Mining web data: Techniques for understanding the user behavior in the Web

Juan D. Velásquez Silva PhD U. of Tokyo
Assistant Professor, U. of Chile
jvelasqu@dii.uchile.cl
http://wi.dii.uchile.cl
Web Intelligence Research Group
Department of Industrial Engineering
University of Chile

Outline

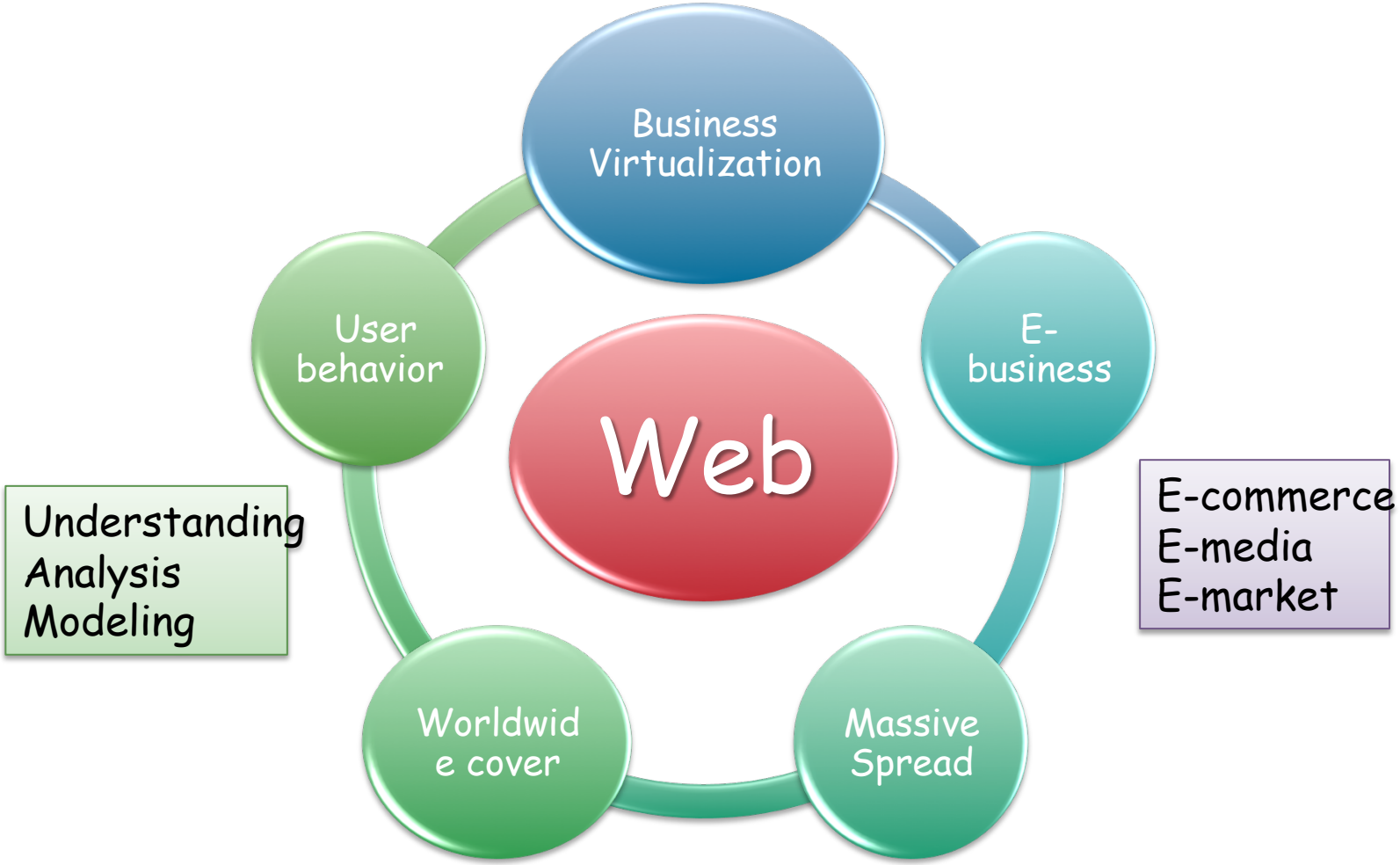
1. Motivation.
2. Web data.
3. Web mining.
4. Applications.
5. Summary.

1.- Motivation

The new sales window

- The markets move towards business virtualization.
- Many companies have begun to use the Web as a new channel of massive spread and worldwide cover.
- How can we improve our web site?
- By understanding what describes the user behavior [Brusilovsky96].

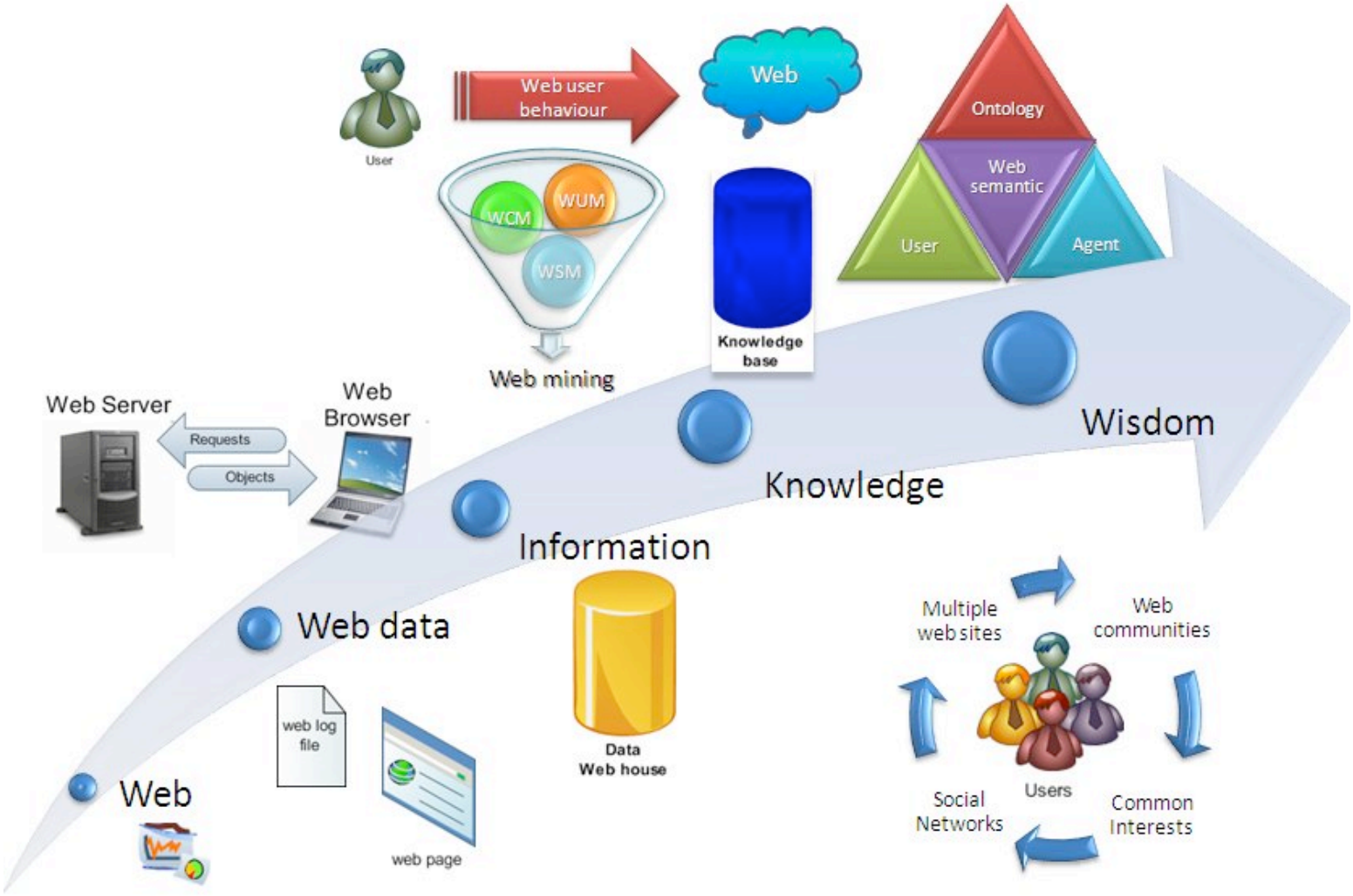
The new sales windows



OK, but how can we do it?

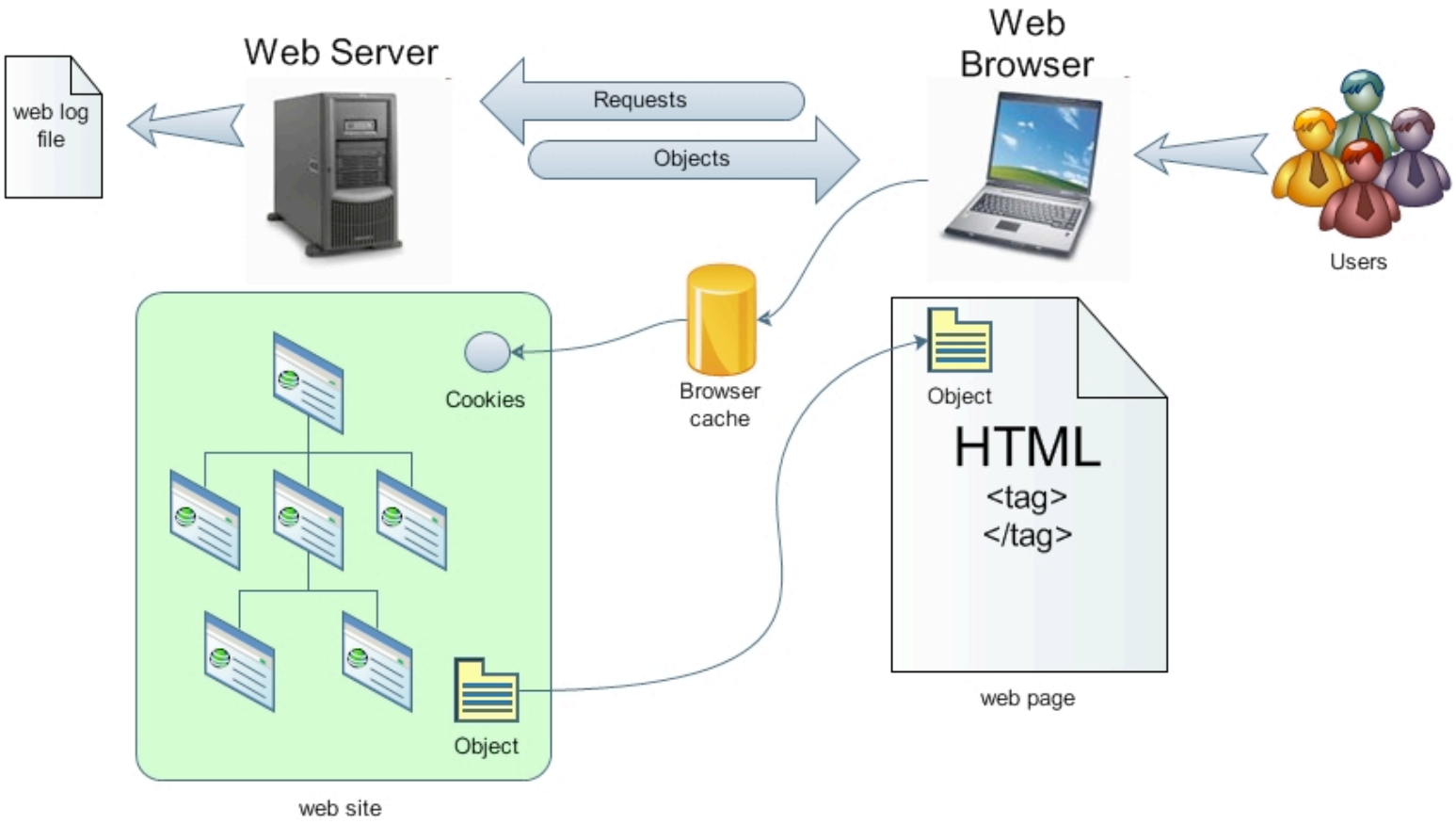
- Analyzing
  - The user's browsing behavior.
  - The web page preferences.
  - The user profile.
- In fact "Extracting knowledge from web data"
- ... applying Web mining techniques.

From data to knowledge and wisdom



2.- Data originated in the Web: Web Data

Web basic operation



Web server and web browser (2)

- Once the document has been read by the browser, the specific tags inside are interpreted.
- When the browser interpreting the tags find a reference about an object, for instance an image, the HTTP gets it and transfers it to the browser.
- The process finishes when the last tag is interpreted and the page is shown to the visitor.

