- Hits or Traffic
- Conversion rate (visitors into customers)
- Batting average (description page to purchase)
- Stickiness - how long users stay
- Slipperiness - how quickly users exit
- Leakage - drop out rate once users have entered
- Take rate - % of visitors that complete some activity
- Repeat visitor rate … etc.

Web Analytics Tools and Services

- Google Analytics
- Yahoo Web Analytics
- Alexa
- Analog Weblog file analyzer
- List of Web Analytics software

Google Analytics



Applications of Web Usage Mining

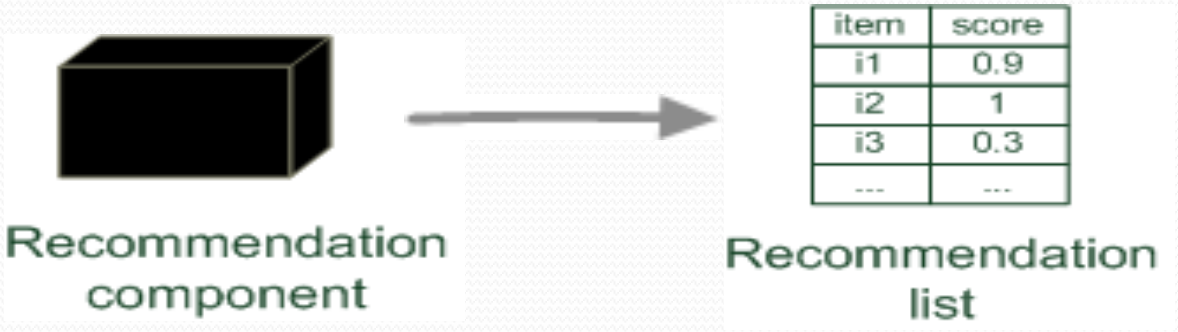
- Allows companies to produce information pertaining to the future of their business function ability.
- Customer analysis and relation management
- IBM's Web Fountain tracks a company's reputation over the web
- Adaptive web site – modified according to how users access the site
  - Personalization
  - Personalized marketing
- Prefetching and caching of web pages
- Security
  - Identifying potential threats or intrusion
  - Anomaly detection

Recommender systems

- **How to answer the following?**
  - *Which digital camera should I buy?*
  - *What is the best holiday for me and my family?*
  - *Which is the best investment for supporting the education of my children?*
  - *Which movie should I watch?*
  - *Which web sites will I find interesting?*
  - *Which book should I buy for my next vacation?*
  - *Which degree and university are the best for my future?*
  - ……
- Too much information: information overload – consumers have too many options.
- **A recommender system is a system which provides recommendations to a user.**

