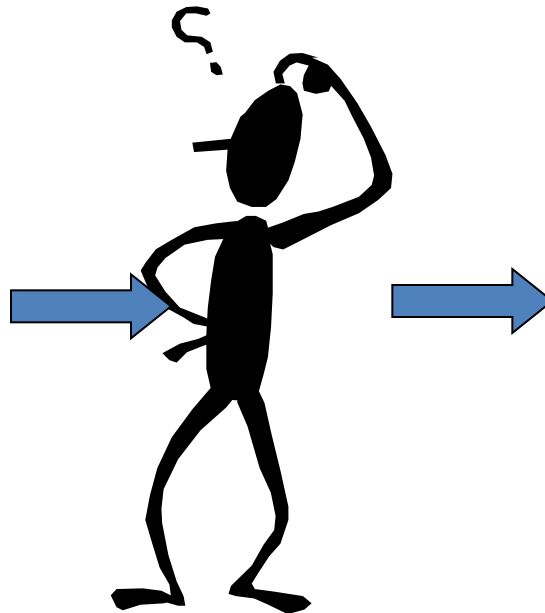
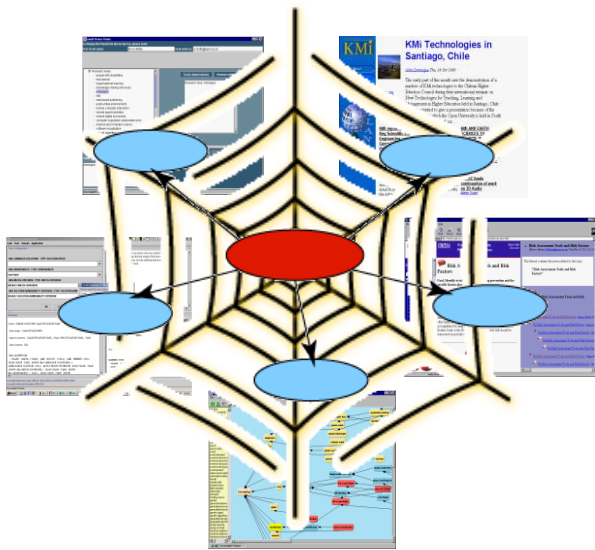


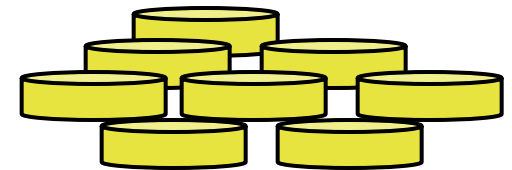
Web Mining

Discovering Knowledge from and about WWW - is one of the basic abilities of an intelligent agent

WWW



Knowledge



Contents

- Introduction
- Web content mining
- Web structure mining
 - Evaluation of Web pages
 - HITS algorithm
 - Discovering cyber-communities on the Web
- Web usage mining
- Search engines for Web mining
- Multi-Layered Meta Web

Introduction

Data Mining and Web Mining

- Data mining: turn data into knowledge.
- Web mining is to apply data mining techniques to extract and uncover knowledge from *web documents* and *services*.

WWW Specifics

- Web: A huge, widely-distributed, highly heterogeneous, semi-structured, hypertext/hypermedia, interconnected information repository
- Web is a huge collection of documents plus
 - Hyper-link information
 - Access and usage information

A Few Themes in Web Mining

- Some interesting problems on Web mining
 - Mining what Web search engine finds
 - Identification of authoritative Web pages
 - Identification of Web communities
 - Web document classification
 - Warehousing a Meta-Web: Web yellow page service
 - Weblog mining (usage, access, and evolution)
 - Intelligent query answering in Web search

Web Mining taxonomy

- Web Content Mining
 - Web Page Content Mining
- Web Structure Mining
 - Search Result Mining
 - Capturing Web's structure using link interconnections
- Web Usage Mining
 - General Access Pattern Mining
 - Customized Usage Tracking

Web Content Mining

What is text mining?

- Data mining in text: find something useful and surprising from a text collection;
- text mining vs. information retrieval;
- data mining vs. database queries.

Types of text mining

- Keyword (or term) based association analysis
- automatic document (topic) classification
- similarity detection
 - cluster documents by a common author
 - cluster documents containing information from a common source
- sequence analysis: predicting a recurring event, discovering trends
- anomaly detection: find information that violates usual patterns

Types of text mining (cont.)

- discovery of frequent phrases
- text segmentation (into logical chunks)
- event detection and tracking

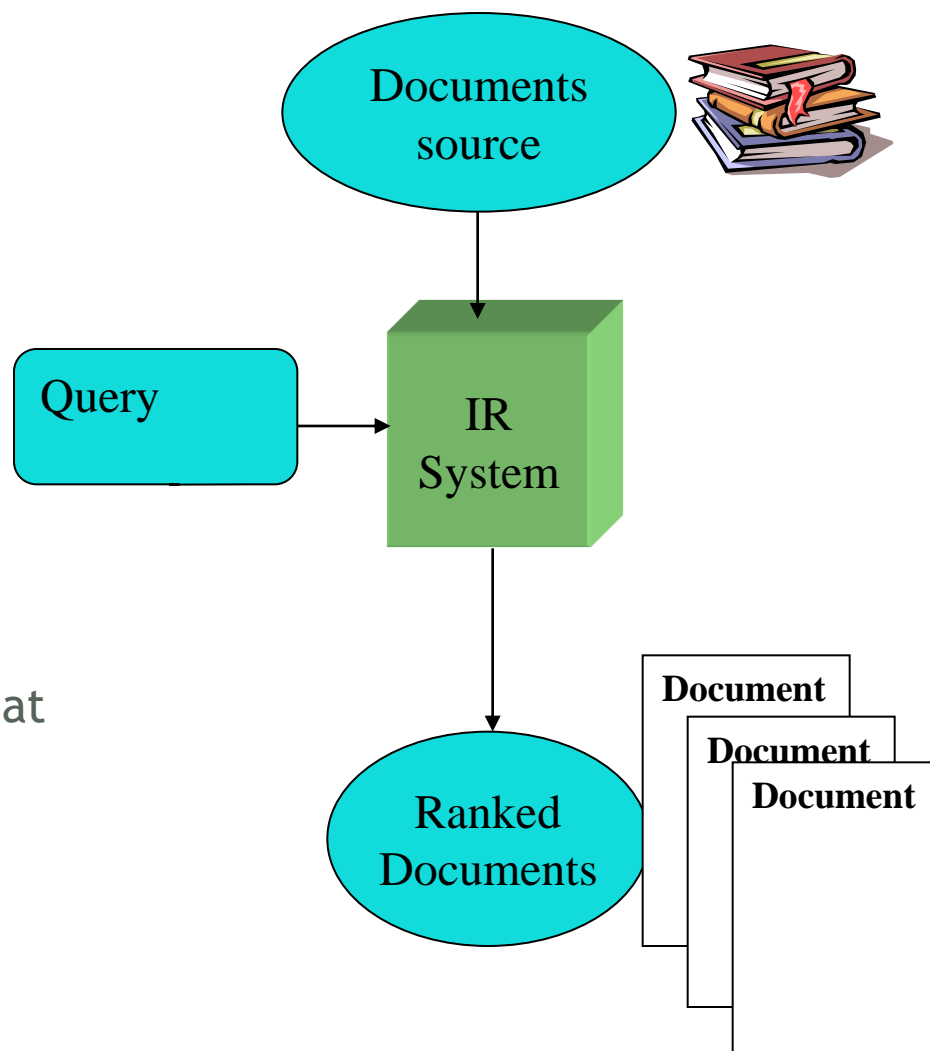
Information Retrieval

■ Given:

- A source of textual documents
- A user query (text based)

● Find:

- A set (ranked) of documents that are relevant to the query



Intelligent Information Retrieval

- meaning of words
 - Synonyms “buy” / “purchase”
 - Ambiguity “bat” (baseball vs. mammal)
- order of words in the query
 - hot dog stand in the amusement park
 - hot amusement stand in the dog park
- user dependency for the data
 - direct feedback
 - indirect feedback
- authority of the source
 - IBM is more likely to be an authorized source than my second far cousin

Intelligent Web Search

- Combine the intelligent IR tools
 - **meaning** of words
 - **order** of words in the query
 - **user dependency** for the data
 - **authority** of the source
- With the unique web features
 - retrieve Hyper-link information
 - utilize Hyper-link as input

What is Information Extraction?

■ **Given:**

- A source of textual documents
- A well defined limited query (text based)