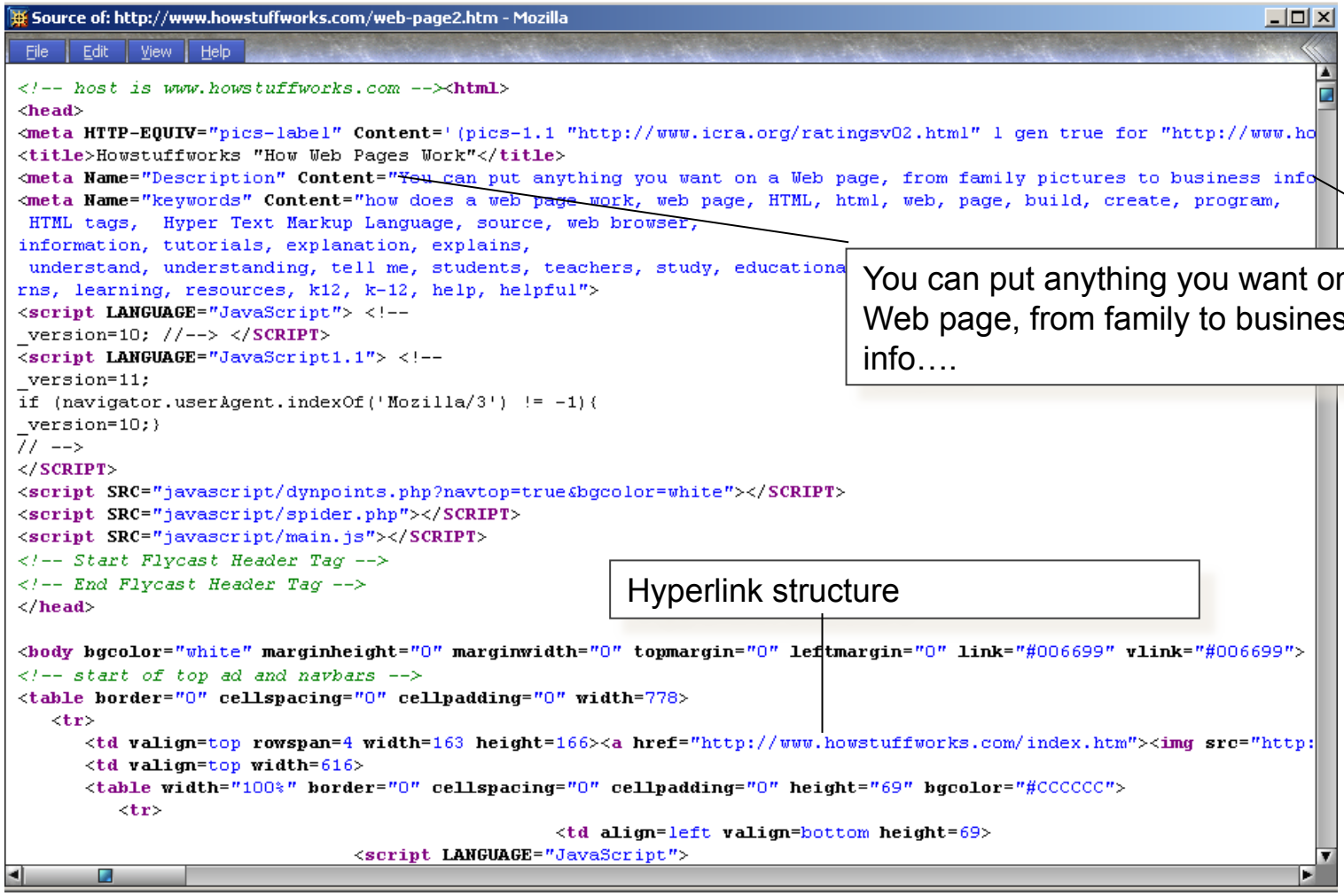
Web server and web browser (3)

- The web log registers contain information about the visitor browsing behavior, in particular the page navigation sequence and the time spent in each page visited.
- When a web page is accessed, the HTML code, with web page tags referring to various web objects, is interpreted in the browser.
- A register is created for the accessed page as well as for each object referred in the page.
- Depending on the web activity, these logs can contain millions of registers and most of them may not hold relevant information.

Web logs

#	IP	Id	Acces	Time	Method/URL/Protocol	Status	Bytes	Referer	Agent
1	165.182.168.101	-	-	16/06/2002:16:24:06	GET p1.htm HTTP/1.1	200	3821	out.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
2	165.182.168.101	-	-	16/06/2002:16:24:10	GET A.gif HTTP/1.1	200	3766	p1.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
3	165.182.168.101	-	-	16/06/2002:16:24:57	GET B.gif HTTP/1.1	200	2878	p1.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
4	204.231.180.195	-	-	16/06/2002:16:32:06	GET p3.htm HTTP/1.1	304	0	-	Mozilla/4.0 (MSIE 6.0; Win98)
5	204.231.180.195	-	-	16/06/2002:16:32:20	GET C.gif HTTP/1.1	304	0	-	Mozilla/4.0 (MSIE 6.0; Win98)
6	204.231.180.195	-	-	16/06/2002:16:34:10	GET p1.htm HTTP/1.1	200	3821	p3.htm	Mozilla/4.0 (MSIE 6.0; Win98)
7	204.231.180.195	-	-	16/06/2002:16:34:31	GET A.gif HTTP/1.1	200	3766	p1.htm	Mozilla/4.0 (MSIE 6.0; Win98)
8	204.231.180.195	-	-	16/06/2002:16:34:53	GET B.gif HTTP/1.1	200	2878	p1.htm	Mozilla/4.0 (MSIE 6.0; Win98)
9	204.231.180.195	-	-	16/06/2002:16:38:40	GET p2.htm HTTP/1.1	200	2960	p1.htm	Mozilla/4.0 (MSIE 6.0; Win98)
10	165.182.168.101	-	-	16/06/2002:16:39:02	GET p1.htm HTTP/1.1	200	3821	out.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
11	165.182.168.101	-	-	16/06/2002:16:39:15	GET A.gif HTTP/1.1	200	3766	p1.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
12	165.182.168.101	-	-	16/06/2002:16:39:45	GET B.gif HTTP/1.1	200	2878	p1.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
13	165.182.168.101	-	-	16/06/2002:16:39:58	GET p2.htm HTTP/1.1	200	2960	p1.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
14	165.182.168.101	-	-	16/06/2002:16:42:03	GET p3.htm HTTP/1.1	200	4036	p2.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
15	165.182.168.101	-	-	16/06/2002:16:42:07	GET p2.htm HTTP/1.1	200	2960	p1.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
16	165.182.168.101	-	-	16/06/2002:16:42:08	GET C.gif HTTP/1.1	200	3423	p2.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
17	204.231.180.195	-	-	16/06/2002:17:34:20	GET p3.htm HTTP/1.1	200	2342	out.htm	Mozilla/4.0 (MSIE 6.0; Win98)
18	204.231.180.195	-	-	16/06/2002:17:34:48	GET C.gif HTTP/1.1	200	3423	p2.htm	Mozilla/4.0 (MSIE 6.0; Win98)
19	204.231.180.195	-	-	16/06/2002:17:35:45	GET p4.htm HTTP/1.1	200	3523	p3.htm	Mozilla/4.0 (MSIE 6.0; Win98)
20	204.231.180.195	-	-	16/06/2002:17:35:56	GET D.gif HTTP/1.1	200	3231	p4.htm	Mozilla/4.0 (MSIE 6.0; Win98)
21	204.231.180.195	-	-	16/06/2002:17:36:06	GET E.gif HTTP/1.1	404	0	p4.htm	Mozilla/4.0 (MSIE 6.0; Win98)

Web site contents



Nature of web data

- **Content**. The web page content, i.e., pictures, free text, sounds, etc.
- **Structure**. Data that show the internal web page structure. In general, they are HTML or XML tags, some of them contain information about hyperlink connections with other web pages.
- **Usage**. Data that describe the visitor preferences while browsing in a web site. It is possible to find them inside web log files.
- **User profile**. A collection of information about a user, combining personal information (e.g. name, age etc.), usage information (e.g. page visited) and interests.

Understanding the visitor behavior in a web site

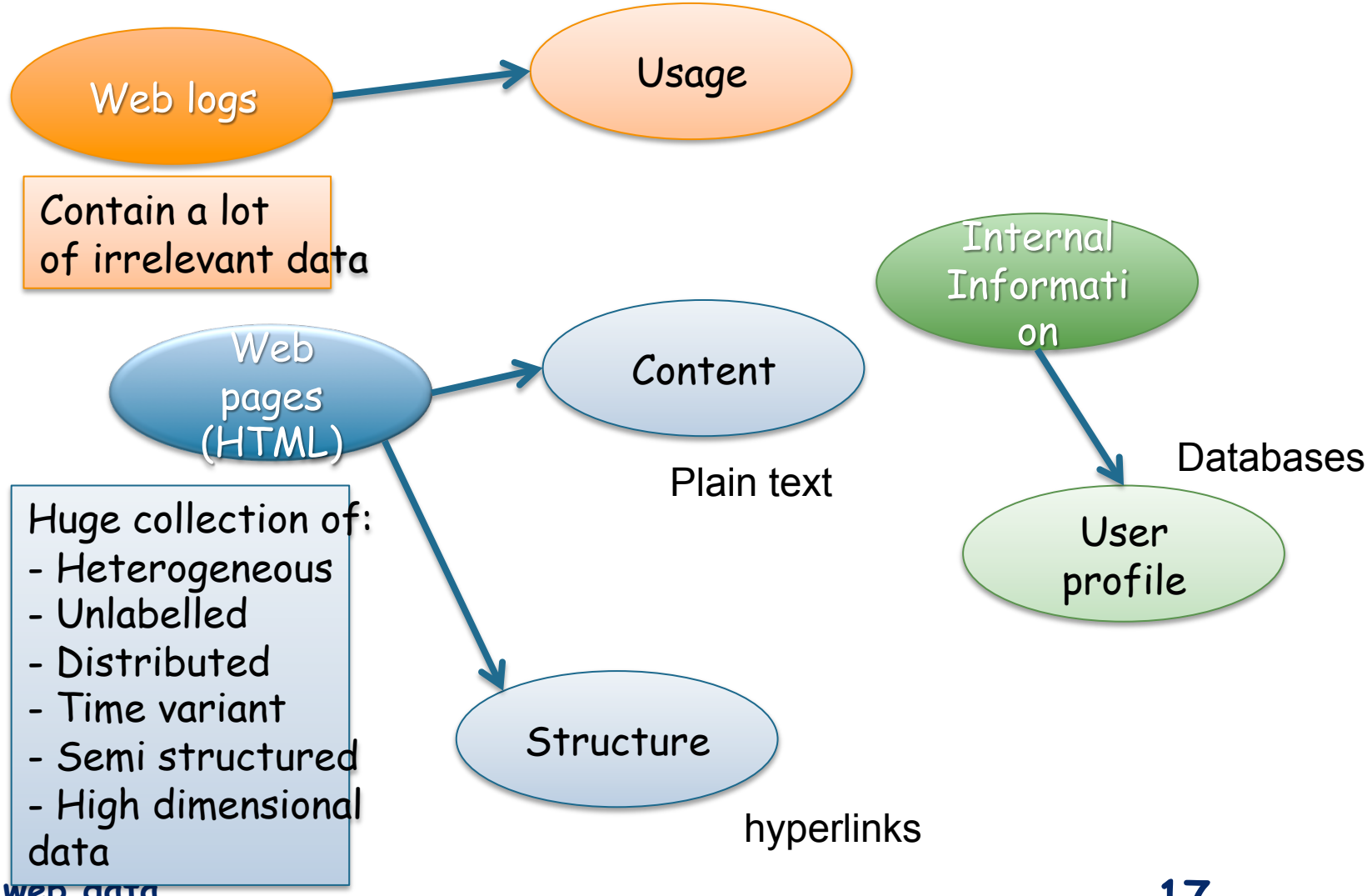
- Visitor browsing behavior: Web logs.
- Visitor preferences: Web pages
- Problems:
  - Web logs contain a lot of irrelevant data.
  - A Web site is a huge collection of heterogeneous, unlabelled, distributed, time variant, semi-structured and high dimensional data.

The dream

"Transform the visitors into customers and retain the existing ones"

- Some solutions:
  - Continuous improvement of the web site structure and content.
  - Personalization of the relationship between the user and the web site.
  - Understanding the user behavior in the web site.

Collecting data from the Web



Data cleaning and preprocessing

- Web logs. Applying the sessionization process.
  - To identify the real visitor session.
- Web pages. Vector space model.
  - Transforming the web pages into feature vectors.
- Web hyperlinks structure.
  - Identifying social networks

Web logs

#	IP	Id	Acces	Time	Method/URL/Protocol	Status	Bytes	Referer	Agent
1	165.182.168.101	-	-	16/06/2002:16:24:06	GET p1.htm HTTP/1.1	200	3821	out.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
2	165.182.168.101	-	-	16/06/2002:16:24:10	GET A.gif HTTP/1.1	200	3766	p1.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
3	165.182.168.101	-	-	16/06/2002:16:24:57	GET B.gif HTTP/1.1	200	2878	p1.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
4	204.231.180.195	-	-	16/06/2002:16:32:06	GET p3.htm HTTP/1.1	304	0	-	Mozilla/4.0 (MSIE 6.0; Win98)
5	204.231.180.195	-	-	16/06/2002:16:32:20	GET C.gif HTTP/1.1	304	0	-	Mozilla/4.0 (MSIE 6.0; Win98)
6	204.231.180.195	-	-	16/06/2002:16:34:10	GET p1.htm HTTP/1.1	200	3821	p3.htm	Mozilla/4.0 (MSIE 6.0; Win98)
7	204.231.180.195	-	-	16/06/2002:16:34:31	GET A.gif HTTP/1.1	200	3766	p1.htm	Mozilla/4.0 (MSIE 6.0; Win98)
8	204.231.180.195	-	-	16/06/2002:16:34:53	GET B.gif HTTP/1.1	200	2878	p1.htm	Mozilla/4.0 (MSIE 6.0; Win98)
9	204.231.180.195	-	-	16/06/2002:16:38:40	GET p2.htm HTTP/1.1	200	2960	p1.htm	Mozilla/4.0 (MSIE 6.0; Win98)
10	165.182.168.101	-	-	16/06/2002:16:39:02	GET p1.htm HTTP/1.1	200	3821	out.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
11	165.182.168.101	-	-	16/06/2002:16:39:15	GET A.gif HTTP/1.1	200	3766	p1.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
12	165.182.168.101	-	-	16/06/2002:16:39:45	GET B.gif HTTP/1.1	200	2878	p1.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
13	165.182.168.101	-	-	16/06/2002:16:39:58	GET p2.htm HTTP/1.1	200	2960	p1.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
14	165.182.168.101	-	-	16/06/2002:16:42:03	GET p3.htm HTTP/1.1	200	4036	p2.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
15	165.182.168.101	-	-	16/06/2002:16:42:07	GET p2.htm HTTP/1.1	200	2960	p1.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
16	165.182.168.101	-	-	16/06/2002:16:42:08	GET C.gif HTTP/1.1	200	3423	p2.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
17	204.231.180.195	-	-	16/06/2002:17:34:20	GET p3.htm HTTP/1.1	200	2342	out.htm	Mozilla/4.0 (MSIE 6.0; Win98)
18	204.231.180.195	-	-	16/06/2002:17:34:48	GET C.gif HTTP/1.1	200	3423	p2.htm	Mozilla/4.0 (MSIE 6.0; Win98)
19	204.231.180.195	-	-	16/06/2002:17:35:45	GET p4.htm HTTP/1.1	200	3523	p3.htm	Mozilla/4.0 (MSIE 6.0; Win98)
20	204.231.180.195	-	-	16/06/2002:17:35:56	GET D.gif HTTP/1.1	200	3231	p4.htm	Mozilla/4.0 (MSIE 6.0; Win98)
21	204.231.180.195	-	-	16/06/2002:17:36:06	GET E.gif HTTP/1.1	404	0	p4.htm	Mozilla/4.0 (MSIE 6.0; Win98)

Session reconstruction: the need

- If we want to understand the user behavior in a web site, web need to know his/her real browsing behavior.
- The quality of patterns extracted by using a mining technique depend on the input data.
- Elements like proxies servers, dynamic IP, missing references and the inability of servers to identify different users make difficult to reconstruct a real session.