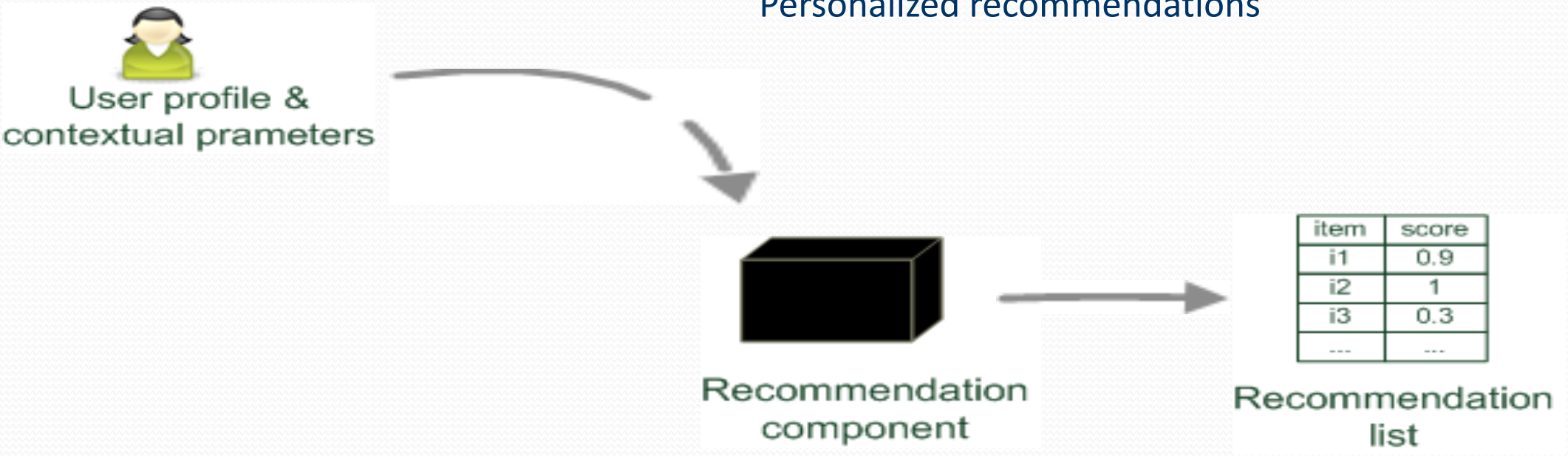
Paradigms of recommender systems

Recommender systems reduce information overload by estimating relevance



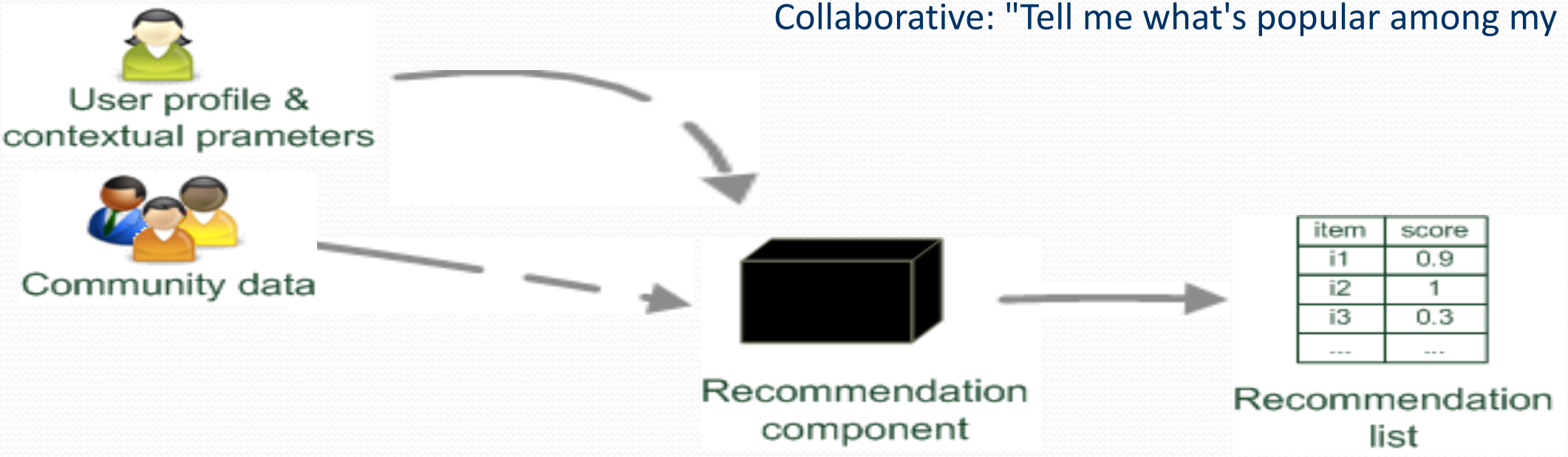