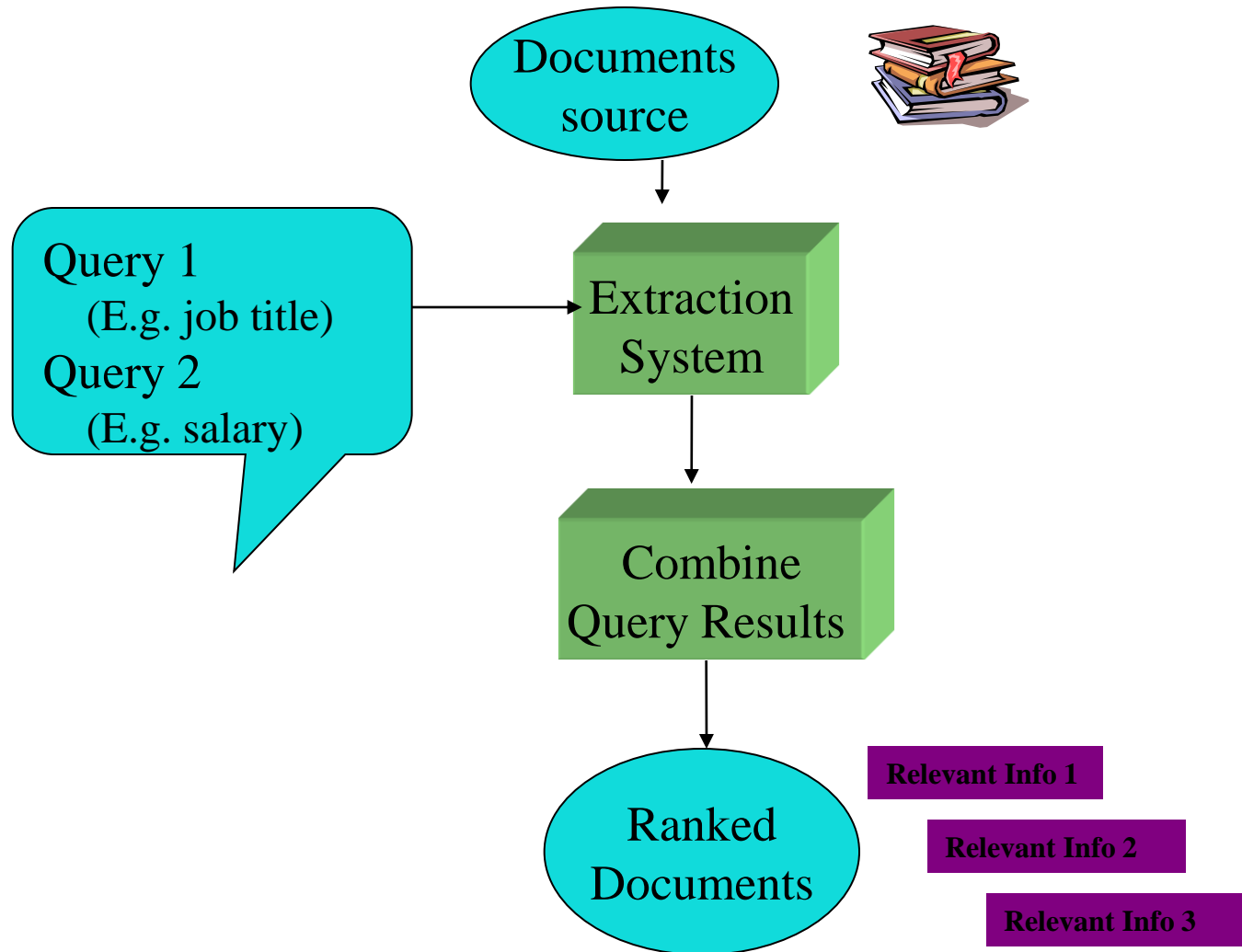
■ **Find:**

- Sentences with **relevant** information
- Extract the relevant information and ignore non-relevant information (important!)
- Link related information and output in a predetermined format

Information Extraction: Example

- Salvadoran President-elect Alfredo Cristiana condemned the terrorist killing of Attorney General Roberto Garcia Alvarado and accused the Farabundo Marti Natinal Liberation Front (FMLN) of the crime. ... Garcia Alvarado, 56, was killed when a bomb placed by urban guerillas on his vehicle exploded as it came to a halt at an intersection in downtown San Salvador. ... According to the police and Garcia Alvarado's driver, who escaped unscathed, the attorney general was traveling with two bodyguards. One of them was injured.
- **Incident Date:** 19 Apr 89
- **Incident Type:** Bombing
- **Perpetrator Individual ID:** “urban guerillas”
- **Human Target Name:** “Roberto Garcia Alvarado”
- ...

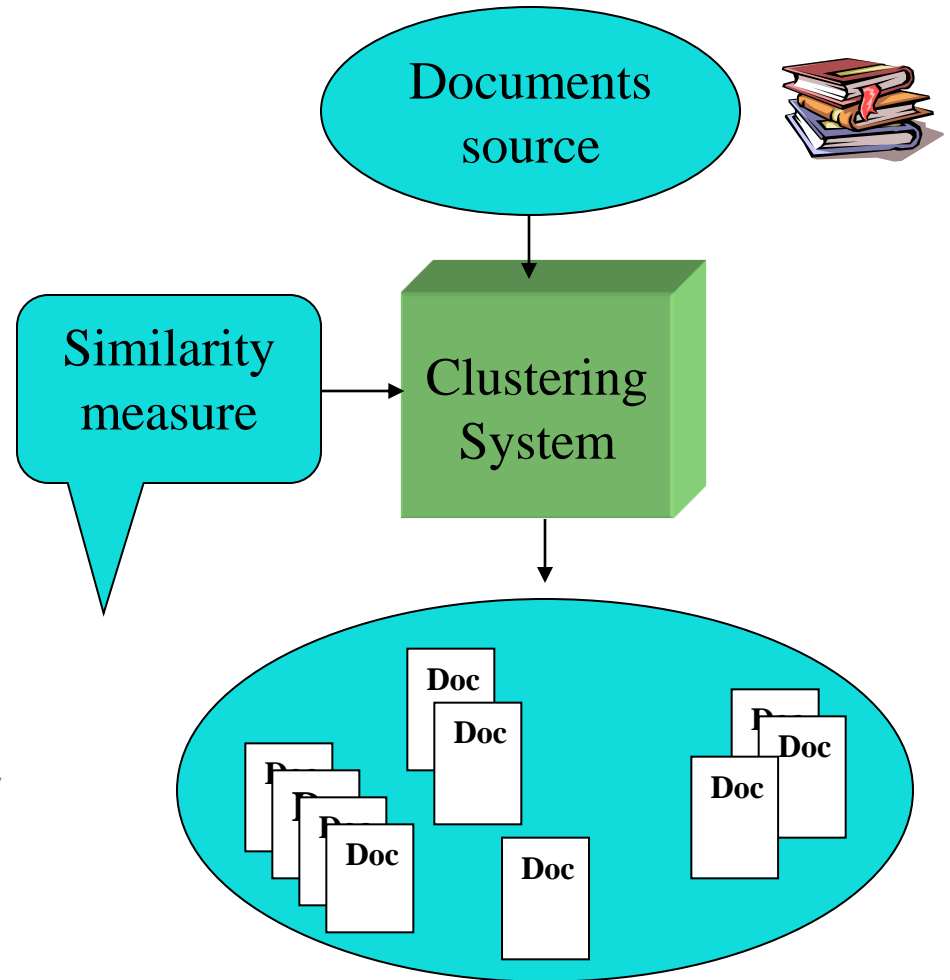
Querying Extracted Information



What is Clustering ?

■ Given:

- A source of textual documents
- Similarity measure
 - e.g., how many words are common in these documents
- Find:
 - Several clusters of documents that are **relevant** to each other



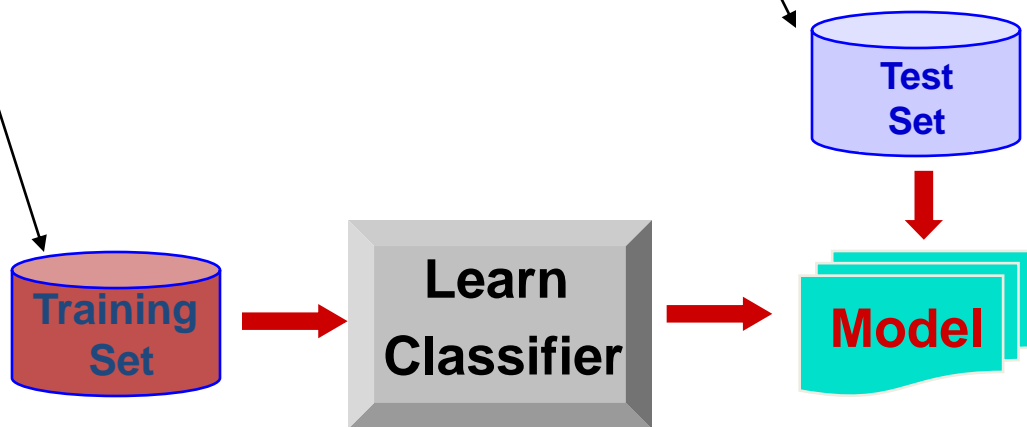
Text Classification definition

- **Given:** a collection of labeled records (*training set*)
 - Each record contains a set of features (*attributes*), and the true class (*label*)
- **Find:** a **model** for the class as a function of the values of the features
- **Goal:** previously unseen records should be assigned a class as accurately as possible
 - A **test set** is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it

Text Classification: An Example

Ex#	text	class
1	An English football fan ...	Yes
2	During a game in Italy ...	Yes
3	England has been beating France ...	Yes
4	Italian football fans were cheering ...	No
5	An average USA salesman earns 75K	No
6	The game in London was horrific	Yes
7	Manchester city is likely to win the championship	Yes
8	Rome is taking the lead in the football league	Yes

Hooligan	
A Danish football fan	?
Turkey is playing vs. France. The Turkish fans ...	?



Discovery of frequent sequences (1)

- Find all frequent maximal sequences of words (=phrases) from a collection of documents
 - frequent: frequency threshold is given; e.g. a phrase has to occur in at least 15 documents
 - maximal: a phrase is not included in another longer frequent phrase
 - other words are allowed between the words of a sequence in text

Discovery of frequent sequences (2)

- Frequency of a sequence cannot be decided locally: all the instances in the collection has to be counted
- however: already a document of length 20 contains over million sequences
- only small fraction of sequences are frequent

Basic idea: bottom-up

- 1. Collect all pairs from the documents, count them, and select the frequent ones
- 2. Build sequences of length $p + 1$ from frequent sequences of length p
- 3. Select sequences that are frequent
- 4. Select maximal sequences

Summary

- There are many **scientific and statistical text mining methods** developed, see e.g.:
 - <http://www.cs.utexas.edu/users/pebronia/text-mining/>
 - http://filebox.vt.edu/users/wfan/text_mining.html
- Also, it is important to study **theoretical foundations** of data mining.
 - **Data Mining Concepts and Techniques / J.Han & M.Kamber**
 - **Machine Learning, / T.Mitchell**