Session reconstruction: I/O

- Input: The complete set of log registers.
- Output: A real user session identified.
- Noise source.
  - Crawler logs.
  - Logs don't related directly with the page (gif, sounds, etc)
  - Bad sessions.
  - Short sessions.
  - Large sessions.

Sessionization process

IP	Agent
165.182.168.101	MSIE 5.01
165.182.168.101	MSIE 5.01
165.182.168.101	MSIE 5.01
165.182.168.101	MSIE 5.5
165.182.168.101	MSIE 5.5
165.182.168.101	MSIE 5.5
165.182.168.101	MSIE 5.5
204.231.180.195	MSIE 6.0
204.231.180.195	MSIE 6.0
204.231.180.195	MSIE 6.0
204.231.180.195	MSIE 6.0
204.231.180.195	MSIE 6.0

	Date	
.....	16–Jun–02	16:39:02
.....	16–Jun–02	16:39:58
	16–Jun–02	16:42:03
	16–Jun–02	16:24:06
	16–Jun–02	16:26:05
.....	16–Jun–02	16:42:07
	16–Jun–02	16:58:03
.....	16–Jun–02	16:32:06
	16–Jun–02	16:34:10
	16–Jun–02	16:38:40
	16–Jun–02	17:34:20
.....	16–Jun–02	17:35:45

IP	Agent	Date		Sess
165.182.168.101	MSIE 5.01	16–Jun–02	16:39:02	1
165.182.168.101	MSIE 5.01	16–Jun–02	16:39:58	1
165.182.168.101	MSIE 5.01	16–Jun–02	16:42:03	1
165.182.168.101	MSIE 5.5	16–Jun–02	16:24:06	2
165.182.168.101	MSIE 5.5	16–Jun–02	16:26:05	2
165.182.168.101	MSIE 5.5	16–Jun–02	16:42:07	2
204.231.180.195	MSIE 6.0	16–Jun–02	16:32:06	3
204.231.180.195	MSIE 6.0	16–Jun–02	16:34:10	3
204.231.180.195	MSIE 6.0	16–Jun–02	16:38:40	3
204.231.180.195	MSIE 6.0	16–Jun–02	17:34:20	4
204.231.180.195	MSIE 6.0	16–Jun–02	17:35:45	4

L: Set of logs registers $R = \{r_1,...,r_n\}$ Real sessions

$$\forall r_i \in R, \forall j = 2,...,length(r_i) \quad r_{i,j}.timestamp > r_{i,j-1}.timestamp$$

$$\bigcup_{r_i \in R} \left(\bigcup_{j=1}^{length(r_i)} r_{i,j} \right) = L$$

$$\forall r_i \in R, \forall j = 1,...,length(r_i) : \nexists i' \neq i, j' / r_{i,j} = r_{i',j'}$$

Some problems [Berendt01, Cooley99]

- Single IP address/Multiple Server Sessions.
- Multiple IP addresses/Single Server Sessions. For privacy reasons or ISP configuration, it is possible to assign a random IP address to a visitor request.
- Multiple IP address/Single Visitor. A visitor that accesses a web site from different machines, but has the same behavior each time.
- Multiple Agent/Single User. As before, when a visitor uses different machines that may have different agents.

User and Session Identification Issues

- Distinguish among different users to a site
- Reconstruct the activities of the users within the site
- Proxy servers and anonymizers
- Rotating IP addresses connections through ISPs
- Missing references due to caching
- Inability of servers to distinguish among different visits

Some solutions

- Remote Agent
  - A remote agent is implemented in Java Applet
  - It is loaded into the client only once when the first page is accessed
  - The subsequent requests are captured and send back to the server
- Modified Browse
  - The source code of the existing browser can be modified to gain user specific data at the client side
- Dynamic page rewriting
  - When the user first submit the request, the server returns the requested page rewritten to include a session specific ID
  - Each subsequent request will supply this ID to the server
- Heuristics
  - use a set of assumptions to identify user sessions and find the missing cache hits in the server log

Mechanisms for session identification [Berendt 2002]

Method	Description	Privace Concerns	Advantages	Disadvantages
IP Adress + Agent	Assume each unique IP address/Agent pair is a unique user	Low	Always available. No Additional technology required.	Not guaranteed to be unique. Defeated by rotating Ips.
Embedded Sessions Ids	Use dinamically generated pages to associate ID with every hyperlink	Low to Medium	Always available. Independent of IP address	Cannot capture repeat visitors. Additional overhead for dynamic pages
Registration	User explicity logs into the site	Medium	Can track individuals not just browsers	Many users won't register. Not available before registration
Cookie	Save ID on the client machine	Medium to High	Can track repeat visit from same browser	Can be turned off by users
Software Agents	Program loaded into browser and sends back usage data	High	Accurrate usage data for a single site	Likely to be rejected by users

WUM – Heuristics

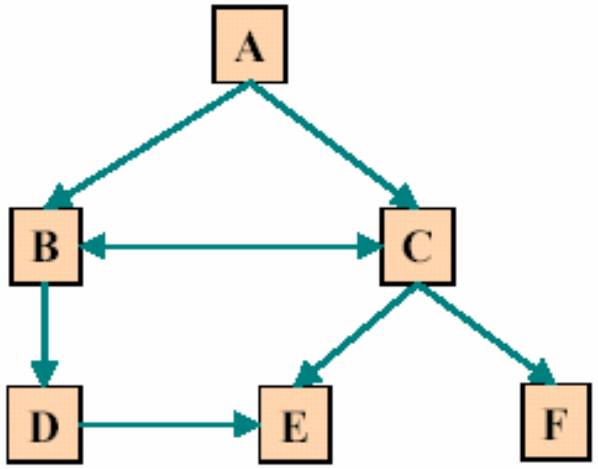
- The session identification heuristics
  - Timeout: if the time between pages requests exceeds a certain limit, it is assumed that the user is starting a new session
  - IP/Agent: Each different agent type for an IP address represents a different sessions
  - Referring page: If the referring page file for a request is not part of an open session, it is assumed that the request is coming from a different session.
  - Same IP-Agent/different sessions (Closest): Assigns the request to the session that is closest to the referring page at the time of the request.
  - Same IP-Agent/different sessions (Recent): In the case where multiple sessions are same distance from a page request, assigns the request to the session with the most recent referrer access in terms of time

WUM – Heuristics (2)

- The path completion heuristics
  - If the referring page file of a session is not part of the previous page file of that session, the user must have accessed a cached page
  - The "back" button method is used to refer a cached page.
  - Assigns a constant view time for each of the cached page file

Based on examples in tutorial PKDD-2002 by Berendt et. al.

Sessionization- Example



Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:10	2.3.4.5	C	-	IE4;Win98
0:12	2.3.4.5	B	C	IE4;Win98
0:15	2.3.4.5	E	C	IE4;Win98
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE4;Win98
0:22	1.2.3.4	A	-	IE4;Win98
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE4;Win98
0:33	1.2.3.4	B	C	IE4;Win98
0:58	1.2.3.4	D	B	IE4;Win98
1:10	1.2.3.4	E	D	IE4;Win98
1:15	1.2.3.4	A	-	IE5;Win2k
1:16	1.2.3.4	C	A	IE5;Win2k
1:17	1.2.3.4	F	C	IE4;Win98
1:25	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

Based on examples in tutorial PKDD-2002 by Berendt et. al.

Sessionization- Example (2)

Sort the users (IP+Agent)

Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:10	2.3.4.5	C	-	IE4;Win98
0:12	2.3.4.5	B	C	IE4;Win98
0:15	2.3.4.5	E	C	IE4;Win98
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE4;Win98
0:22	1.2.3.4	A	-	IE4;Win98
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE4;Win98
0:33	1.2.3.4	B	C	IE4;Win98
0:58	1.2.3.4	D	B	IE4;Win98
1:10	1.2.3.4	E	D	IE4;Win98
1:15	1.2.3.4	A	-	IE5;Win2k
1:16	1.2.3.4	C	A	IE5;Win2k
1:17	1.2.3.4	F	C	IE4;Win98
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:19	1.2.3.4	C	A	IE5;Win2k
0:25	1.2.3.4	E	C	IE5;Win2k
1:15	1.2.3.4	A	-	IE5;Win2k
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

0:10	2.3.4.5	C	-	IE4;Win98
0:12	2.3.4.5	B	C	IE4;Win98
0:15	2.3.4.5	E	C	IE4;Win98
0:22	2.3.4.5	D	B	IE4;Win98

0:22	1.2.3.4	A	-	IE4;Win98
0:25	1.2.3.4	C	A	IE4;Win98
0:33	1.2.3.4	B	C	IE4;Win98
0:58	1.2.3.4	D	B	IE4;Win98
1:10	1.2.3.4	E	D	IE4;Win98
1:17	1.2.3.4	F	C	IE4;Win98

Based on examples in tutorial PKDD-2002 by Berendt et. al.

Sessionization- Example (3)

Sessionize using heuristics (h1 with 30 min)

0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:19	1.2.3.4	C	A	IE5;Win2k
0:25	1.2.3.4	E	C	IE5;Win2k
1:15	1.2.3.4	A	-	IE5;Win2k
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:19	1.2.3.4	C	A	IE5;Win2k
0:25	1.2.3.4	E	C	IE5;Win2k

1:15	1.2.3.4	A	-	IE5;Win2k
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

The h1 heuristic (timeout=30 min) will result in the two sessions

Based on examples in tutorial PKDD-2002 by Berendt et. al.

Sessionization- Example (4)

Sessionize using heuristics (with href)

0:22	1.2.3.4	A	-	IE4;Win98
0:25	1.2.3.4	C	A	IE4;Win98
0:33	1.2.3.4	B	C	IE4;Win98
0:58	1.2.3.4	D	B	IE4;Win98
1:10	1.2.3.4	E	D	IE4;Win98
1:17	1.2.3.4	F	C	IE4;Win98

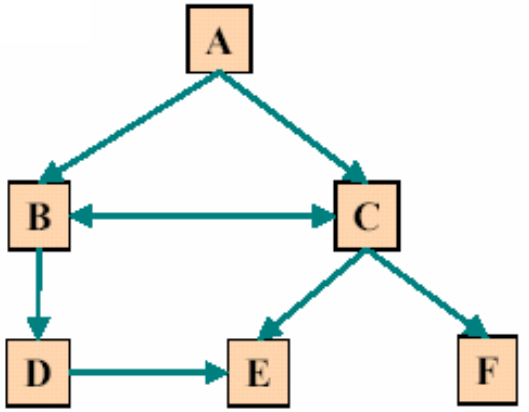
By using the reffer-based heuristics, we have only a single session

Based on examples in tutorial PKDD-2002 by Berendt et. al.

Sessionization- Example (5)

Path completion

0:22	1.2.3.4	A	-	IE4;Win98
0:25	1.2.3.4	C	A	IE4;Win98
0:33	1.2.3.4	B	C	IE4;Win98
0:58	1.2.3.4	D	B	IE4;Win98
1:10	1.2.3.4	E	D	IE4;Win98
1:17	1.2.3.4	F	C	IE4;Win98



A=>C , C=>B , B=>D , D=>E , C=>F

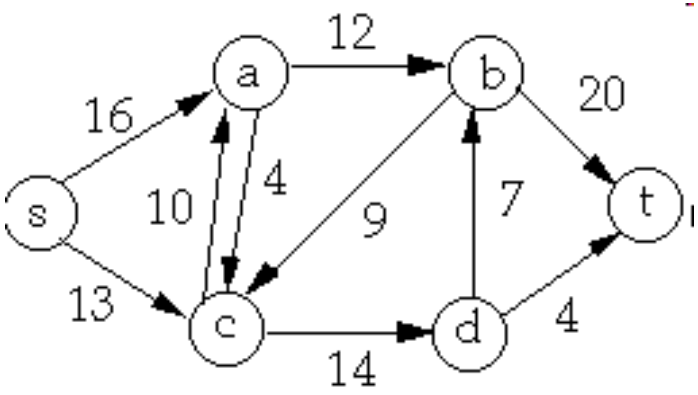
Need to look for the shortest backwards path from E to C based on the site topology. Note, however, that the elements of the path need to have occurred in the user trail previously.