Paradigms of recommender systems

Personalized recommendations



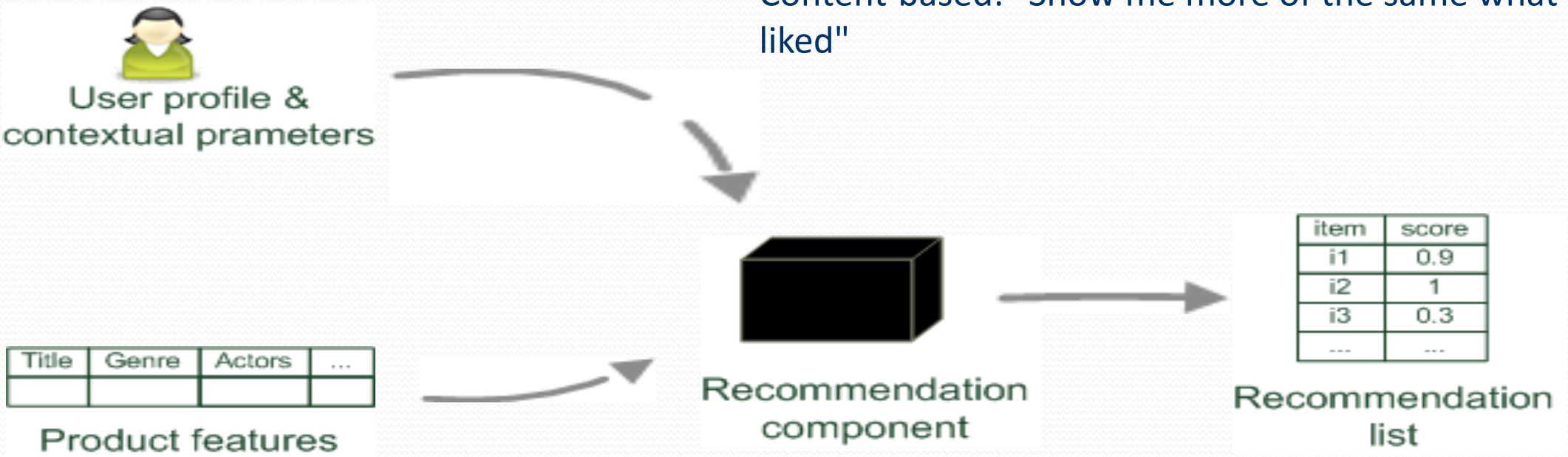
Paradigms of recommender systems

Collaborative: "Tell me what's popular among my peers"