Web Structure Mining

Web Structure Mining

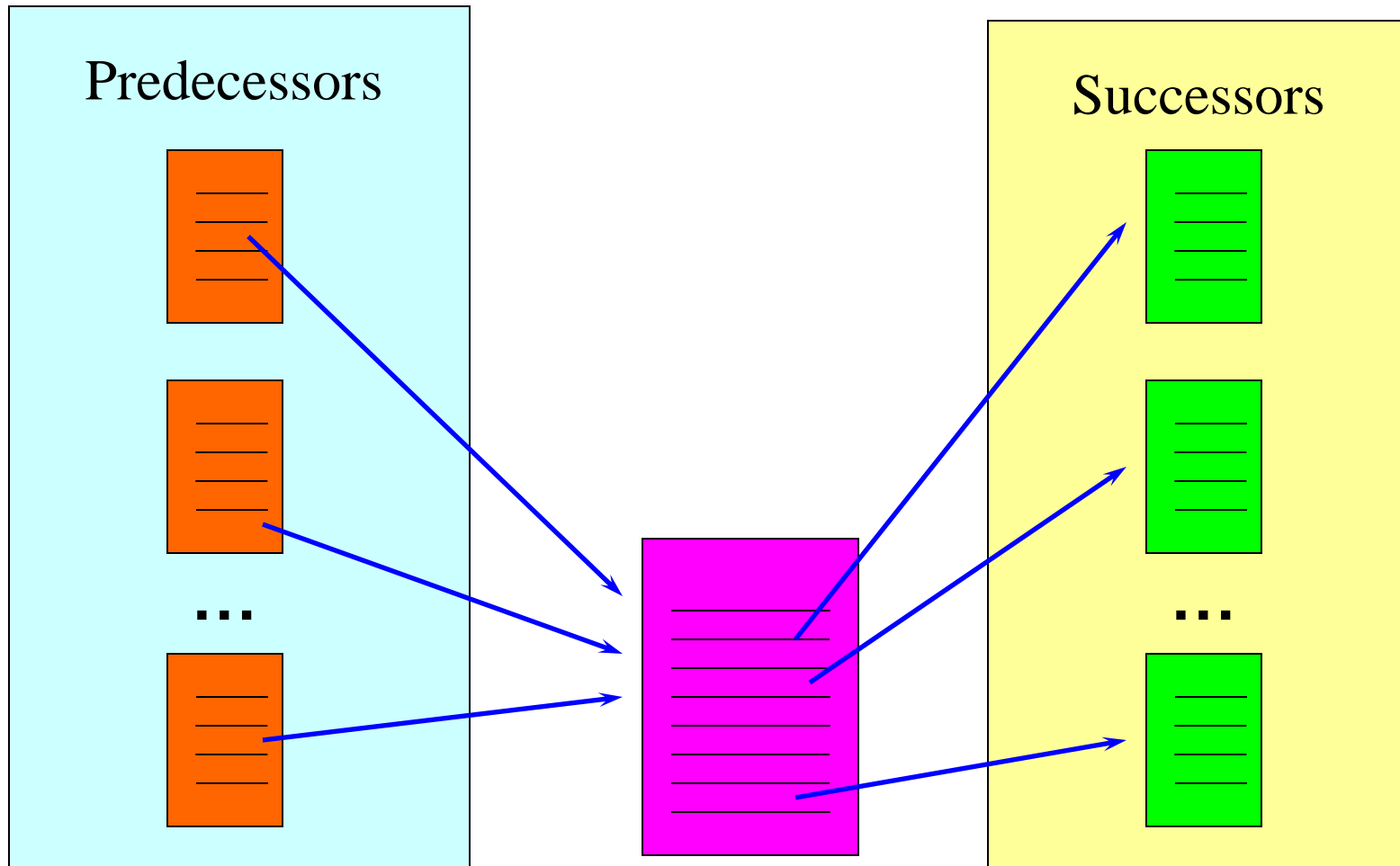
- (1970) Researchers proposed methods of using citations among journal articles to evaluate the quality of research papers.
- Customer behavior – evaluate a quality of a product based on the opinions of other customers (instead of product's description or advertisement)
- Unlike journal citations, the Web linkage has some unique features:
 - not every hyperlink represents the endorsement we seek
 - one authority page will seldom have its Web page point to its competitive authorities (CocaCola → Pepsi)
 - authoritative pages are seldom descriptive (Yahoo! may not contain the description „Web search engine”)

Evaluation of Web pages

Web Search

- There are two approaches:
 - **page rank**: for discovering the most important pages on the Web (as used in Google)
 - **hubs and authorities**: a more detailed evaluation of the importance of Web pages
- Basic definition of importance:
 - A page is important if important pages link to it

Predecessors and Successors of a Web Page



Page Rank (1)

Simple solution: create a stochastic matrix of the Web:

- Each page i corresponds to row i and column i of the matrix
- If page j has n successors (links) then the ij^{th} cell of the matrix is equal to $1/n$ if page i is one of these n successors of page j , and 0 otherwise.

Page Rank (2)

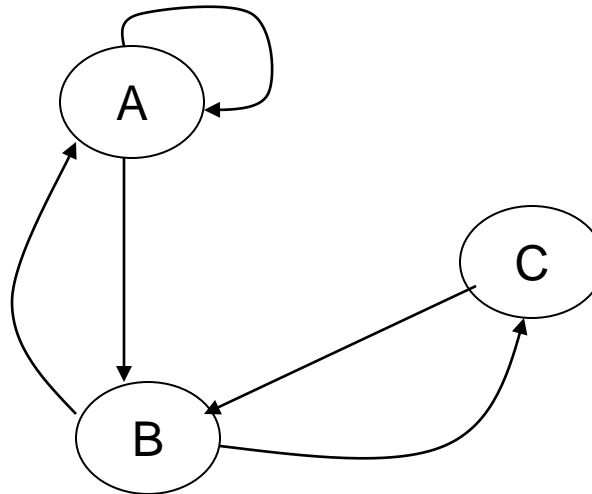
The intuition behind this matrix:

- initially each page has 1 unit of importance. At each round, each page shares importance it has among its successors, and receives new importance from its predecessors.
- The importance of each page reaches a limit after some steps
- That importance is also the probability that a Web surfer, starting at a random page, and following random links from each page will be at the page in question after a long series of links.

Page Rank (3) – Example 1

- Assume that the Web consists of only three pages - A, B, and C. The links among these pages are shown below.

Let $[a, b, c]$ be the vector of importances for these three pages



	A	B	C
A	$1/2$	$1/2$	0
B	$1/2$	0	1
C	0	$1/2$	0

Page Rank – Example 1 (cont.)

- The equation describing the asymptotic values of these three variables is:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \times \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

We can solve the equations like this one by starting with the assumption $a = b = c = 1$, and applying the matrix to the current estimate of these values repeatedly. The first four iterations give the following estimates:

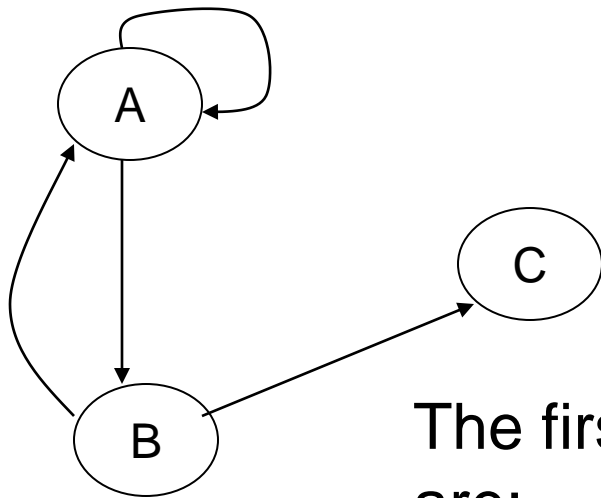
a =	1	1	5/4	9/8	5/4	...	6/5
b =	1	3/2	1	11/8	17/16	...	6/5
c =	1	1/2	3/4	1/2	11/16	...	3/5

Problems with Real Web Graphs

- In the limit, the solution is $a=b=6/5$, $c=3/5$. That is, a and b each have the same importance, and twice of c .
- **Problems with Real Web Graphs**
 - **dead ends**: a page that has no successors has nowhere to send its importance.
 - **spider traps**: a group of one or more pages that have no links out.

Page Rank – Example 2

- Assume now that the structure of the Web has changed. The new matrix describing transitions is:



$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

The first four steps of the iterative solution are:

$$a = 1 \quad 1 \quad 3/4 \quad 5/8 \quad 1/2$$

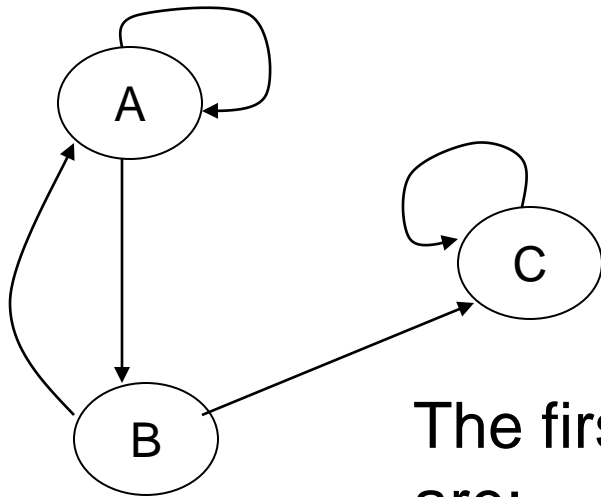
$$b = 1 \quad 1/2 \quad 1/2 \quad 3/8 \quad 5/16$$

$$c = 1 \quad 1/2 \quad 1/4 \quad 1/4 \quad 3/16$$

Eventually, each of a, b, and c become 0.

Page Rank – Example 3

- Assume now once more that the structure of the Web has changed. The new matrix describing transitions is:



$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \\ 0 & 1 & \frac{1}{2} \end{bmatrix} \times \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

The first four steps of the iterative solution are:

$$\begin{array}{l} a = 1 \quad 1 \quad 3/4 \quad 5/8 \quad 1/2 \\ b = 1 \quad 1/2 \quad 1/2 \quad 3/8 \quad 5/16 \\ c = 1 \quad 3/2 \quad 7/4 \quad 2 \quad 35/16 \end{array}$$

c converges to 3, and a=b=0.

Google Solution

- Instead of applying the matrix directly, „tax” each page some fraction of its current importance, and distribute the taxed importance equally among all pages.
- Example: if we use 20% tax, the equation of the previous example becomes:

$$a = 0.8 * (\frac{1}{2} * a + \frac{1}{2} * b + 0 * c)$$

$$b = 0.8 * (\frac{1}{2} * a + 0 * b + 0 * c)$$

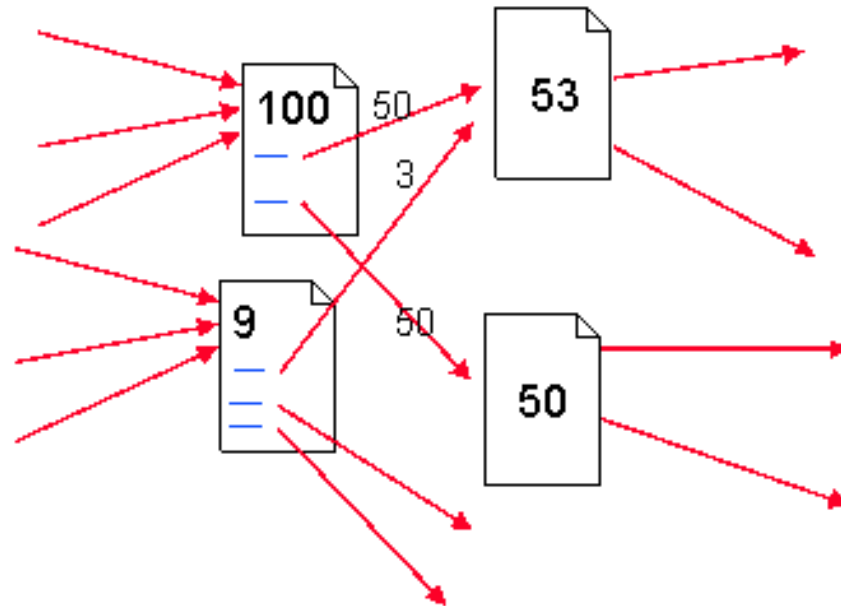
$$c = 0.8 * (0 * a + \frac{1}{2} * b + 1 * c)$$

The solution to this equation is $a=7/11$, $b=5/11$, and $c=21/11$

Google Anti-Spam Solution

- „Spamming” is the attempt by many Web sites to appear to be about a subject that will attract surfers, without truly being about that subject.
- Solutions:
 - Google tries to match words in your query to the words on the Web pages. Unlike other search engines, Google tends to believe what others say about you in their anchor text, making it harder for you to appear to be about something you are not.
 - The use of Page Rank to measure importance also protects against spammers. The naive measure (number of links into the page) can easily be fooled by the spammers who create 1000 pages that mutually link to one another, while Page Rank recognizes that none of the pages have any real importance.