E=>D, D=>B, B=>C

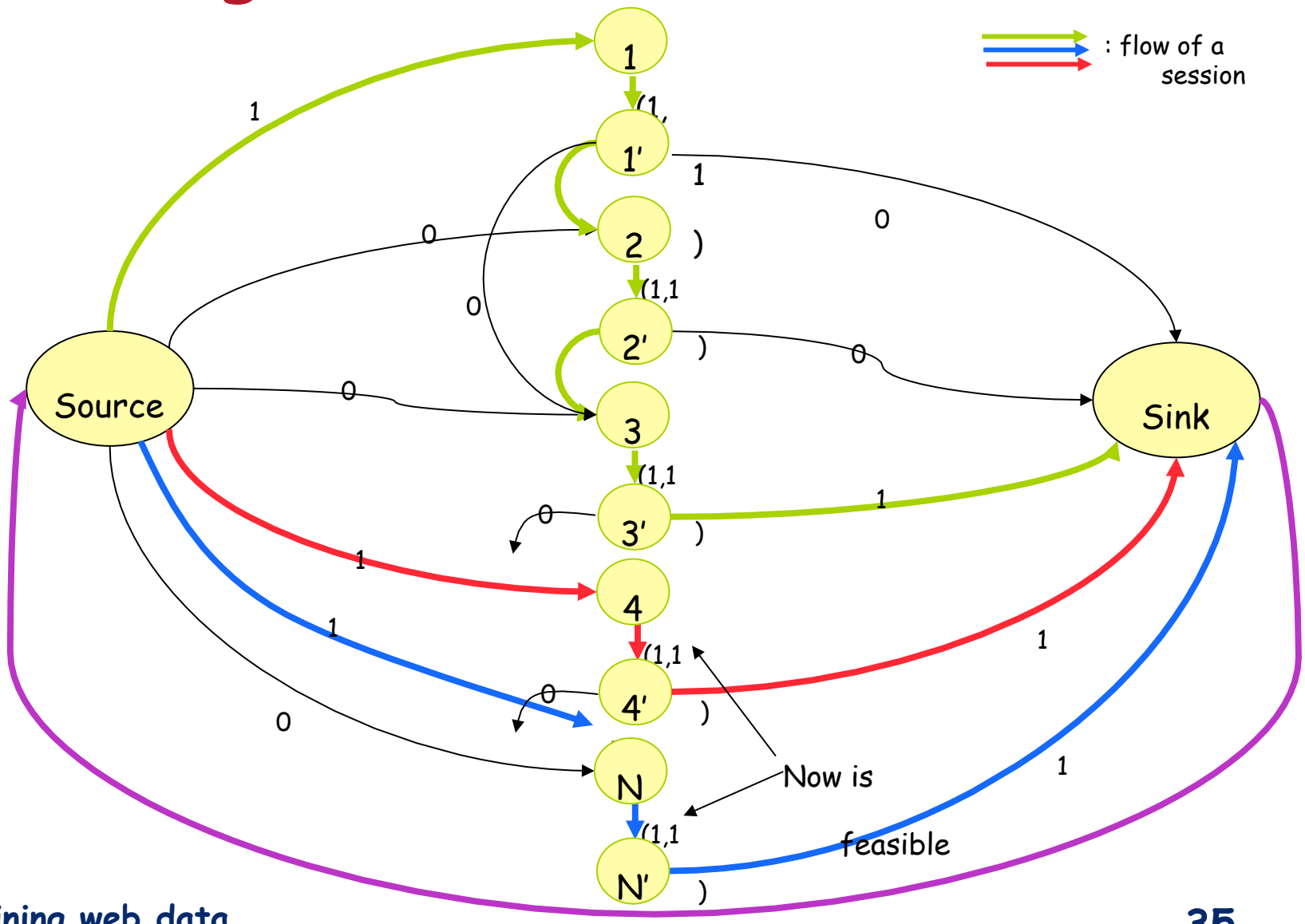
A novel solution: Network Flow Model

[Román P., et All., Submitted]

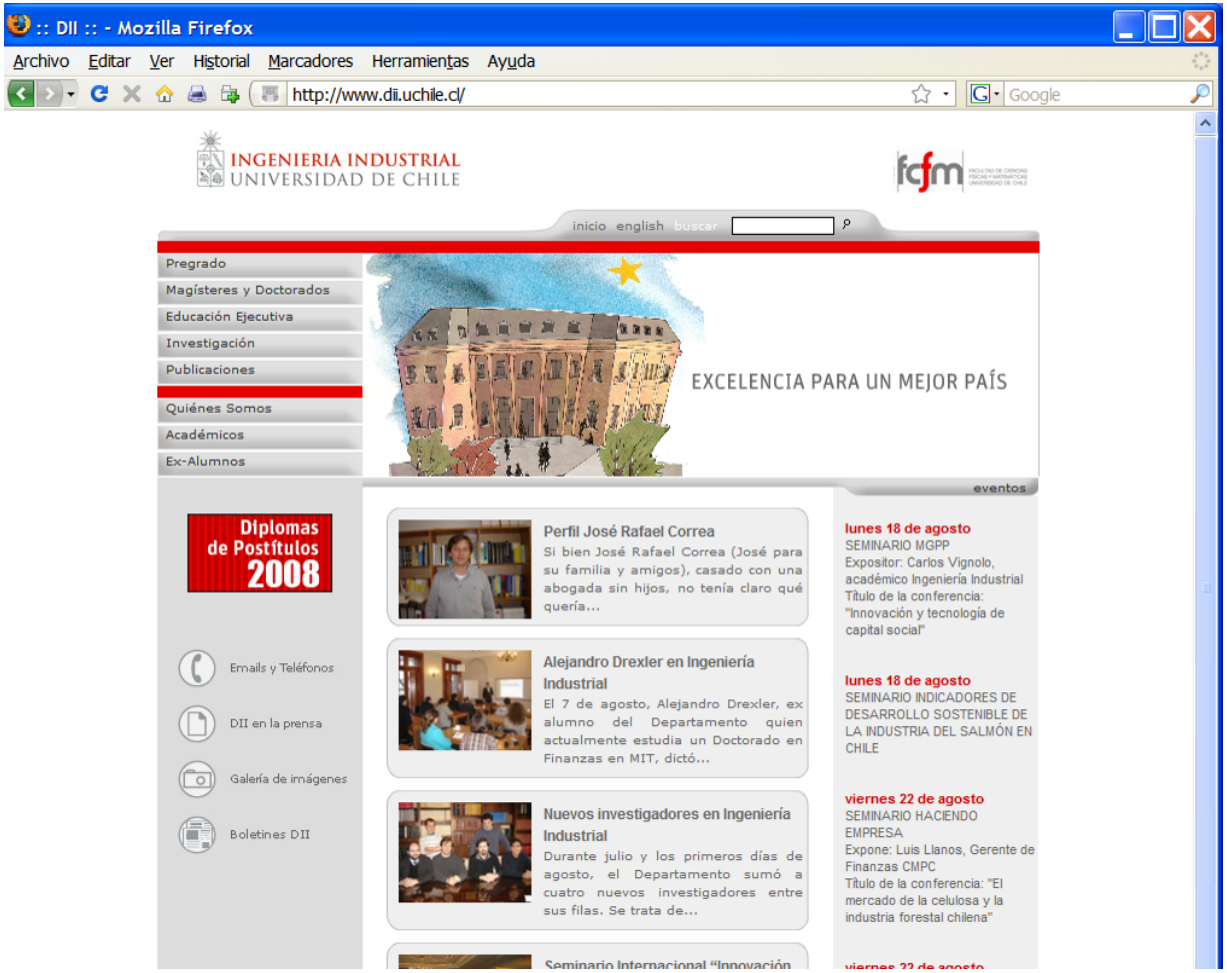
- Solutions prove to be solved in polynomial time (i.e. Ford-Fulkerson)
- Same linear restriction converted to build an network (links, time ordering).
- Minimize the flow: Number of session
- Each session is a unit of flow.



Minimizing the number of session

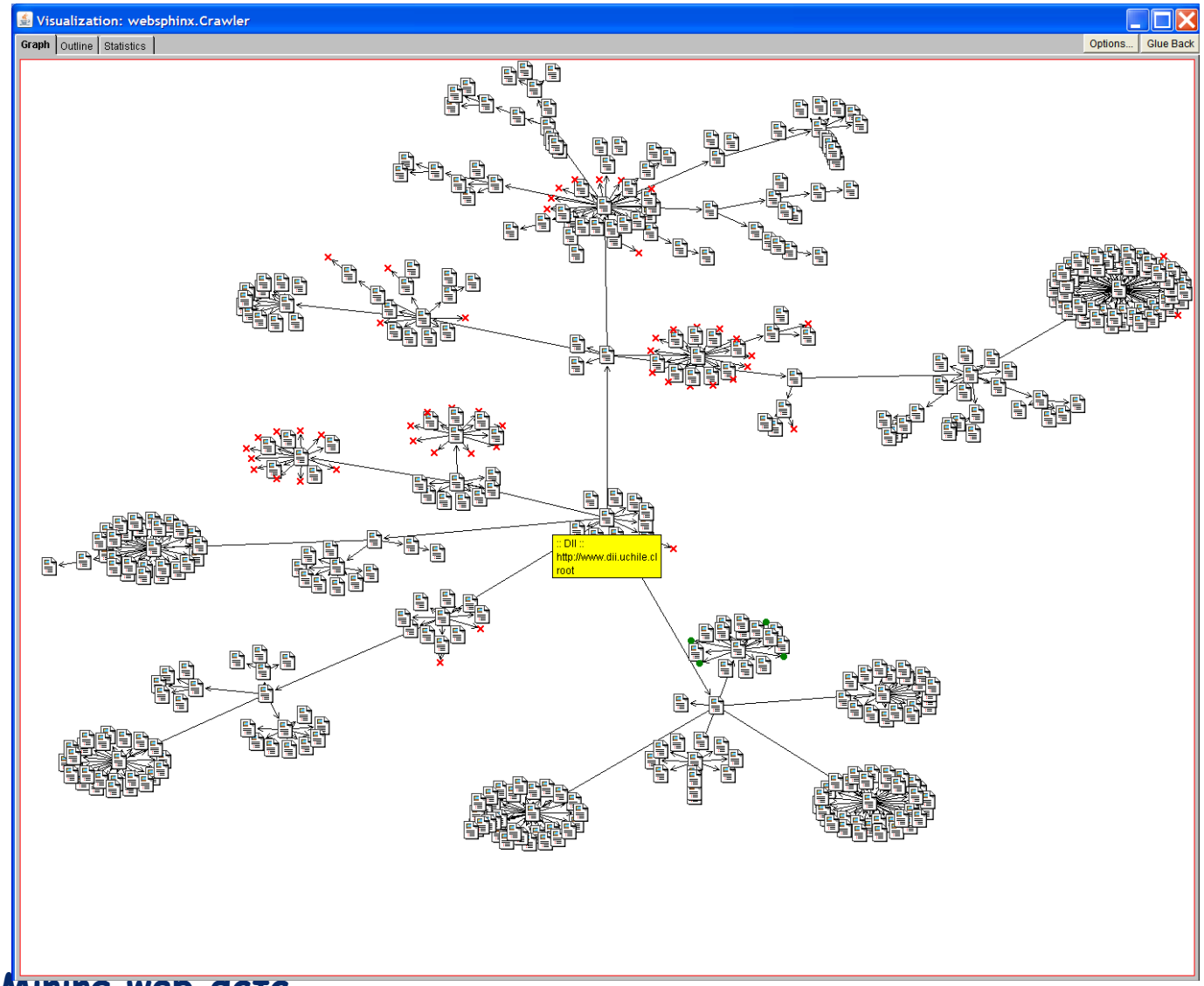


The Web Site: http://www.dii.uchile.cl



- Departmental information
- Department of industrial engineering
- Project pages
- Research group pages
- Personal pages
- Student organizations
- Web mail
- Over 3,500,000 raw registers per month

Partial view of web page links for DII site

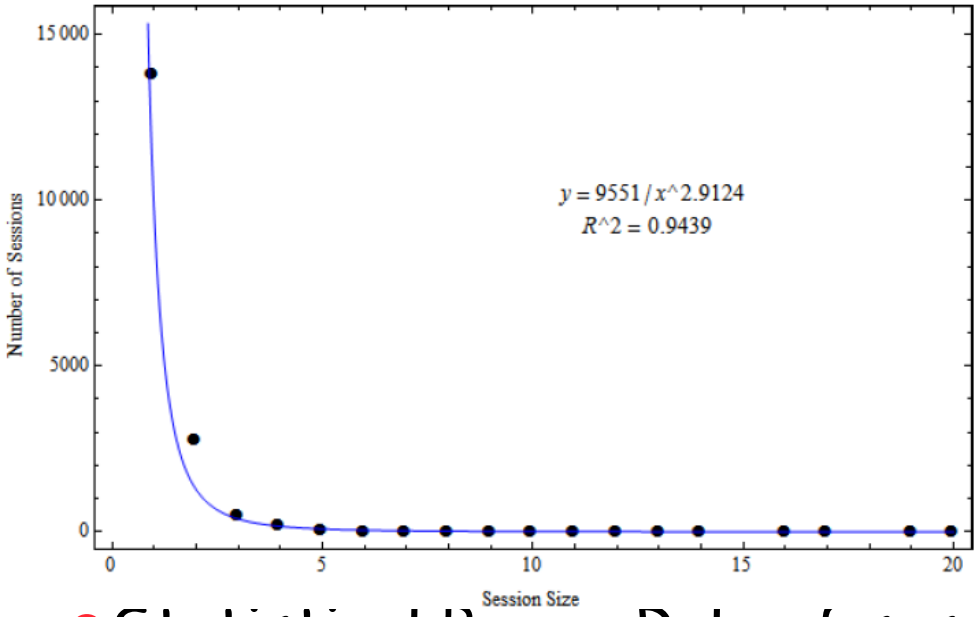


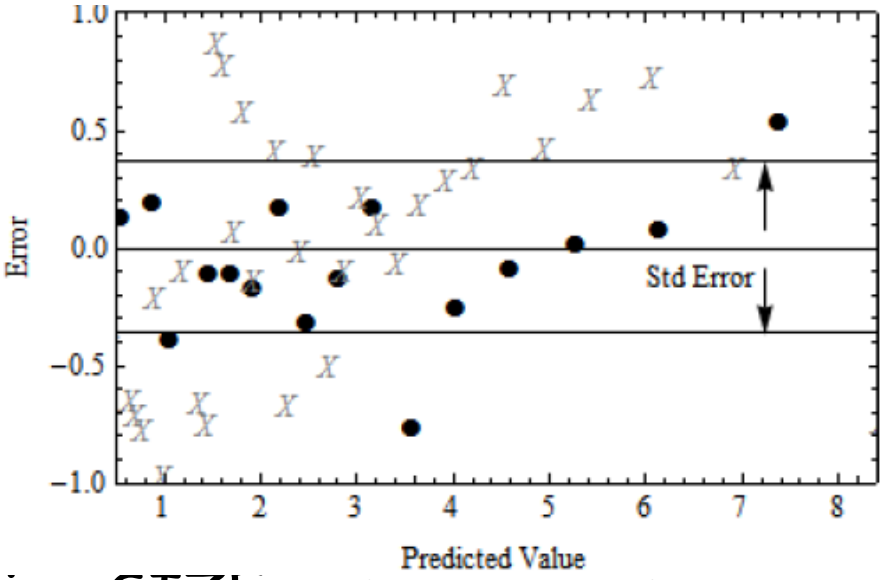
- 172 different text pages with links.
- 1,228 links between pages.