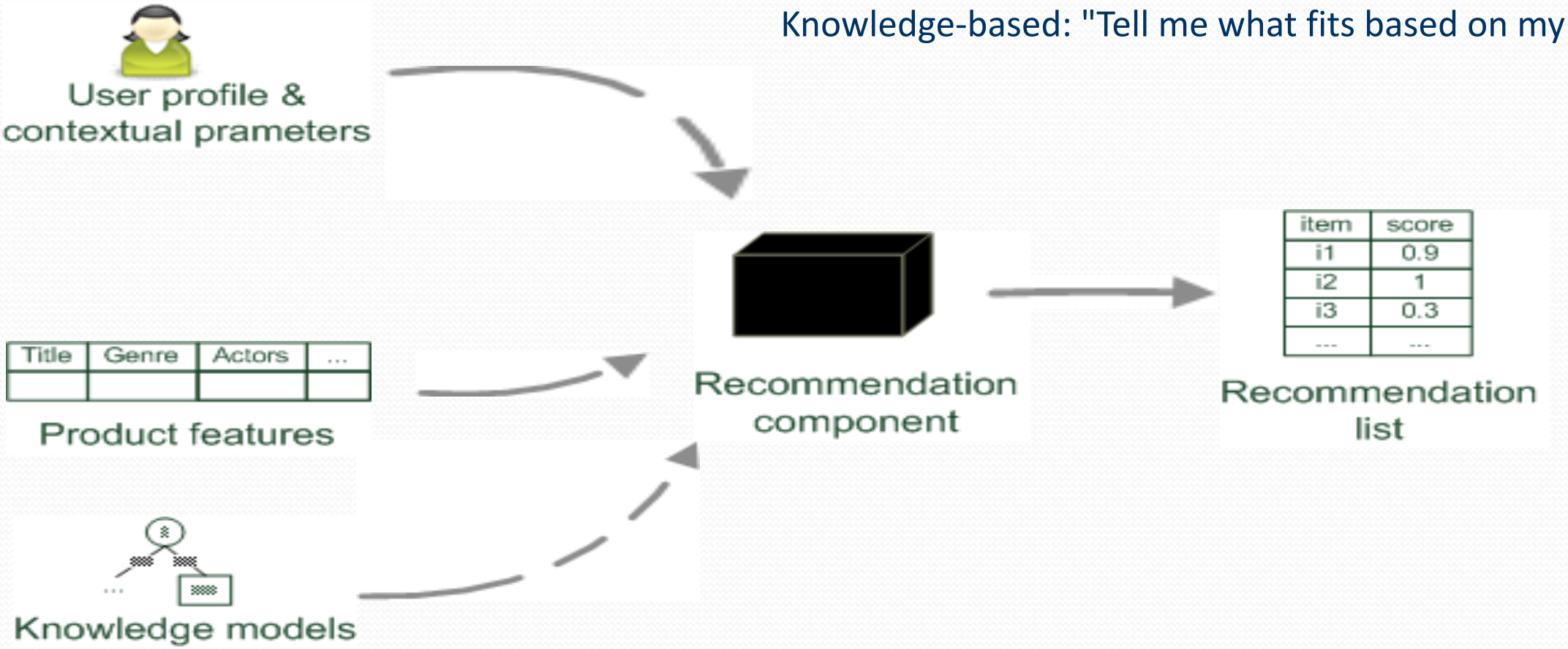
Paradigms of recommender systems

Content-based: "Show me more of the same what I've liked"