PageRank Calculation



HITS Algorithm

--Topic Distillation on WWW

- Proposed by Jon M. Kleinberg
- **H**yperlink-**I**nduced **T**opic **S**earch

Key Definitions

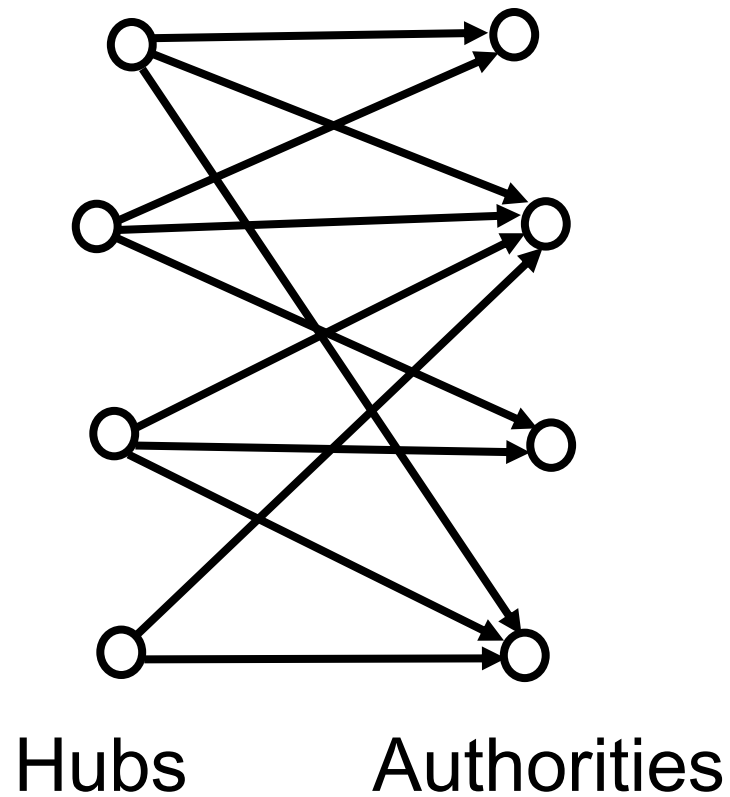
- **Authorities**

Relevant pages of the highest quality on a broad topic

- **Hubs**

Pages that link to a collection of authoritative pages on a broad topic

Hub-Authority Relations



Hyperlink-Induced Topic Search (HITS)

The approach consists of two phases:

- It uses the query terms to collect a starting set of pages (200 pages) from an index-based search engine – **root set of pages**.
- The root set is expanded into a **base set** by including all the pages that the root set pages link to, and all the pages that link to a page in the root set, up to a designed size cutoff, such as 2000-5000.
- A weight-propagation phase is initiated. This is an iterative process that determines numerical estimates of hub and authority weights

Hub and Authorities

- Define a matrix \mathbf{A} whose rows and columns correspond to Web pages with entry $\mathbf{A}_{ij}=1$ if page i links to page j , and 0 if not.
- Let \mathbf{a} and \mathbf{h} be vectors, whose i^{th} component corresponds to the degrees of authority and hubbiness of the i^{th} page. Then:
 - $\mathbf{h} = \mathbf{A} \times \mathbf{a}$. That is, the hubbiness of each page is the sum of the authorities of all the pages it links to.
 - $\mathbf{a} = \mathbf{A}^T \times \mathbf{h}$. That is, the authority of each page is the sum of the hubbiness of all the pages that link to it (\mathbf{A}^T - transposed matrix).

Then, $\mathbf{a} = \mathbf{A}^T \times \mathbf{A} \times \mathbf{a}$ $\mathbf{h} = \mathbf{A} \times \mathbf{A}^T \times \mathbf{h}$

Hub and Authorities - Example

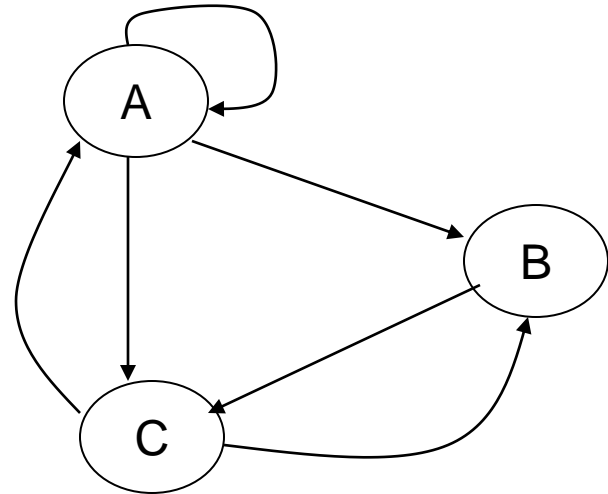
Consider the Web presented below.

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

$$A^T = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

$$AA^T = \begin{bmatrix} 3 & 1 & 2 \\ 1 & 1 & 0 \\ 2 & 0 & 2 \end{bmatrix}$$

$$A^T A = \begin{bmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$



Hub and Authorities - Example

If we assume that the vectors

$h = [h_a, h_b, h_c]$ and $a = [a_a, a_b, a_c]$ are each initially $[1,1,1]$, the first three iterations of the equations for a and h are the following:

$$a_a = 1 \quad 5 \quad 24 \quad 114$$

$$a_b = 1 \quad 5 \quad 24 \quad 114$$

$$a_c = 1 \quad 4 \quad 18 \quad 84$$

$$h_a = 1 \quad 6 \quad 28 \quad 132$$

$$h_b = 1 \quad 2 \quad 8 \quad 36$$

$$h_c = 1 \quad 4 \quad 20 \quad 96$$

Discovering cyber-communities on the web

Based on link structure

What is cyber-community

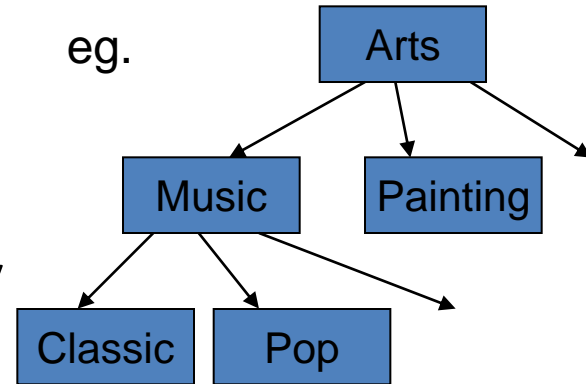
- Defn: a *community on the web* is a group of web pages sharing a common interest
 - Eg. A group of web pages talking about POP Music
 - Eg. A group of web pages interested in data-mining
- Main properties:
 - Pages in the same community should be similar to each other in contents
 - The pages in one community should differ from the pages in another community
 - Similar to cluster

Recursive Web Communities

- **Definition:** A *community* consists of members that have more links within the community than outside of the community.
- Community identification is NP-complete task

Two different types of communities

- Explicitly-defined communities
 - They are well known ones, such as the resource listed by Yahoo!



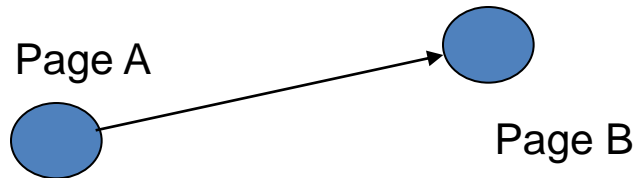
- Implicitly-defined communities
 - They are communities unexpected or invisible to most users

eg. The group of web pages interested in a particular singer

Similarity of web pages

- Discovering web communities is similar to clustering. For clustering, we must define the similarity of two nodes
- A Method I:
 - For page A and page B, A is related to B if there is a hyper-link from A to B, or from B to A

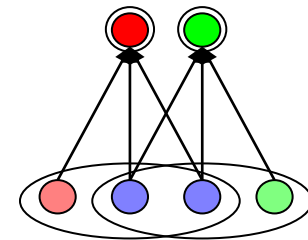
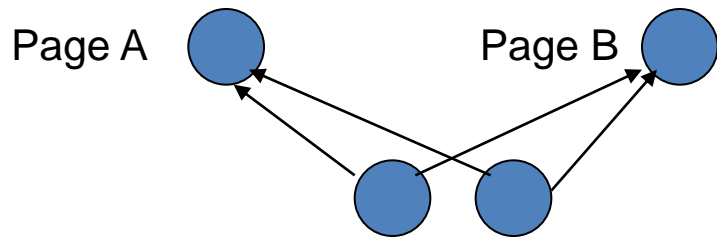
– Not so good
Microsoft.



f IBM and

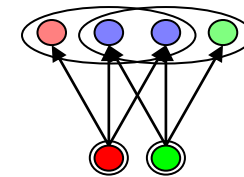
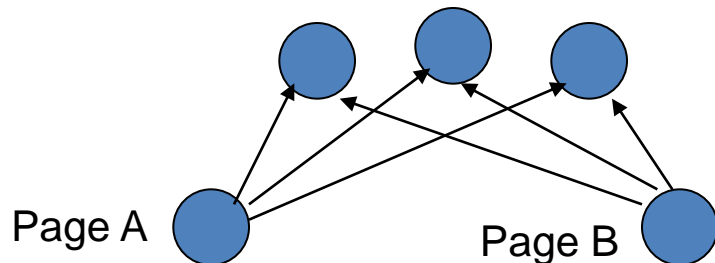
Similarity of web pages

- Method II (from Bibliometrics)
 - **Co-citation:** the similarity of A and B is measured by the number of pages cite both A and B



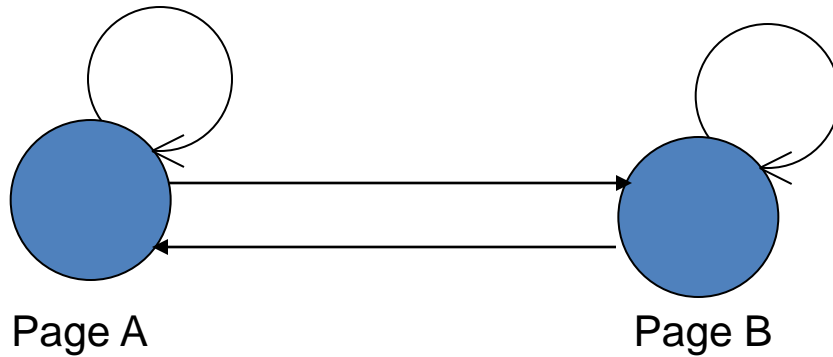
The normalized degree of overlap in inbound links

- **Bibliographic coupling:** the similarity of A and B is measured by the number of pages cited by both A and B.