A network model to minimize the number of sessions

- Takes less than 10 minutes to solve the 403 chunks (each 30 minutes).
  - Overall it finds results similar to the Integer Program.
  - A total of 12,367 sessions.

Quality of a session





- Statistical Power Rule of session SIZE **[Huberman et al. 1998]**.
- QUALITY == FITING TO POWER RULE
- RESULT: R^2=0.94 and σ=0.38.
- OLD SESIONIZATION: R^2=0.91 and σ= 0.64.

Final comments

- What happen if the web page content is changed during the study period?
- A→B, B→D but there are two versions of D.
- If we want study the user behavior, it is necessary to consider to maintain a change register.
- Proposal solution LOGML [Punin WEBKDD'01]

Web text content

- From different web page content, special attention receive the free text.
- For the moment, a searching is performed by using key words.
- It is necessary to represent the text information in a feature vector, before to apply a mining process.
- The representation must consider that the words in the web page don't have the same importance.

Web text content: filters

- There are words that don't have information (article, prepositions, conjunctions, etc)
- It is necessary a cleaning process:
  - HTML tags.
  - Stop words (i.e. pronouns, prepositions, conjunctions, etc.)
  - Word stemming. Process of suffix removal, to generate word

Processing the web site: Vector space model [Salton75]

- The model associates a weight to each word in the page, based on its frequency in the whole web site.
- Let n_i and Q the amount of pages, a simple estimation of the relevance of a word is $w_{i,j} = n_i / Q$
- The *inverse document frequency* $IDF = \log(Q / n_i)$ can be used like a weight.
- *A variation of the last expression is known as TF*IDF*

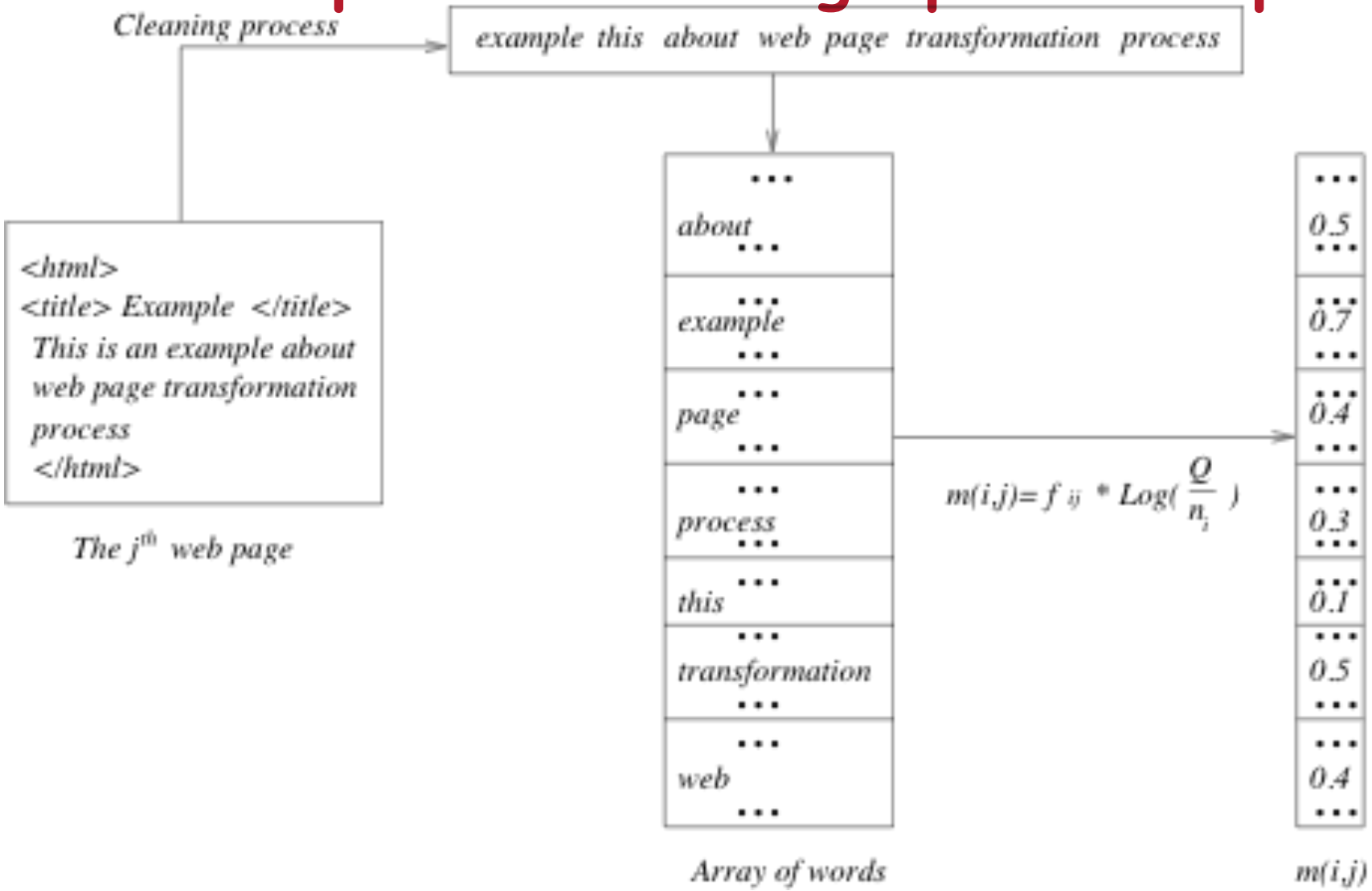
$$TF * IDF = f_{ij} * \log(Q / n_i)$$

Web page: vectorial representation

- Its vectorial representation would be a matrix of *RxQ*.
- *Q* is the number of pages in the web site and *R* is the number of different words in *P*.

	Word	1	2	…	Q
1	advise	1	0	…	1
2	business	0	1	…	0
.	…	.	.	…	.
.	…	.	.	…	.
.	…	.	.	…	.
.	…	.	.	…	.
.	…	.	.	…	.
R	zambia	1	0	…	0

Vector space model: a graphic example

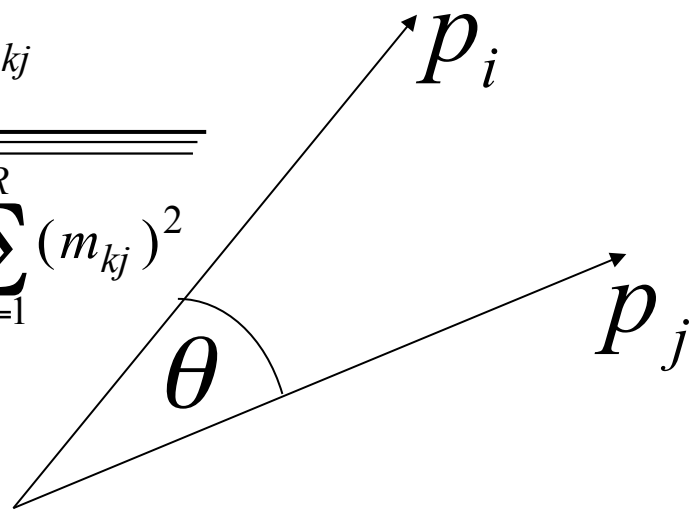


Comparing web pages

$$M = (m_{ij}) = f_{ij} * \log(\frac{Q}{n_i})$$

$$p_i \rightarrow (m_{1i}, \ldots, m_{Ri}) \qquad p_j \rightarrow (m_{1j}, \ldots, m_{Rj})$$

$$dp(p_i, p_j) = \cos\theta = \frac{\sum_{k=1}^{R} m_{ki} m_{kj}}{\sqrt{\sum_{k=1}^{R} (m_{ki})^2} \sqrt{\sum_{k=1}^{R} (m_{kj})^2}}$$



Vector Model, Summarized

- The best term-weighting schemes tf-idf weights:

  $w_{ij} = f(i,j) * log(Q/n_i)$

- For the query term weights, a suggestion is

  $w_{iq} = (0.5 + [0.5 * freq(i,q) / max(freq(l,q)]) * log(N / n_i)$

- This model is very good in practice:
  - tf-idf works well with general collections
  - Simple and fast to compute
  - Vector model is usually as good as the known ranking alternatives

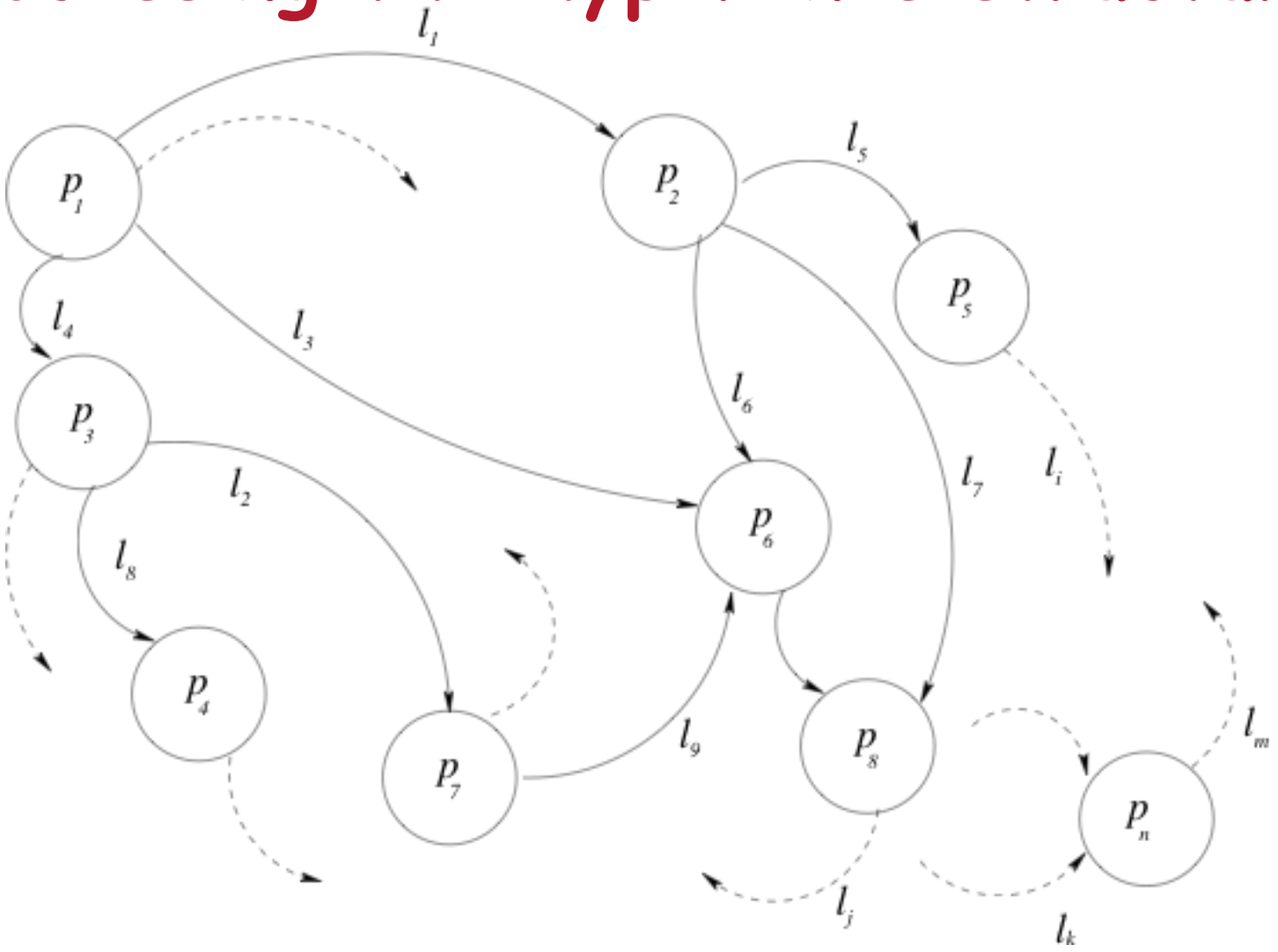
Pros & Cons of Vector Model

- Advantages:
  - term-weighting improves quality of the answer set
  - partial matching allows retrieval of docs that approximate the query conditions
  - cosine ranking formula sorts documents according to degree of similarity to the query
- Disadvantages:
  - assumes independence of index terms; not clear if this is a good or bad assumption

Web hyperlinks structure

- Why a web page point to another one?
- Link analysis: use link structure to determine credibility.
- If a web page is pointed by other ones, maybe it is because the page contains relevant information.
- We can understand the formation of a web community.
- We can improve our web site.

Processing the hyperlinks structure



Processing the hyperlinks structure (2)

- Identifying which pages contain more relevant information that others [Kleinberg99]:
  - Authorities. A natural information repository for the community.
  - Hub. These concentrate links to authorities web pages, for instance, "my favorite sites".

$$x_p = \sum_{q \;\; such \;\; that \;\; q \rightarrow p} y_q$$

$$y_p = \sum_{q \;\; such \;\; that \;\; q \rightarrow p} x_q$$

4.- Web mining

Data mining techniques

- Association rules
- Clustering
- Classification.
- Prediction.
- Probabilistic models.
- Regression.

What happen with web data?