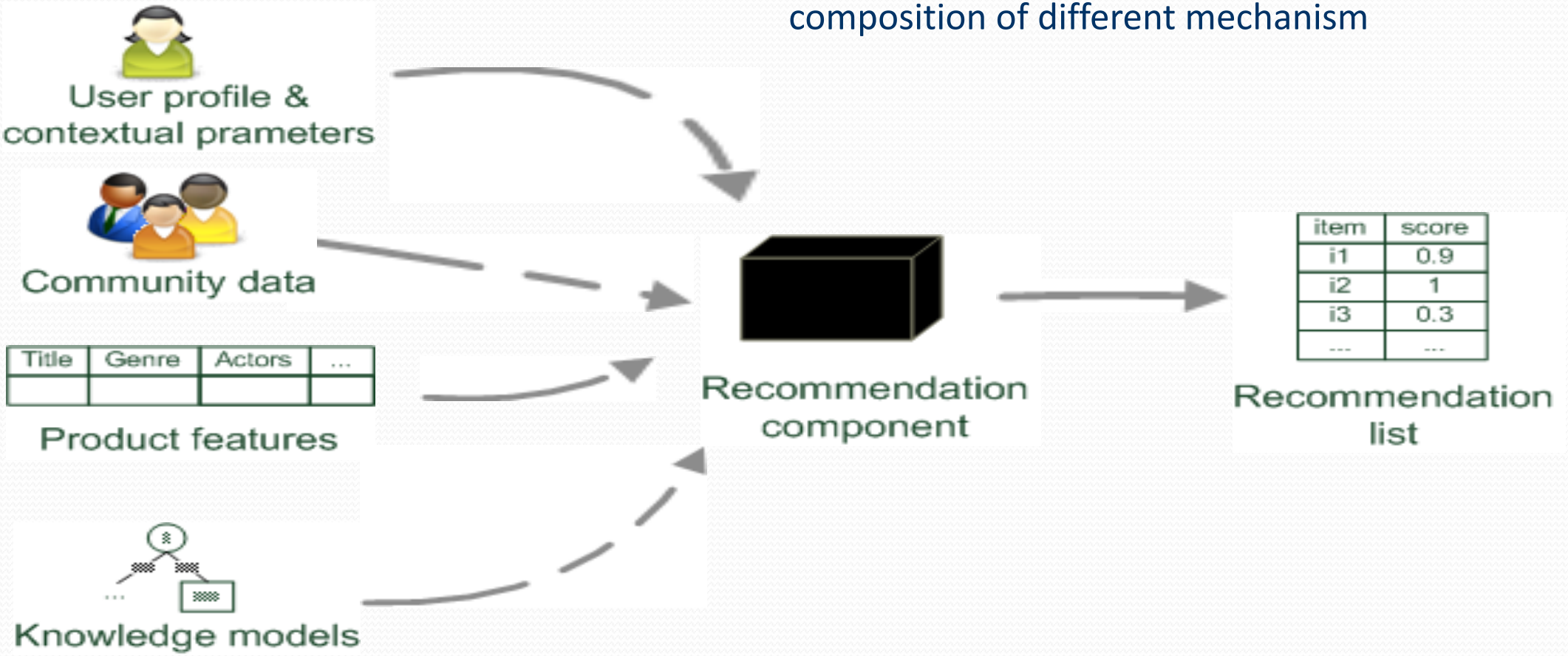
Paradigms of recommender systems

Knowledge-based: "Tell me what fits based on my needs"



Paradigms of recommender systems

Hybrid: combinations of various inputs and/or composition of different mechanism



Content-based Recommendation Systems

- Content-based Recommendation systems
  - Build profile of user's interest - explicitly or implicitly
  - Check similarity of contents (items) with the user interest profile
  - Much like a search process where the profile acts as the query
- Pros
  - Requirements explicit
  - Precise recommendation is possible – directly related contents
- Cons
  - Difficulty of keeping user profiles
    - Creating and updating
  - Difficulty of designing profiles
  - Limited recommendation scope
  - Not adaptive for new contents

Required Information

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
  - *Demographic data: age, gender, salary, profession, country of residence, country of origin, religion ...*
  - *Site behaviour: Purchase history at the site; viewing history, perhaps including time spent on certain pages/items; clickstream sequence*

Collaborative Filtering

- Recommending items/information to a user by collecting preferences of many users (collaborating)

- Amazon.com

How Collaborative Filtering Works?

- How to estimate missing information from the given matrix?
  - Each vector represents preference of each person
  - Some values are missing because she has not experience them
  - Estimate these values
- Solution: Use similarity between users

Two ways
•User-based collaborative filtering
•Item-based collaborative filtering

- **User-based:** Recommend things that were purchased or viewed by users who are ***similar to you.***
- **Item-based:** Recommend things that are ***similar to the items that you have viewed/ purchased before***