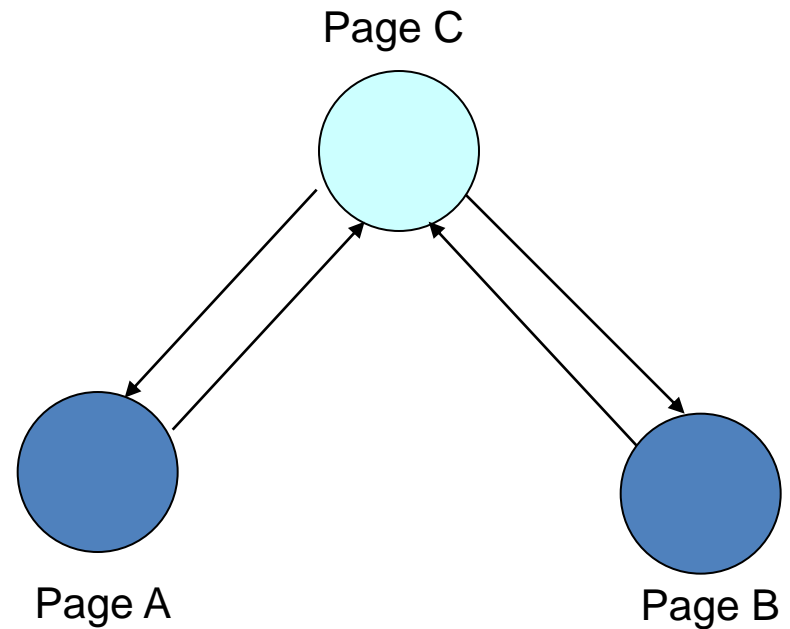


The normalized degree of overlap in outbound links

Simple Cases (co-citations and coupling)



Better not to account self-citations



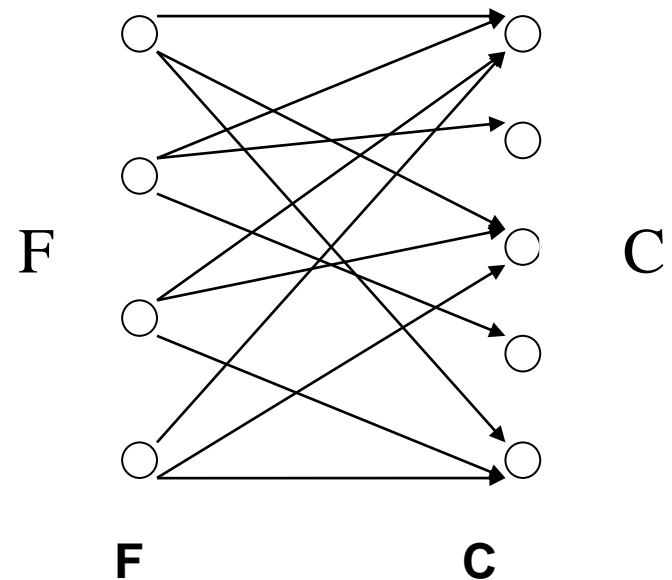
Number of pages for similarity decision should be big enough

Example method of clustering

- The method from R. Kumar, P. Raghavan, S. Rajagopalan, Andrew Tomkins
 - IBM Almaden Research Center
- They call their method ***communities trawling (CT)***
- They implemented it on the graph of 200 millions pages, it worked very well

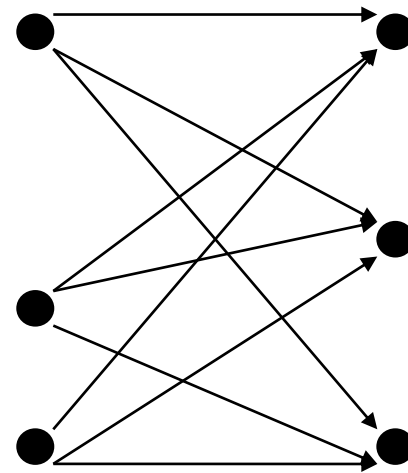
Basic idea of CT

- **Bipartite graph**: Nodes are partitioned into two sets, **F** and **C**
- Every directed edge in the graph is directed from a node in **F** to a node in **C**



Basic idea of CT

- Definition **Bipartite cores**
 - a complete bipartite subgraph with at least i nodes from **F** and at least j nodes from **C**
 - i and j are tunable parameters
 - $A(i, j)$ Bipartite core



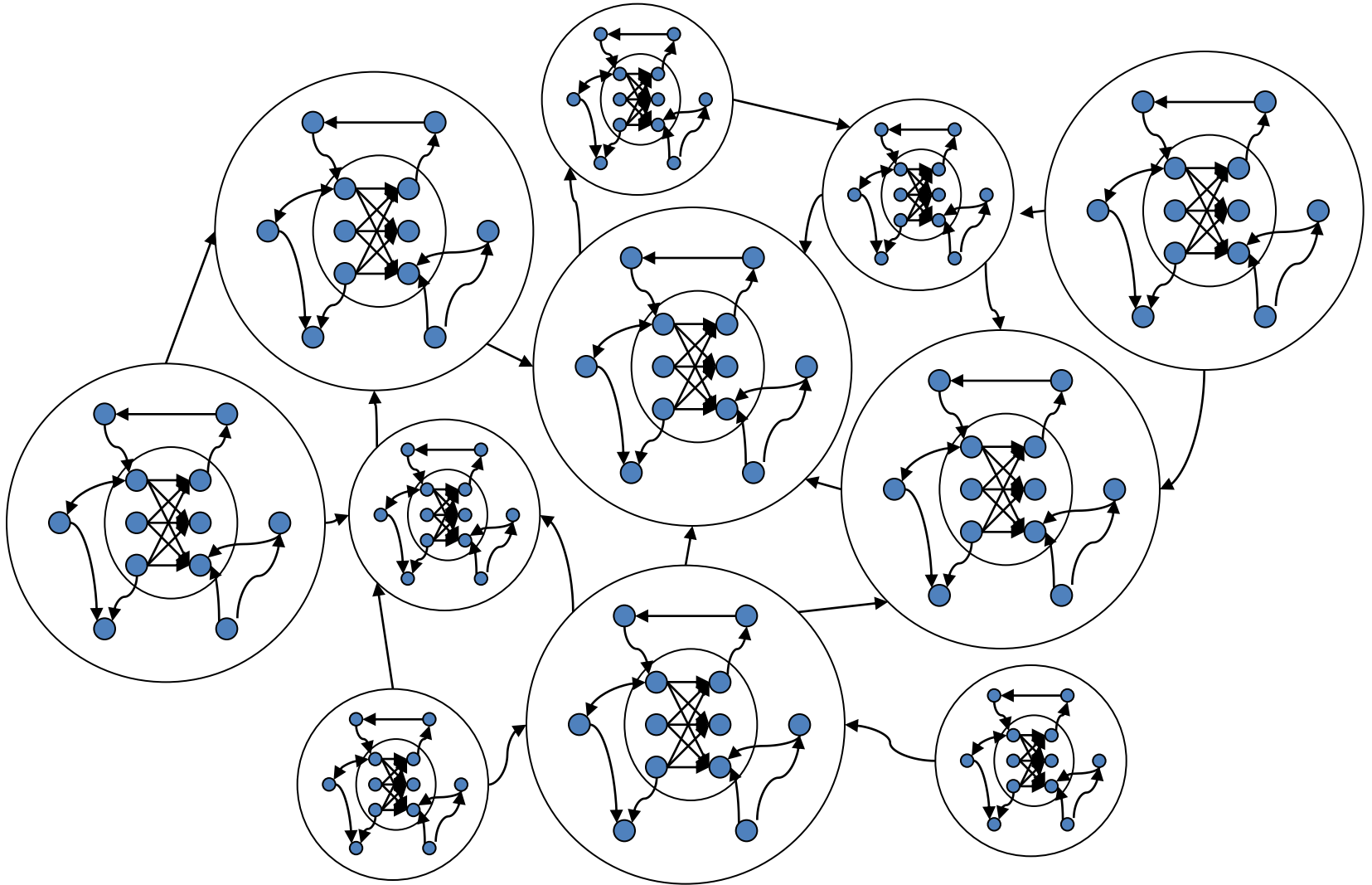
$A(i=3, j=3)$ bipartite core

- Every community have such a core with a certain i and j

Basic idea of CT

- A bipartite core is the identity of a community
- To extract all the communities is to enumerate all the bipartite cores on the web

Web Communities



Web Usage Mining

What is Web Usage Mining?