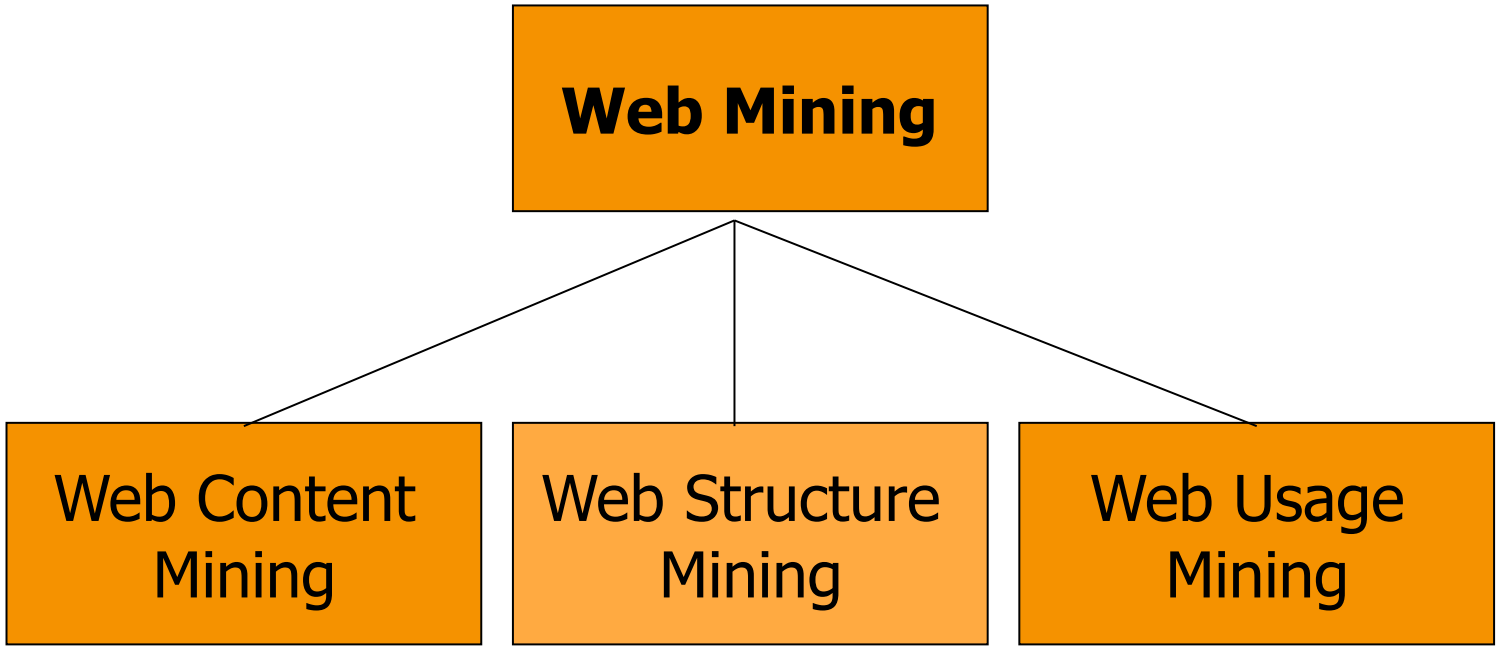
" The web is a huge collection of heterogeneous, unlabelled, distributed, time variant, semi-structured and high dimensional data"

S.K. Pal 2002

Mining the web

- Web mining techniques are the application of data mining theory in order to discovery patterns from web data.
- Web mining must consider three important steps:
  - Preprocessing.
  - Pattern discovery.
  - Pattern analysis.

Web Mining Taxonomy [Jooshi00,Lu03]



Web Structure Mining

- It deals with the mining of the web hyperlink structure (inter document structure).
- A website is represented by a graph of its links, within the site or between sites.
- Facts like the popularity of a web page can be studied, for instance, if a page is referred by a lot of other pages in the web.
- The web link structure allows to develop a notion of hyperlinked communities.
- It can be used by search engines, like Google or Altavista, in order to get the set of pages more cited for a particular subject.

WSM (2)

- To discover the link structure of the hyperlinks at the inter-document level to generate structural summary about the Website and Web page.
  - Direction 1: based on the hyperlinks, categorizing the Web pages and generated information.
  - Direction 2: discovering the structure of Web document itself.
  - Direction 3: discovering the nature of the hierarchy or network of hyperlinks in the Website of a particular domain.

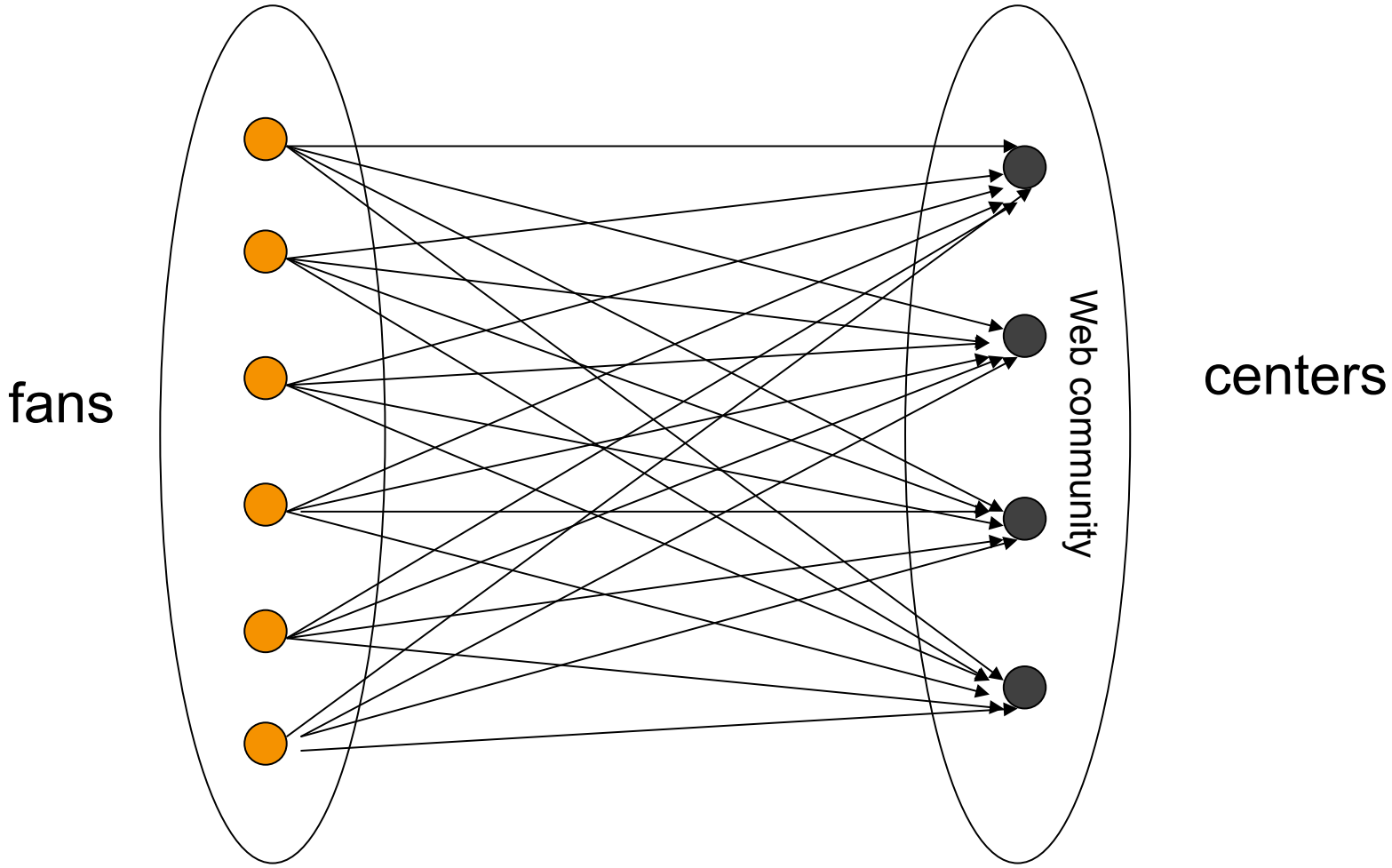
WSM (3)

- Finding authoritative Web pages
  - Retrieving pages that are not only relevant, but also of high quality, or authoritative on the topic
- Hyperlinks can infer the notion of authority
  - The Web consists not only of pages, but also of hyperlinks pointing from one page to another
  - These hyperlinks contain an enormous amount of latent human annotation
  - A hyperlink pointing to another Web page, this can be considered as the author's endorsement of the other page

WSM (4)

- Web pages categorization (Chakrabarti, et al., 1998)
- Discovering micro communities on the web
  - Example: Clever system (Chakrabarti, et al., 1999), Google (Brin and Page, 1998)
- Schema Discovery in Semi-structured Environment

WSM: Example

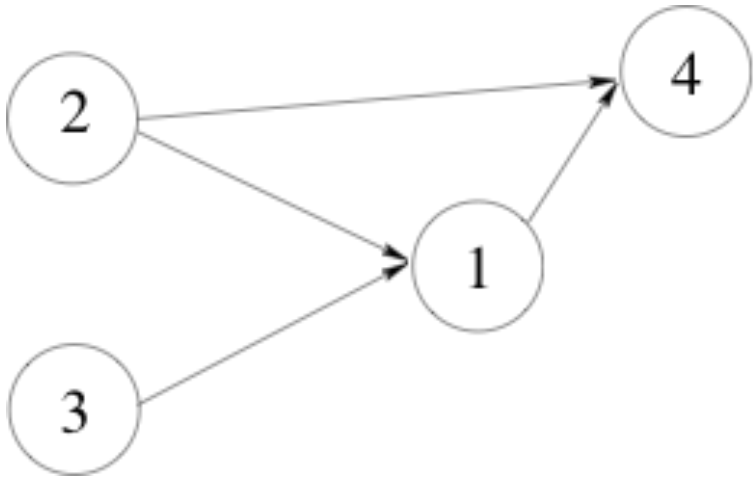


Web Community Centers: many web pages go there

HITS (Kleinberg99), PageRank (Google,Bring and Page98)

- Assumptions:
  - Credible sources will mostly point to credible sources
  - Names of hyperlinks suggest meaning
  - Ranking is a function of the *query terms and of the hyperlink structure*
- An example of why this makes sense:
  - The official Lord of the Rings site will be linked to by most high-quality sites about movies, Lord of the Rings, etc.
  - The biggest LoTR fan clubs probably are also frequently linked
  - A spammer who adds "Lord of the Rings" to his/her web site *probably* won't have many links to it

A simple web graph for a web community



Let $P = \{p_1, \ldots, p_n\}$ be the set of pages in the

Web community (in the example n=4)

HITS Algorithm

- Assumes that the authority came from in-edges.
- A good authority cam from good hubs and a good hub contains links that point to good authorities.
- A simple method to differentiate the page's relevance is by first assigning non-negative weights, depending if the page is hub or authoritative.
- Next, the weights are adjusted by an iterative process and the relative page's importance in the community is calculated.

HITS Algorithm (2)

- Let a_p and h_p be the weights associate to authority and hub pages, with $p \in P$. These weights can be calculated as

$$a_p = \sum_{\forall q, p \in P / q \to p} h_q$$

$$h_p = \sum_{\forall q, p \in P / q \to p} a_q$$

HITS Algorithm (3)

- Let $A = (a_1, \ldots, a_n)$ and $H = (h_1, \ldots, h_n)$ be the vector with the weights for authorities and hubs pages in the community.
- Then $A = M^T H \quad H = MA$
- Whit

$$M = (m_{ij}) = \begin{cases} 1 & i \rightarrow j \\ 0 & otherwise \end{cases}$$

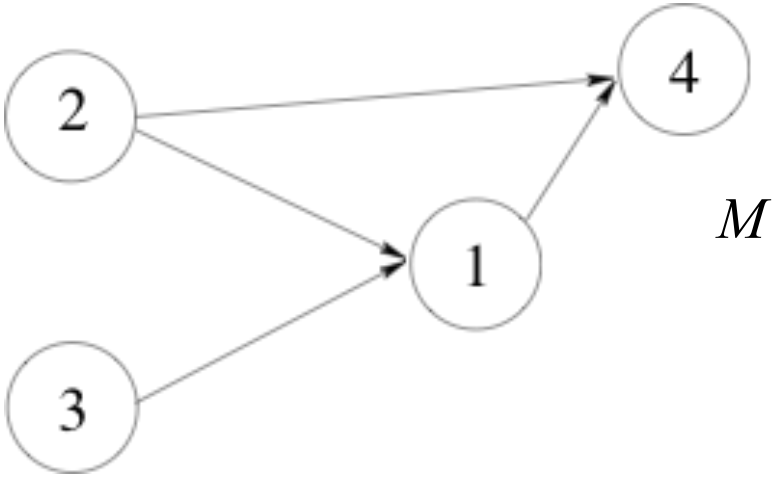
$$A^{(k+1)} \leftarrow M^T H^{(k)} = (M^T M) A^{(k)}$$

$$H^{(k+1)} \leftarrow MA^{(k)} = (MM^T) H^{(k)}$$

HITS Algorithm (4)

- Initialize $A = (1,\ldots,1), H = (1,\ldots,1)$
- Calculate
$$A^{(k+1)} \leftarrow M^T H^{(k)} = (M^T M) A^{(k)}$$
$$H^{(k+1)} \leftarrow M A^{(k)} = (M M^T) H^{(k)}$$
- Normalize A and H.
- If $A^{(k+1)} \approx A^{(k)}, H^{(k+1)} \approx H^{(k)}$ stop.
- Else $A^{(k)} = A^{(k+1)}, H^{(k)} = H^{(k+1)}$ go to point 2.

HITS Algorithm (5)



$$M = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad and \quad M^T = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

$$M^T M = \begin{bmatrix} 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2 \end{bmatrix} \quad and \quad MM^T = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

HITS Algorithm (6)

$$A^{(0)} = H^{(0)} = (1,1,1,1)$$

$$A^{(1)} = \begin{pmatrix} 0.5 \\ 0 \\ 0 \\ 0.5 \end{pmatrix}, A^{(2)} = \begin{pmatrix} 0.4 \\ 0 \\ 0 \\ 0.6 \end{pmatrix}, A^{(3)} = \begin{pmatrix} 0.47 \\ 0 \\ 0 \\ 0.53 \end{pmatrix}, A^{(4)} = \begin{pmatrix} 0.49 \\ 0 \\ 0 \\ 0.51 \end{pmatrix}$$

$$H^{(1)} = \begin{pmatrix} 0.33 \\ 0.44 \\ 0.22 \\ 0 \end{pmatrix}, H^{(2)} = \begin{pmatrix} 0.37 \\ 0.43 \\ 0.2 \\ 0 \end{pmatrix}, H^{(3)} = \begin{pmatrix} 0.37 \\ 0.43 \\ 0.19 \\ 0 \end{pmatrix}$$

- In general the algorithm will converge in few iterations (around five [Chakrabarti98]).