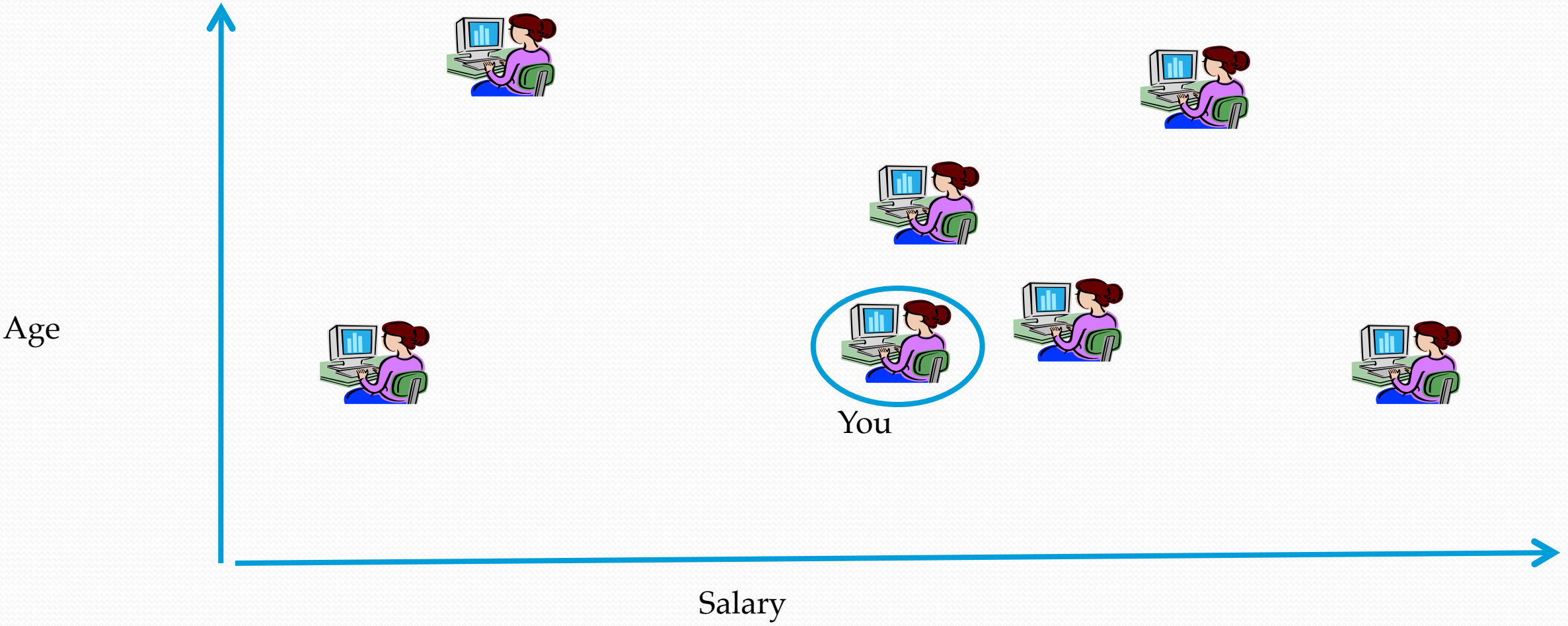
Ratings -See next slide

Article	Person A	Person B	Person C	Person D
1	1	4	2	2
2	5	2	4	4
3			3	
4	2	5		5
5	4	1		1
6	?	2	5	

How Collaborative Filtering Works?