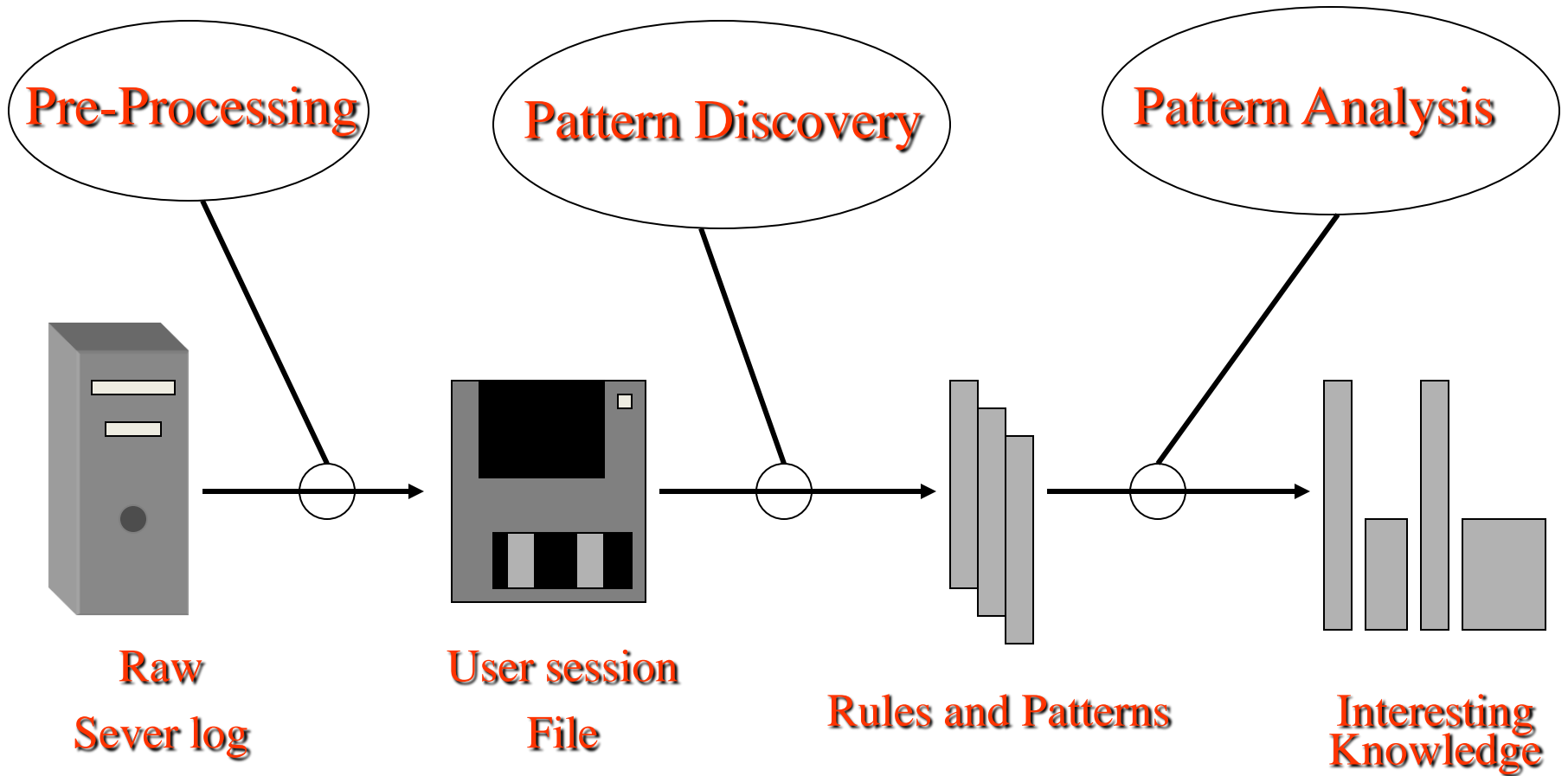
- A *Web* is a collection of inter-related files on one or more *Web servers*.
- *Web Usage Mining*.
 - ➔ Discovery of meaningful patterns from data generated by client-server transactions.
- Typical Sources of Data:
 - ➔ automatically generated data stored in server *access logs*, *referrer logs*, *agent logs*, and client-side *cookies*.
 - ➔ user profiles.
 - ➔ metadata: page attributes, content attributes, usage data.

Web Usage Mining (WUM)

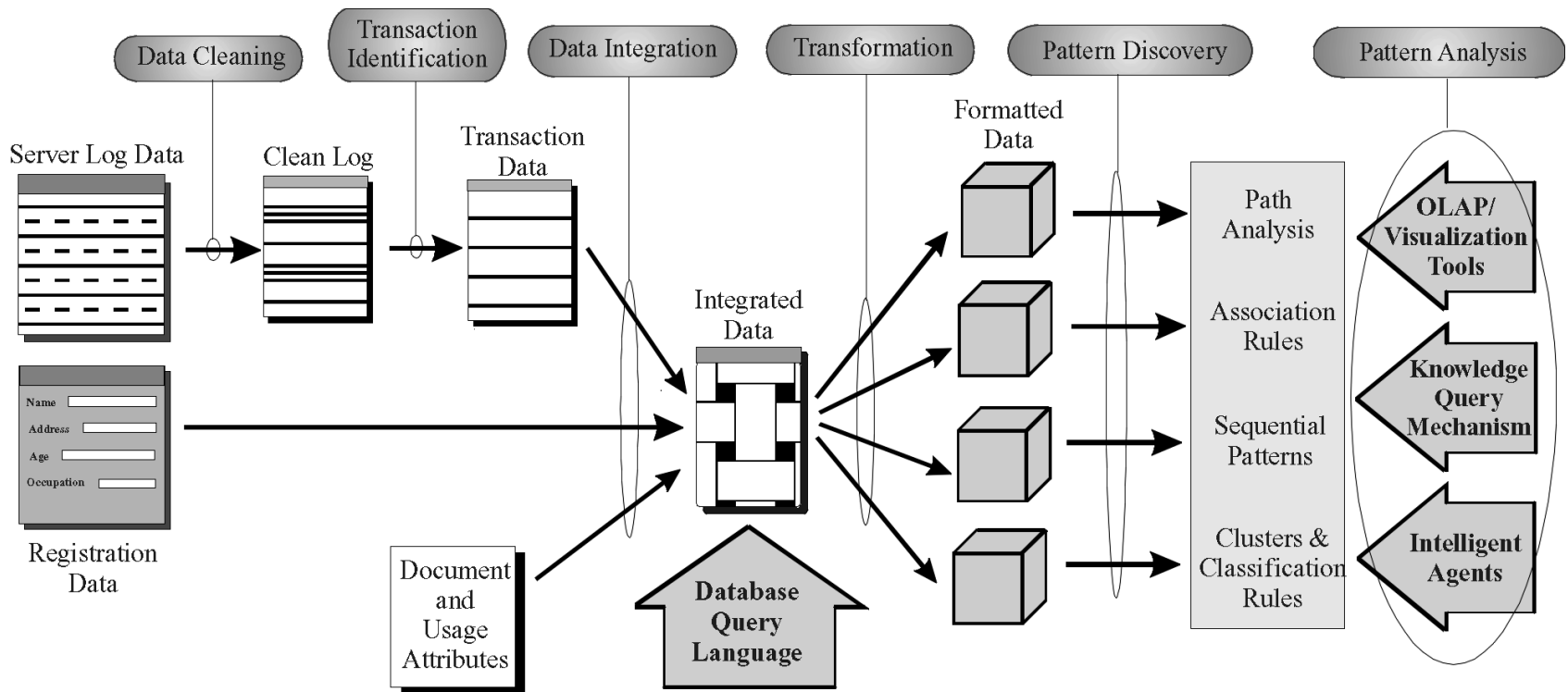
The discovery of interesting user access patterns from Web server logs

- **Generate simple statistical reports:**
 - A summary report of hits and bytes transferred
 - A list of top requested URLs
 - A list of top referrers
 - A list of most common browsers used
 - Hits per hour/day/week/month reports
 - Hits per domain reports
- **Learn:**
 - Who is visiting you site
 - The path visitors take through your pages
 - How much time visitors spend on each page
 - The most common starting page
 - Where visitors are leaving your site

Web Usage Mining – Three Phases



The Web Usage Mining Process



- General Architecture for the WEBMINER -

Web Server Access Logs

- Typical Data in a Server Access Log

```
looney.cs.umn.edu han - [09/Aug/1996:09:53:52 -0500] "GET mobasher/courses/cs5106/cs5106l1.html HTTP/1.0" 200
mega.cs.umn.edu njain - [09/Aug/1996:09:53:52 -0500] "GET / HTTP/1.0" 200 3291
mega.cs.umn.edu njain - [09/Aug/1996:09:53:53 -0500] "GET /images/backgnds/paper.gif HTTP/1.0" 200 3014
mega.cs.umn.edu njain - [09/Aug/1996:09:54:12 -0500] "GET /cgi-bin/Count.cgi?df=CS home.dat&dd=C\&ft=1 HTTP
mega.cs.umn.edu njain - [09/Aug/1996:09:54:18 -0500] "GET advisor HTTP/1.0" 302
mega.cs.umn.edu njain - [09/Aug/1996:09:54:19 -0500] "GET advisor/ HTTP/1.0" 200 487
looney.cs.umn.edu han - [09/Aug/1996:09:54:28 -0500] "GET mobasher/courses/cs5106/cs5106l2.html HTTP/1.0" 200
...           ...           ...
```

◆ Access Log Format

IP address userid time method url protocol status size

Example: Session Inference with Referrer Log

	IP	Time	URL	Referrer	Agent
1	www.aol.com	08:30:00	A	#	Mozilla/2.0; AIX 4.1.4
2	www.aol.com	08:30:01	B	E	Mozilla/2.0; AIX 4.1.4
3	www.aol.com	08:30:02	C	B	Mozilla/2.0; AIX 4.1.4
4	www.aol.com	08:30:01	B	#	Mozilla/2.0; Win 95
5	www.aol.com	08:30:03	C	B	Mozilla/2.0; Win 95
6	www.aol.com	08:30:04	F	#	Mozilla/2.0; Win 95
7	www.aol.com	08:30:04	B	A	Mozilla/2.0; AIX 4.1.4
8	www.aol.com	08:30:05	G	B	Mozilla/2.0; AIX 4.1.4

Identified Sessions:

$S_1: \# \implies A \implies B \implies G$

from references 1, 7, 8

$S_2: E \implies B \implies C$

from references 2, 3

$S_3: \# \implies B \implies C$

from references 4, 5

$S_4: \# \implies F$

from reference 6

Data Mining on Web Transactions

- Association Rules:

- ➔ discovers similarity among sets of items across transactions

$$X \xrightarrow{\alpha, \sigma} Y$$

where X, Y are sets of items, $\alpha = \textit{confidence}$ or $P(X \vee Y)$,
 $\sigma = \textit{support}$ or $P(X \wedge Y)$

- Examples:

- ➔ 60% of clients who accessed [/products/](#), also accessed [/products/software/webminer.htm](#).

- ➔ 30% of clients who accessed [/special-offer.html](#), placed an online order in [/products/software/](#).