HITS Algorithm (7)

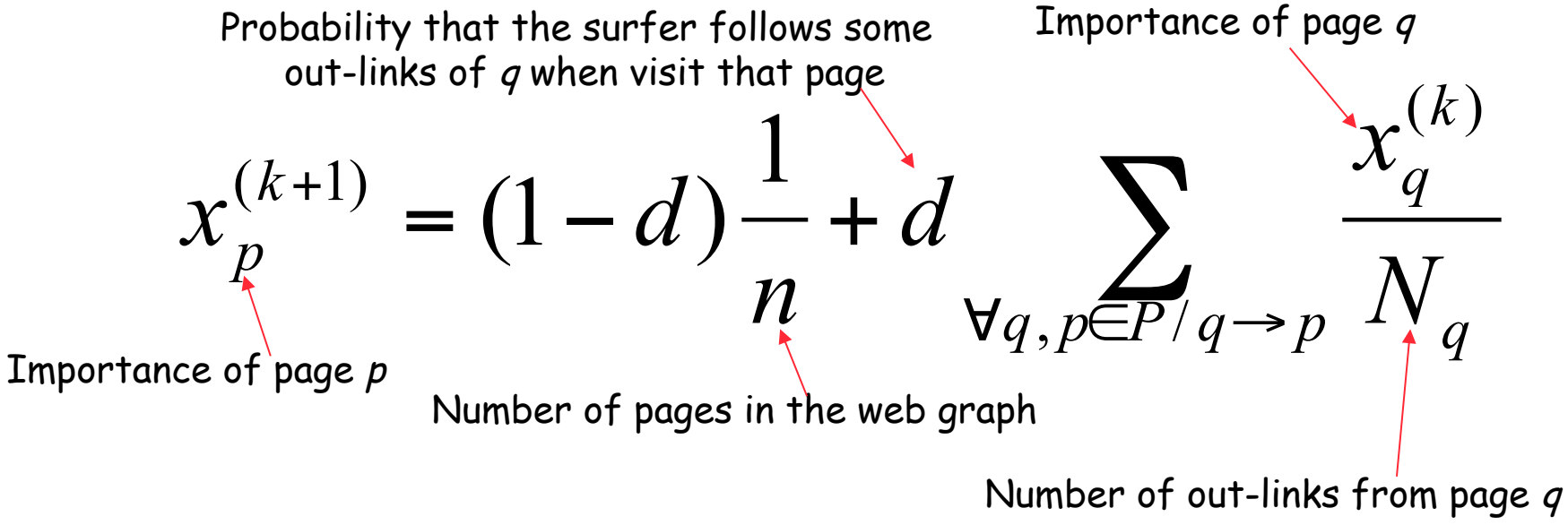
- By construction of HIST, after using the query to extract the pages related with the query's terms, the page's text content is ignored in the page's rank task, being the algorithm purely hyperlink-based computation.
- CLEVER system [Chakrabarti98b] address the problem by considering query's terms in the calculus of the above equations.
- In each link, a non-negative weight, whose initial value is basis on the text that surround the hyperlink expression (href tag in HTML) (more details in [Chakrabarti98a]).

Google's PageRank (Brin/Page 98)

- Mine structure of *web graph* independently of the query!
  - Each web page is a node, each hyperlink is a directed edge
- Assumes a random walk (surf) through the web:
  - Start at a random page
  - At each step, the surfer proceeds
    - to a randomly chosen web page with probability *d*
    - to a randomly chosen successor of the current page with probability 1- *d*
- **The PageRank of a page p is *the fraction of steps the surfer spends at p in the limit***

PageRank Algorithm

- The assumption is the importance of a page is given for the importance of the pages that pointed it.



PageRank Algorithm (2)

- By using a matrix representation, the PageRank equation is rewrite as:

$$X^{(k+1)} = (1-d)D + dMX^{(k)}$$

- where

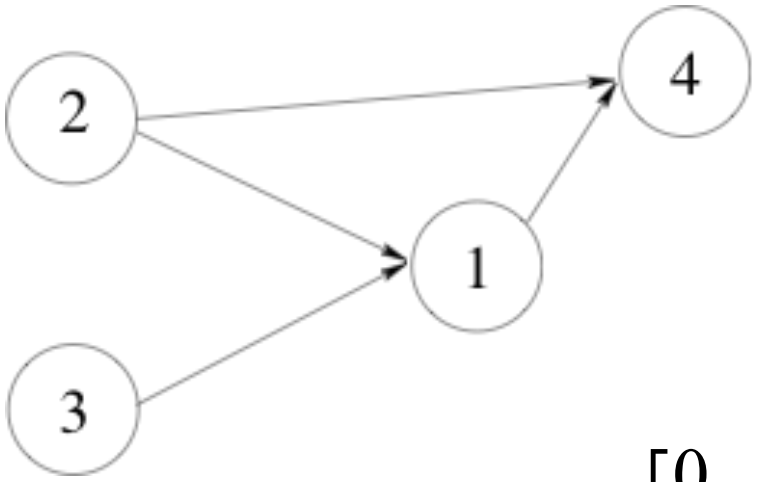
$$X^{(k)} = (x_{p_1}^{(k)},...,x_{p_n}^{(k)}), D = (\frac{1}{n},...,\frac{1}{n}) \quad and \quad M = (m_{ij} = \frac{1}{N_j})$$

- With N_j amount of out-links from page j such as $j \rightarrow i$

PageRank Algorithm (3)

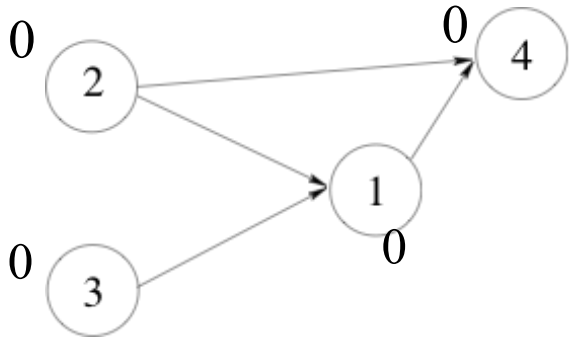
- Initialize $X^{(0)} = (0,...0) \quad and \quad D = (\frac{1}{n},...,\frac{1}{n})$
- Set *d*, usually d=0.85.
- Calculate $X^{(k+1)} = (1-d)D + dMX^{(k)}$
- If $\| X^{(k+1)} - X^{(k)} \| < \xi$ stop and return $X^{(k+1)}$.
- Else $X^{(k)} = X^{(k+1)}$ and go to point 3.

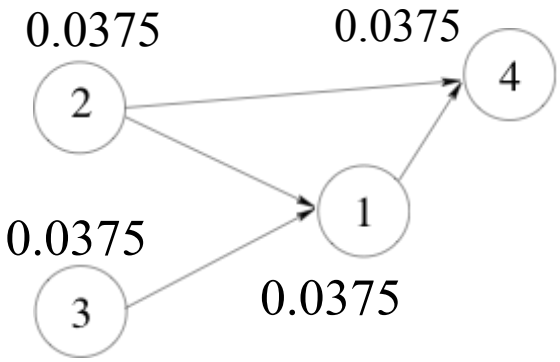
PageRank Algorithm (4)

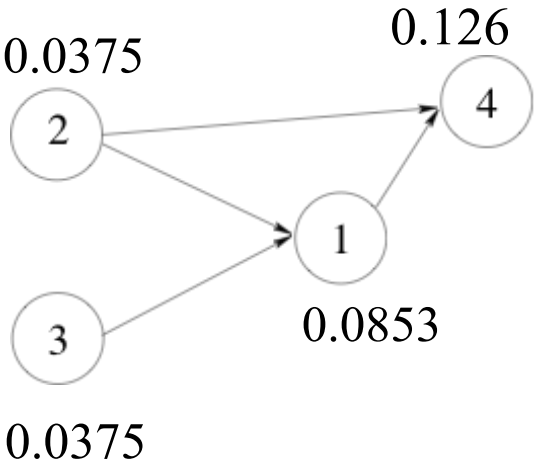


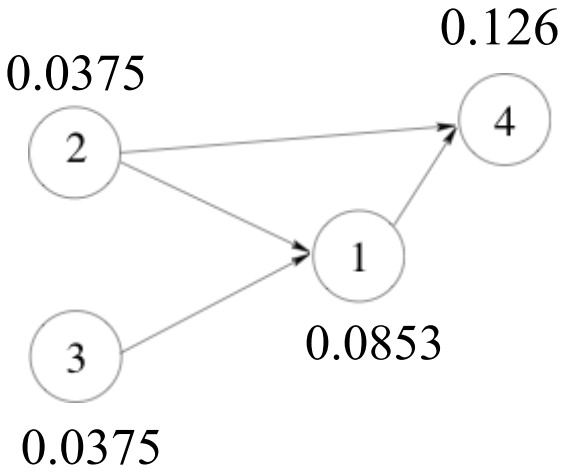
$$M = \begin{bmatrix} 0 & 0.5 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0.5 & 0 & 0 \end{bmatrix} \quad and \quad D = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}$$

PageRank Algorithm (5)









$$X^{(1)} = \begin{pmatrix} 0.0375 \\ 0.0375 \\ 0.0375 \\ 0.0375 \end{pmatrix}, X^{(2)} = \begin{pmatrix} 0.0853 \\ 0.0375 \\ 0.0375 \\ 0.126 \end{pmatrix}, X^{(3)} = \begin{pmatrix} 0.0853 \\ 0.0375 \\ 0.0375 \\ 0.126 \end{pmatrix}$$

PageRank Algorithm (6)

- About the disadvantages:
  - If a page is pointed by another one, it mean the page receive a vote for the PageRank calculus.
  - If a page is pointed by a lot of pages, it mean that the page is important.
- "Only the good pages are pointed by others one", but:
  - Reciprocal links. If the page A link page B, then page B will link page A.
  - Link Requirements. Some web pages give electronic gifts, like a program, document etc., if another page point it.
  - Near persons community. For instance friends and relatives that from their pages point another friend or relative, only for the human relationship between them.

Identifying web communities

- "*Set of sites that have more links (in either direction) to members of the community than non-members*" [Flake02]
- The web community identification have several practical applications, like focalized search engine, content filters, and complement of text-based searches [Staab05].
- The most important continue being the analysis of the entire Web for studying the relationship within and between communities, like scientific, research and in general *Social Networks* [Kumar02,Mika04]

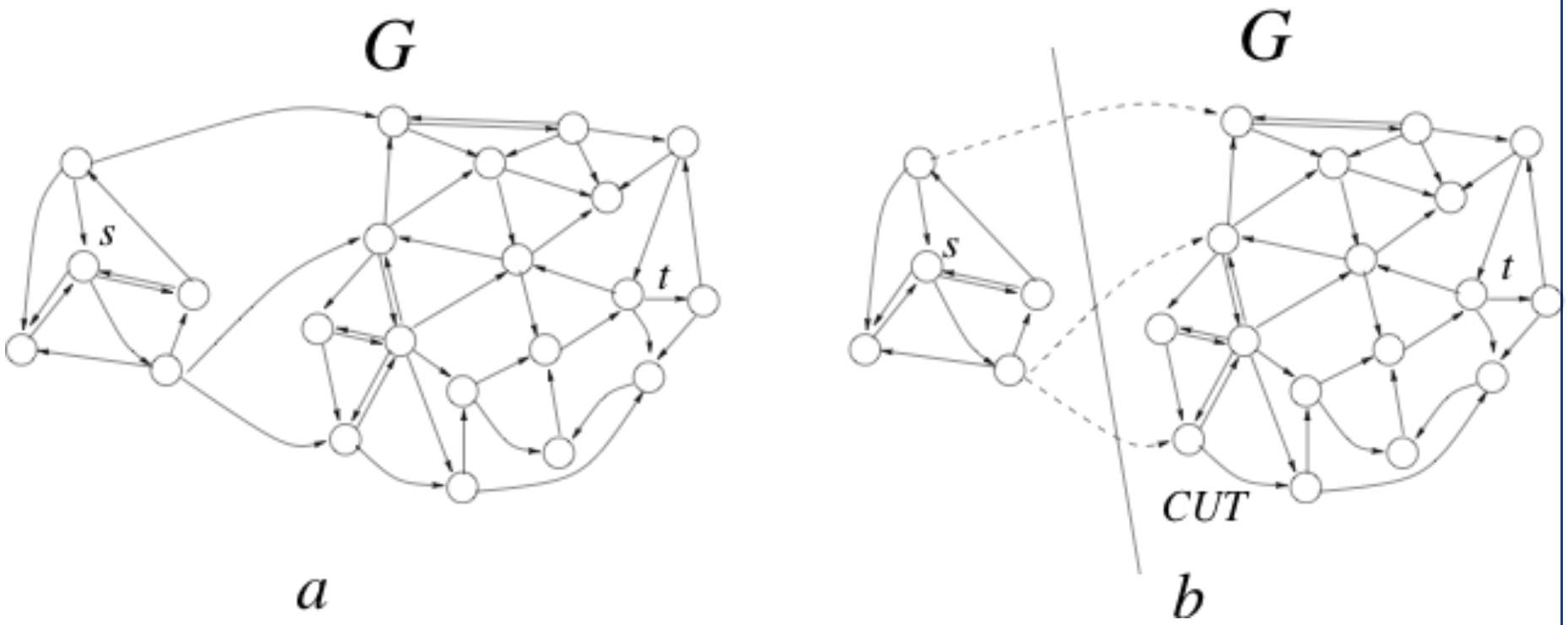
Identifying web communities (2)

- If the Web is represented as a directed graph or web-graph, then the "Max Flow-Min Cut" method [Ford56] can be used to identify web communities [Flake02].
- Let *G=(V,E)* be the directed web-graph with edge capacities $c(u_i, u_j) \in Z^+$ and two vertices $s, t \in V$
- A web community is defined as a set a vertex

$$U \subset V \,/\, \forall u_i \in U, \exists u_j \in U \,/\, u_i \rightarrow u_j \quad with \quad i \neq j \quad and \quad u_j \notin (V - U)$$

- Find the maximum flow that it is possible to route from *s* to the sink *t*, maintaining the capacity constrains.