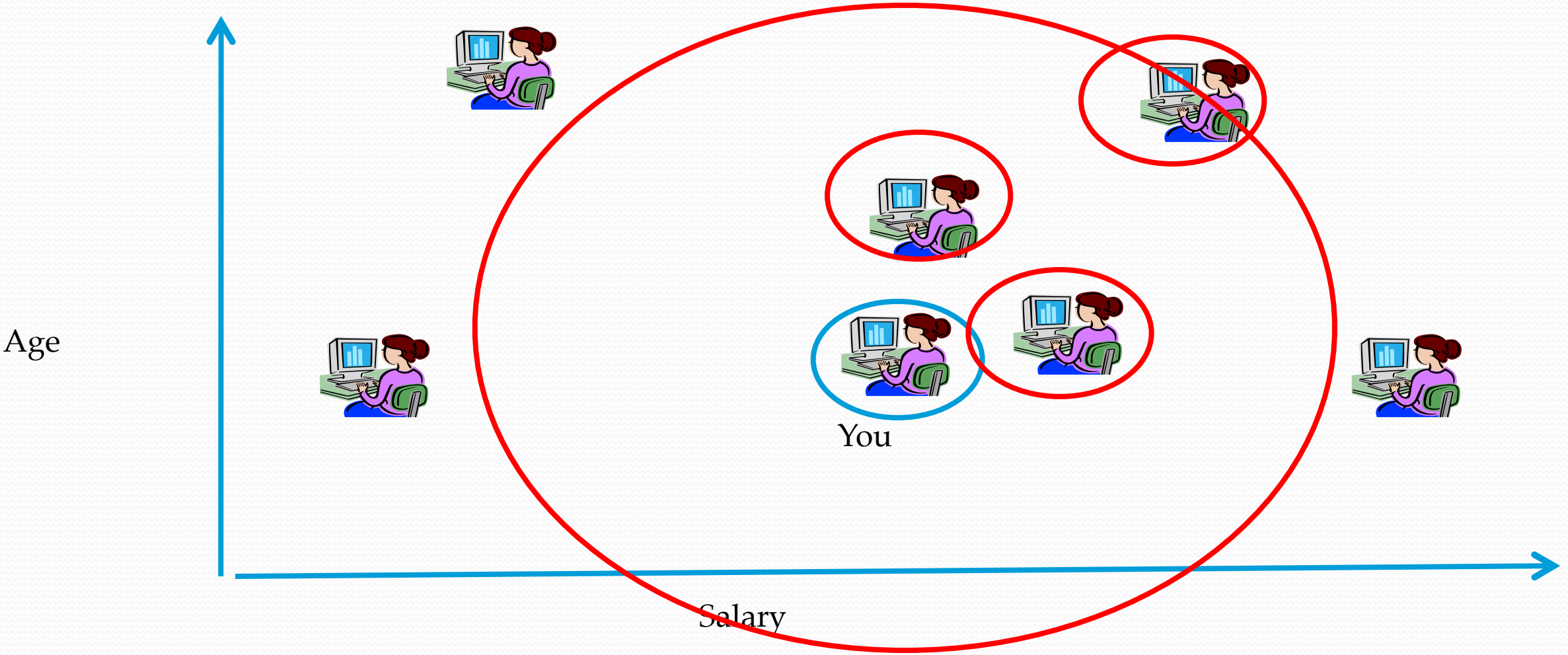
- Users rate items – user interests recorded.
- Ratings may be:
  - Explicit, e.g. buying or rating an item
  - Implicit, e.g. browsing time, no. of mouse clicks
- **Nearest neighbour** matching used to find people with similar interests
- Items that neighbours rate highly but that you have not rated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

K-Nearest Neighbour



(Think in terms of many dimensions, not just these two)

K-Nearest Neighbour



Your neighbours: recommend things that they have viewed / purchased

Ratings on 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).

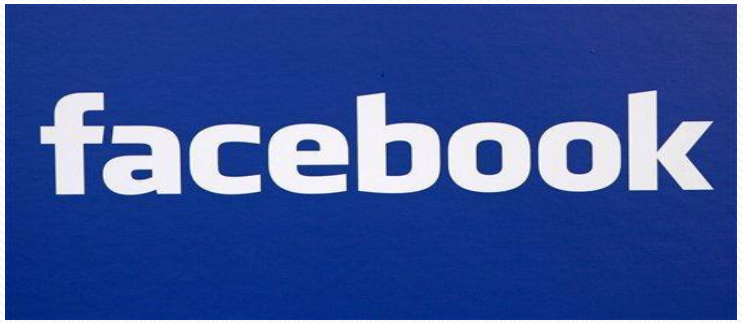
Collaborative Filtering: Pros and Cons

- Pros
  - Robust for content change
    - No need for content analysis
    - Applicable for non-text data as well
  - Minimal user input needed
    - Just view/evaluate items
- Cons
  - "Cold start" problem
    - Lot of evaluation data is needed before reliable recommendation
  - No evaluation, no recommendation
    - Items without evaluation / new items are never recommended

Collective Intelligence

- **Emergence of new ideas as a result of a group effort** (usually a huge group)
- **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!

Sites like Facebook, Flickr, Wikipedia and YouTube could not exist without their users.



Web 2.0 - Collective Intelligence

- Blogs - Blogosphere
- Wikis – Wikipedia
- Social networking – Facebook
- Social Bookmarking – delicious, digg ..
- Community question answering (CQA) – answers.yahoo.com
- Social multimedia sharing – YouTube, last.fm, flickr ….
- Tagging
- ***The social networking world*** *is perhaps the most popular of collective intelligence. Friend post status 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.*

Other Examples

- 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

Next Class:

Chapter-8: Scalable and Emerging information System techniques