- ➔ (Actual Example from IBM official Olympics Site)

- {Badminton, Diving} \implies {Table Tennis} ($\alpha = 69.7\%$, $\sigma = 0.35\%$)

Other Data Mining Techniques

- Sequential Patterns:
 - ➔ 30% of clients who visited [/products/software/](#), had done a search in **Yahoo** using the keyword “**software**” before their visit
 - ➔ 60% of clients who placed an online order for WEBMINER, placed another online order for software within 15 days
- Clustering and Classification
 - ➔ clients who often access [/products/software/webminer.html](#) tend to be from educational institutions.
 - ➔ clients who placed an online order for software tend to be students in the 20-25 age group and live in the United States.
 - ➔ 75% of clients who download software from [/products/software/demos/](#) visit between 7:00 and 11:00 pm on weekends.

Path and Usage Pattern Discovery

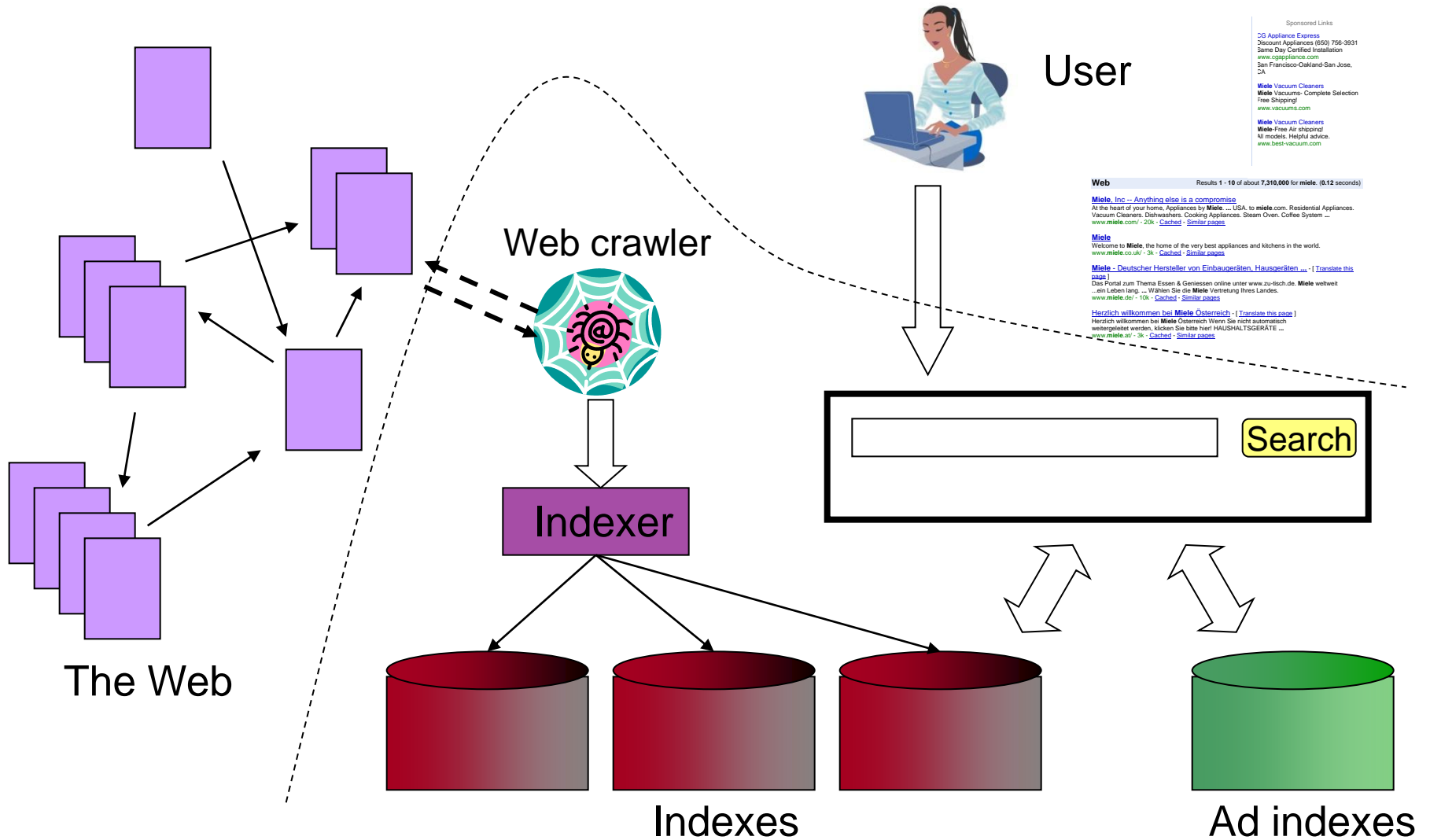
- Types of Path/Usage Information
 - Most Frequent paths traversed by users
 - Entry and Exit Points
 - Distribution of user session duration
- Examples:
 - 60% of clients who accessed `/home/products/file1.html`, followed the path `/home` ==> `/home/whatsnew` ==> `/home/products` ==> `/home/products/file1.html`
 - (Olympics Web site) 30% of clients who accessed [sport specific pages](#) started from the [Sneakpeek](#) page.
 - 65% of clients left the site after 4 or less references.

Search Engines for Web Mining

The number of Internet hosts exceeded...

- 1.000 in 1984
- 10.000 in 1987
- 100.000 in 1989
- 1.000.000 in 1992
- 10.000.000 in 1996
- 100.000.000 in 2000

Web search basics



Search engine components

- Spider (a.k.a. crawler/robot) – builds corpus
 - Collects web pages recursively
 - For each known URL, fetch the page, parse it, and extract new URLs
 - Repeat
 - Additional pages from direct submissions & other sources
- The indexer – creates inverted indexes
 - Various policies wrt which words are indexed, capitalization, support for Unicode, stemming, support for phrases, etc.
- Query processor – serves query results
 - Front end – query reformulation, word stemming, capitalization, optimization of Booleans, etc.
 - Back end – finds matching documents and ranks them

Web Search Products and Services

- Alta Vista
- DB2 text extender
- Excite
- Fulcrum
- Glimpse (Academic)
- Google!
- Inforceek Internet
- Inforceek Intranet
- Inktomi (HotBot)
- Lycos
- PLS
- Smart (Academic)
- Oracle text extender
- Verity
- Yahoo!

Boolean search in AltaVista

Boolean query: [Help](#) | [Customize Settings](#) | [Family Filter is off](#)

Use the terms AND, OR, AND NOT, NEAR

Sort by:
(Enter terms to prioritize your results.)

Date:

by timeframe:

by date range: to (dd/mm/yy)

Display:

one result per site

results per page

[Search Assistant](#) | [Basic Search](#)

Specifying field content in HotBot

<input type="text"/>		SEARCH
Look For	<input type="text" value="all the words"/>	? click to see explanations of how these options work.
Language	<input type="text" value="any language"/>	
Word Filter	<input type="text" value="must contain"/> <input type="text" value="the words"/>	
	<input type="text" value="must not contain"/> <input type="text" value="the phrase"/>	
	<input type="text" value=""/>	
	more terms <input style="font-size: small;" type="button" value="+"/>	
Date	<input checked="" type="radio"/> <input type="text" value="anytime"/>	
	<input type="radio"/> <input type="text" value="After"/> or on	
	<input type="text" value="January"/> <input type="text" value="1"/> , <input type="text" value="2000"/>	
Pages Must Include	<input type="checkbox"/> image <input type="checkbox"/> audio <input type="checkbox"/> MP3 <input type="checkbox"/> video	
	<input type="checkbox"/> Shockwave <input type="checkbox"/> Java <input type="checkbox"/> JavaScript <input type="checkbox"/> ActiveX	
	<input type="checkbox"/> RealAudio/Video <input type="checkbox"/> extension: <input type="text" value=""/> (.gif)	
Location/Domain	<input checked="" type="radio"/> Region <input type="radio"/> Domain	
	<input type="text" value="anywhere"/>	<input type="text" value=""/>
	<small>(.com, .edu) website: (wired.com, etc.) country code: (.uk, .fr, .jp)</small>	

Natural language interface in AskJeeves



The image shows a screenshot of the Ask Jeeves website interface. At the top right, there are navigation links: [About](#) | [Help](#). Below these are four buttons: [Ask Jeeves Home](#), [Browse by Subject](#), [Ask Other People](#), and [Shopping Guide](#). On the left, there is a cartoon character of a man in a suit, gesturing towards the search area. The main logo reads "Ask Jeeves" with "Ask.com" underneath. The central text asks "What can I help you find?". Below this is a search input field with a red "Ask" button to its right. A tip below the field reads: **Tip:** Use a question, phrase, or word - Jeeves is flexible. At the bottom, there are links for international sites: [Ask Jeeves Español \(Pregunta.com\)](#) | [Ask Jeeves UK \(United Kingdom\)](#) | [Ask Jeeves Australia](#). General information links include: [About](#) | [Business-to-Business Solutions](#) | [Advertise](#) | [Investor Relations](#) | [Become an Affiliate](#).

Three examples of search strategies

- Rank web pages based on popularity
- Rank web pages based on word frequency
- Match query to an expert database

All the major search engines use a mixed strategy in ranking web pages and responding to queries

Rank based on word frequency

- Library analogue: Keyword search
- Basic factors in HotBot ranking of pages:
 - words in the title
 - keyword meta tags
 - word frequency in the document
 - document length

Alternative word frequency measures

- Excite uses a thesaurus to search for what you want, rather than what you ask for
- AltaVista allows you to look for words that occur within a set distance of each other
- NorthernLight weighs results by search term sequence, from left to right

Rank based on popularity

- Library analogue: citation index
- The Google strategy for ranking pages:
 - Rank is based on the number of links to a page
 - Pages with a high rank have a lot of other web pages that link to it
 - The formula is on the Google help page 😊

More on popularity ranking

- The Google philosophy is also applied by others, such as NorthernLight
- HotBot measures the popularity of a page by how frequently users have clicked on it in past search results

Expert databases: Yahoo!

- An expert database contains predefined responses to common queries
- A simple approach is subject directory, e.g. in Yahoo!, which contains a selection of links for each topic
- The selection is small, but can be useful

Expert databases: AskJeeves







- AskJeeves has predefined responses to various types of common queries
- These prepared answers are augmented by a meta-search, which searches other SEs
- Library analogue: Reference desk

Best wines in France: AskJeeves

You asked: 

Tip: Try to keep your search questions short and to the point.

Click Ask for Answers! I have found the **answers** to the following questions:

-  Where can I see a list of the top 100 wines of 1999?
 -  Where can I buy from an online wine shop?
 -  Where can I find information on wine from ?
 -  Where can I find ?
 -  Where can I get a crash course in choosing ?
-
-  Post your question and get answers from other users in our Community Forum

People with **similar questions** have found these sites relevant:

 [All French wines, Bordeaux Burgundy and Champagne wine - Buy French Wine](#)

[Online](#)

NouvellePage4

From: <http://www.french-wine-shop.com/>

 [Slow Food Guide to the Wines of the World](#)

Web site by Intesys 2000 wineries 6500 wines described 174 top wines | Info | Catalogue | Top Wines | Search | Help | Copyright © 1996, Arcigola Slow Food and Veronafiere

From: <http://www.veronafiere.it/slowwines>

Best wines in France: HotBot

WEB RESULTS 53,600 Matches 1 - 10 [next](#) >>

1. [Jeroboam : Wine Cellar Management software.](#)

Software which manages your own wine cellar, helps you to serve and to pair your best vintages. Other Features include: Stock management, Pairing wines and dishes, help to serve and age wines. Available in English and French version.

2. [The best wines of South Africa](#)

The best wines of South Africa Unique dans la région Mouscronnoise et Tournaisienne Invitation découverte en seconde page ... Le soleil de l'Afrique du Sud Apporte au fruit de sa terre une saveur qui jalouse les vins que nous connaissons. Les viticu

3. [All French Wine Shop](#)

They are French and they sell wine. Customers can't ask for more when shopping for Bordeaux, Burgundy, Alsace and Cote du Rhone wines.

4. [French Wines](#)

Check out this guide to French wines. Includes a wine vocabulary section and an overview of French wineries.

5. [French Wines - Other](#)

Named as the best wine shop on the Internet by Money magazine and one of the top 10 wine retailers in the nation by the publishers of the Wine Spectator.

Best wines in France: Google

[Au Web du Vin, Cave à vin et vente en ligne](#)

... Accord Mets et Vins. Our cellar shop of finest **french wines** and alcohols, Boutique de vente en ligne et vins et alcools. ...
www.webduvin.com/ - 10k - [Afrit](#) - [Svipaðar síður](#).

[Best Cellars: French Wines](#)

... for its quality and diversity of **wines**, and only Italy can be compared with it in terms of quantity. **Best** Cellars currently stocks these fine **French wines**. ...
www.bestcellars.ag/france.htm - 68k - [Afrit](#) - [Svipaðar síður](#).

[All French wines, Bordeaux Burgundy and Champagne wine - Buy ...](#)

... on line sales of **French wines**. Bordeaux and Burgundy **wines**, Champagne red and white wine. The **best** vineyards of the **french** soils, grown up with skill from the ...
Lýsing: Online sales of **French** great **wines**: Bordeaux and Burgundy red and white, Champagne, Alsace, Loire...
Flokkur: [Regional](#) > [Europe](#) > [France](#) > [Business and Economy](#) > [Shopping](#) > [Wine](#) > [Retailers](#)
www.french-wine-shop.com/ - 6k - [Afrit](#) - [Svipaðar síður](#).

[Best links concerning shopping for french wines](#)

... \$15.50 only on Amazon store !! Why waste your time searching for shops on line ? Here is my Selection among **best** sites about **French wines** on the Internet. ...
www.french-wines.com/Link.htm - 7k - [Afrit](#) - [Svipaðar síður](#).

[Best Cellars - French Red Wines](#)

... Red **Wines** Taste Guide - A (light) to F (full ... bodied Fitou is made from the **best** traditional southern **French** grape varieties grown between Narbonne and ...
www.devon.directory.co.uk/bestcellars/pages/html/wines/france/red2.htm - 22k - [Afrit](#) - [Svipaðar síður](#).

Some possible improvements

- Automatic translation of websites
- More natural language intelligence
- Use meta data on trusty web pages

Predicting the future...

- Association analysis of related documents (a popular data mining technique)
- Graphical display of web communities (both two- and three dimensional)
- Client-adjusted query responses

Multi-Layered Meta-Web

What Role will XML Play?

- XML provides a promising direction for a more structured Web and DBMS-based Web servers
- Promote standardization, help construction of multi-layered Web-base.
- Will XML transform the Web into one unified database enabling structured queries like:
 - “find the cheapest airline ticket from NY to Chicago”
 - “list all jobs with salary > 50 K in the Boston area”
- It is a dream now but more will be minable in the future!

Web Mining in an XML View

- Suppose most of the documents on web will be published in XML format and come with a valid DTD.
- XML documents can be stored in a relational database, OO database, or a specially-designed database
- To increase efficiency, XML documents can be stored in an intermediate format.

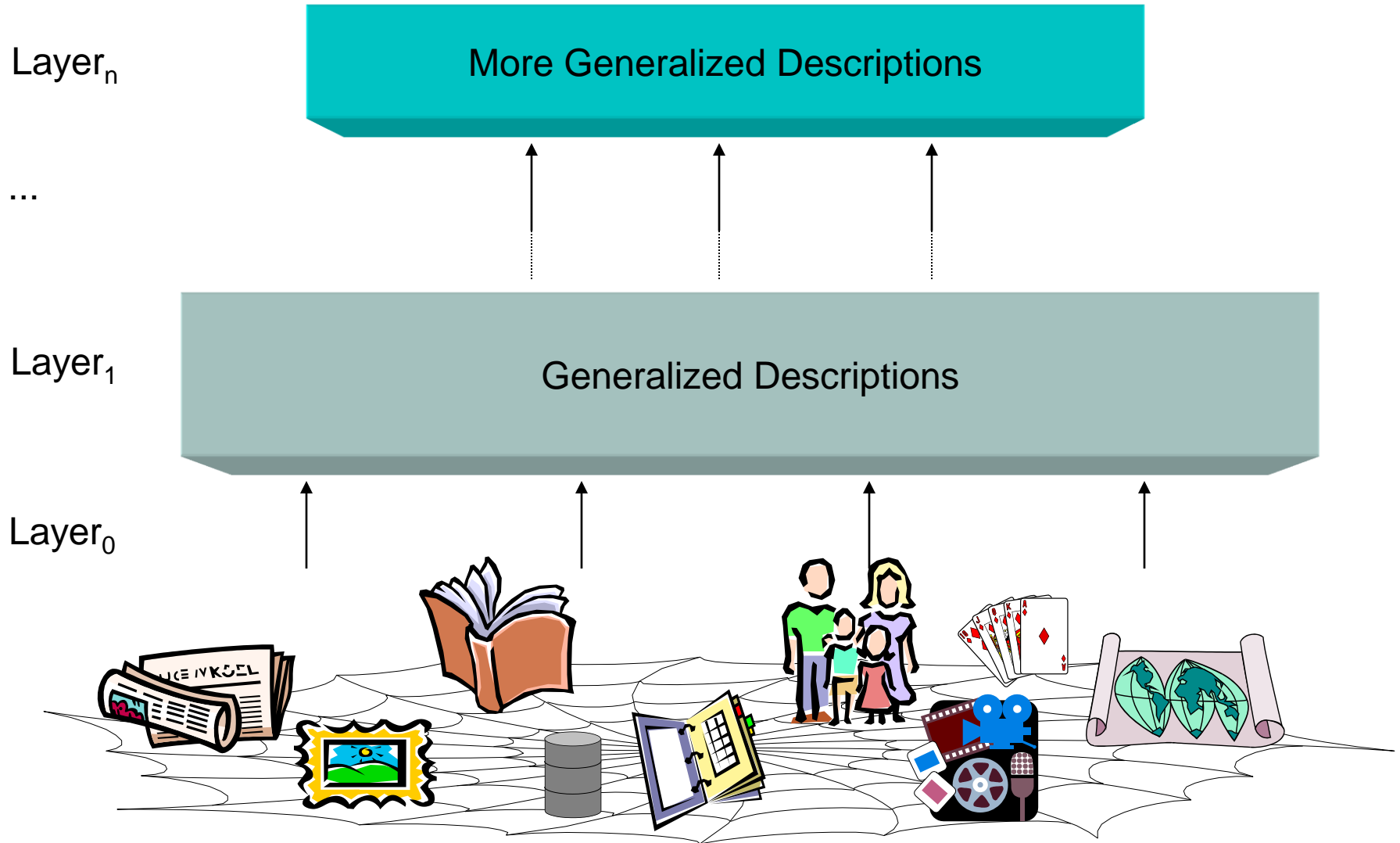
Mine What Web Search Engine Finds

- Current Web search engines: convenient source for mining
 - keyword-based, return too many answers, low quality answers, still missing a lot, not customized, etc.
- Data mining will help:
 - coverage: using synonyms and conceptual hierarchies
 - better search primitives: user preferences/hints
 - linkage analysis: authoritative pages and clusters
 - Web-based languages: XML + WebSQL + WebML
 - customization: home page + Weblog + user profiles

Warehousing a Meta-Web: An MLDB Approach

- *Meta-Web*: A structure which summarizes the contents, structure, linkage, and access of the Web and which evolves with the Web
- Layer₀: the Web itself
- Layer₁: the lowest layer of the Meta-Web
 - an entry: a Web page summary, including class, time, URL, contents, keywords, popularity, weight, links, etc.
- Layer₂ and up: summary/classification/clustering in various ways and distributed for various applications
- Meta-Web can be warehoused and incrementally updated
- Querying and mining can be performed on or assisted by meta-Web (a multi-layer digital library catalogue, yellow page).

A Multiple Layered Meta-Web Architecture



Construction of Multi-Layer Meta-Web

- XML: facilitates structured and meta-information extraction
- Hidden Web: DB schema “extraction” + other meta info
- Automatic classification of Web documents:
 - based on Yahoo!, etc. as training set + keyword-based correlation/classification analysis (AI assistance)
- Automatic ranking of important Web pages
 - authoritative site recognition and clustering Web pages
- Generalization-based multi-layer meta-Web construction
 - With the assistance of clustering and classification analysis

Use of Multi-Layer Meta Web

- **Benefits of Multi-Layer Meta-Web:**
 - Multi-dimensional Web info summary analysis
 - Approximate and intelligent query answering
 - Web high-level query answering (WebSQL, WebML)
 - Web content and structure mining
 - Observing the dynamics/evolution of the Web
- **Is it realistic to construct such a meta-Web?**
 - Benefits even if it is partially constructed
 - Benefits may justify the cost of tool development, standardization and partial restructuring

Conclusion

- Web Mining fills the information gap between web users and web designers