Identifying web communities (3)



WSM summary

- HIST and PageRank use back-links as a means of adjusting the "worthiness" or "importance" of a page
- Both use iterative process over matrix/vector values to reach a convergence point
- HITS is query-dependent (thus too expensive to compute in general)
- PageRank is query-independent and considered more stable
- Just one piece of the overall picture…

Web Content Mining

- The goal is to find useful information from the web content.
- In this sense, WCM is similar to Information Retrieval (IR).
- However, web content is not only free text, other objects like pictures, sound and movies belong also to the content.
- There are two main areas in WCM :
  - the mining of document contents (web page content mining)
  - the improvement of content search in tools like search engines (search result mining).

WCM (2)

- Web content;
  - text, image, audio, video, metadata and hyperlinks.
- Information Retrieval View (Structured + Semi-Structured)
  - Assist / Improve information finding
  - Filtering Information to users on user profiles
- Database View
  - Model Data on the web
  - Integrate them for more sophisticated queries

WCM (3)

- Developing Web query systems
  - WebOQL, XML-QL
- Mining multimedia data
  - Mining image from satellite (Fayyad, et al. 1996)
  - Mining image to identify small volcanoes on Venus (Smyth, et al 1996) .

Issues in Web Content Mining

- Developing intelligent tools for IR
  - Finding keywords and key phrases
  - Discovering grammatical rules and collocations
  - Hypertext classification/categorization
  - Extracting key phrases from text documents
  - Learning extraction models/rules
  - Hierarchical clustering
  - Predicting (words) relationship

Web page in vector space model

- Each web page can be consider as a document text with tags.
- Applying filters, the web page is transformed to feature vectors.
- Let $P = \{p_1, \ldots, p_Q\}$ b the set of Q pages in a web site.
- The i-th page is represented by

$$wp^i = (wp_1^i, \ldots, wp_R^i) \in WP$$

with R the amount of words after a stop word and stemming process.

Classification

- A text classifier . Given a document d and return a scalar value with a category $c_i \in C / \bigcup c_i = C$ [Sebastiani99].
- The function is known as "Categorization Status Value''
  $CSV_i : D \rightarrow [0,1]$
- The $CSV_i(d)$ takes un different expressions, according with the classifier in use.
- For instance, it can be a probability approach [Lewis92] basis on Naive Bayes theorem or a distance between vectors in a r-dimensional space [Schutze95].

Classification (2)

- The classification was implemented by semi-automatic of full-automatic [Asirvatham05] approaches, like Neighbor [Kwon03] Bayessian models [McCallum98] , Support Vector Machines [Joachims97] , Artificial Neural Networks [Honkela97] and Decision Trees [Apte04] .
- The web pages classification algorithms can be grouped in [Asirvatham05] :
  - Manual categorization
  - Applying clustering approaches. Previous to classify the web pages, a clustering algorithm is used to find the possible clusters in a training set.
  - Meta tags. It use the information contained in the web page tags (<META name=`` keywords''> and <META name=`` description'' >).
  - Text content based categorization.
  - Link and content analysis. It is basis on the fact that the hyperlink contain the information about which kind of pages is pointed (href tag)

Clustering

- To group web pages allows perform efficient searching task and semi-automatic or full-automatic document's categorizations.
- The clustering techniques need a similarity measure in order to compare two vectors by common characteristics [Strehl00] .
- It is necessary a similarity of distortion measure to compare the vectors in a training set.
- For instance a simple distance like the angle's cosine between two pages in a vector representation.

$$dp(p_i, p_j) = \cos\theta = \frac{\sum_{k=1}^{R} m_{ki} m_{kj}}{\sqrt{\sum_{k=1}^{R} (m_{ki})^2} \sqrt{\sum_{k=1}^{R} (m_{kj})^2}}$$

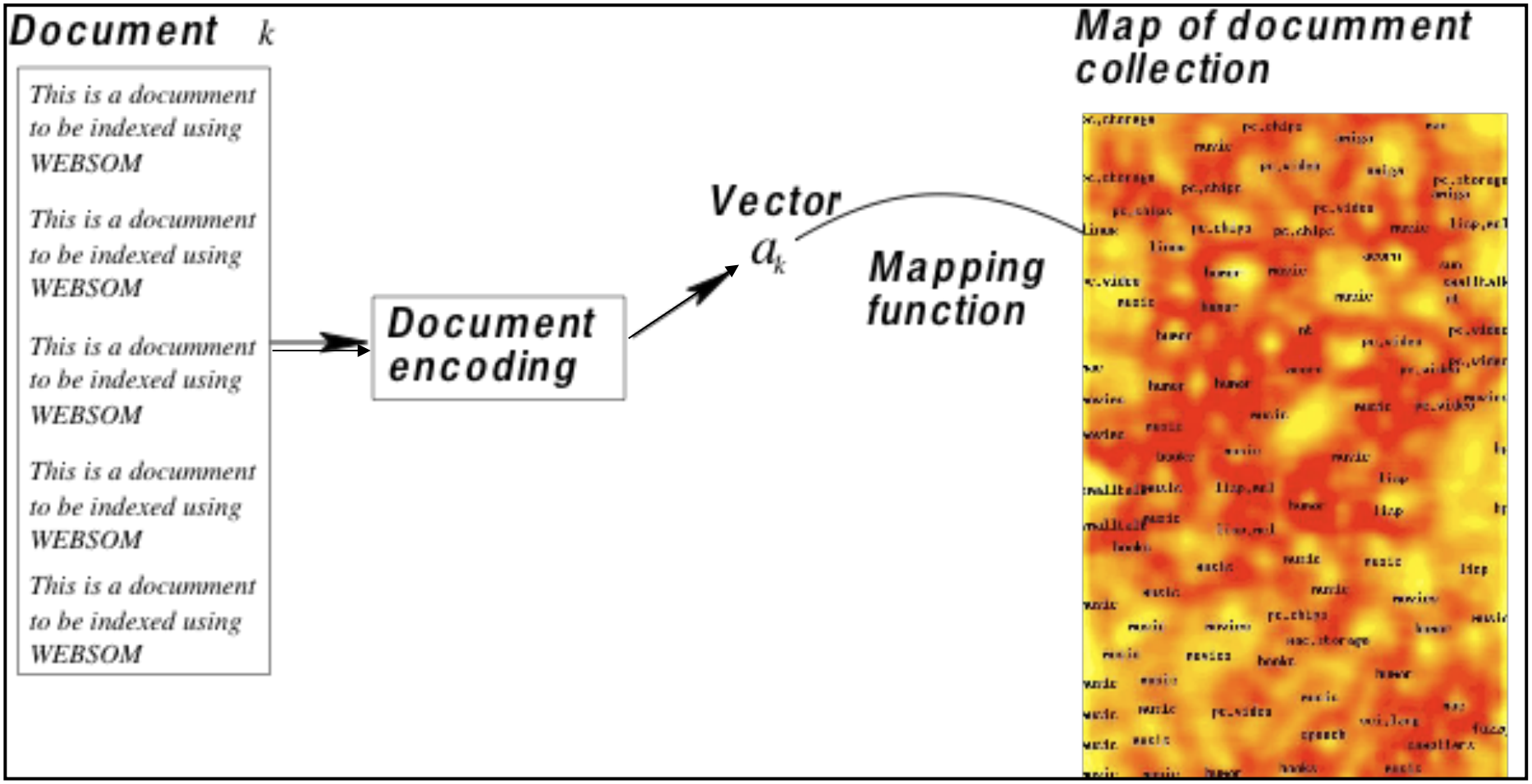
Clustering (2)

- For document clustering, more complex and semantic based similarity have been proposed [Strehl00] .
- Let $C = \{c_1, \ldots, c_l\}$ be the set of l clusters extracted from WP.
- Since the hard clustering point of view $\exists! c_k \in C / wp^i \in c_k$
- Whereas in soft clustering, a vector can belong to two or more clusters [Karypis99, Koutri04] .
- Several document clustering algorithms have appeared in the last years [Feldman95, Willet88] .
- An interesting approaches is the utilization of K-means and its variations in overlapping clusters, known as Fuzzy C-means [Jang97] .

WEBSOM

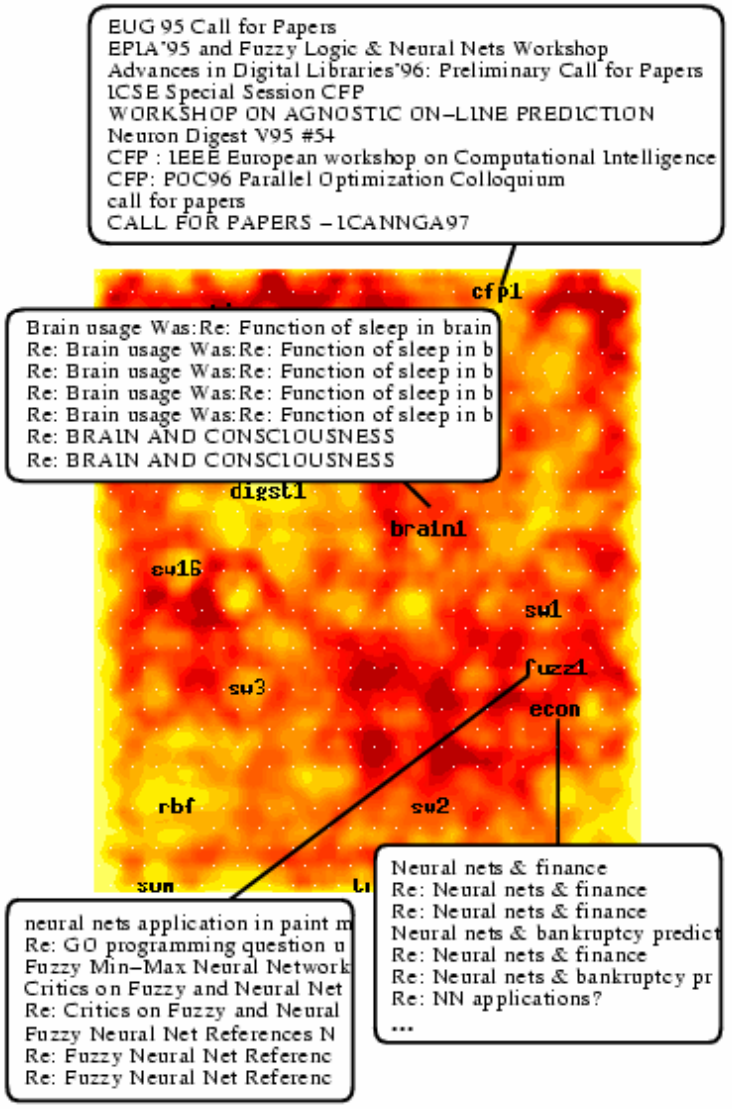
- It is a means for organizing miscellaneous text documents into meaningful maps for exploration and search.
- It is based on SOM (Self-Organizing Map) that automatically organizes documents into a two-dimensional grid so that related documents appear close to each other
- http://www.cis.hut.fi/websom

WEBSOM


Document k
This is a documment to be indexed using WEBSOM
This is a documment to be indexed using WEBSOM
This is a documment to be indexed using WEBSOM
This is a documment to be indexed using WEBSOM
This is a documment to be indexed using WEBSOM
Document encoding
Vector
a_k
Mapping function
Map of documment collection

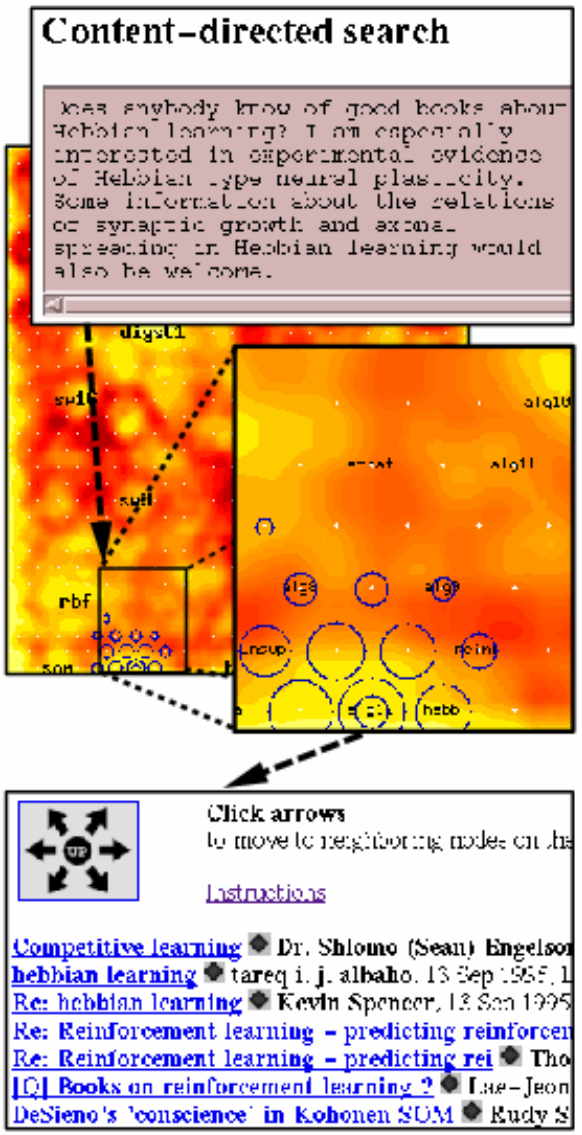
A map of documents

- A set of documents related with neural networks is mapped by using WEBSOM method.
- Browsing for the interface, it is possible to see the "labels" for documents



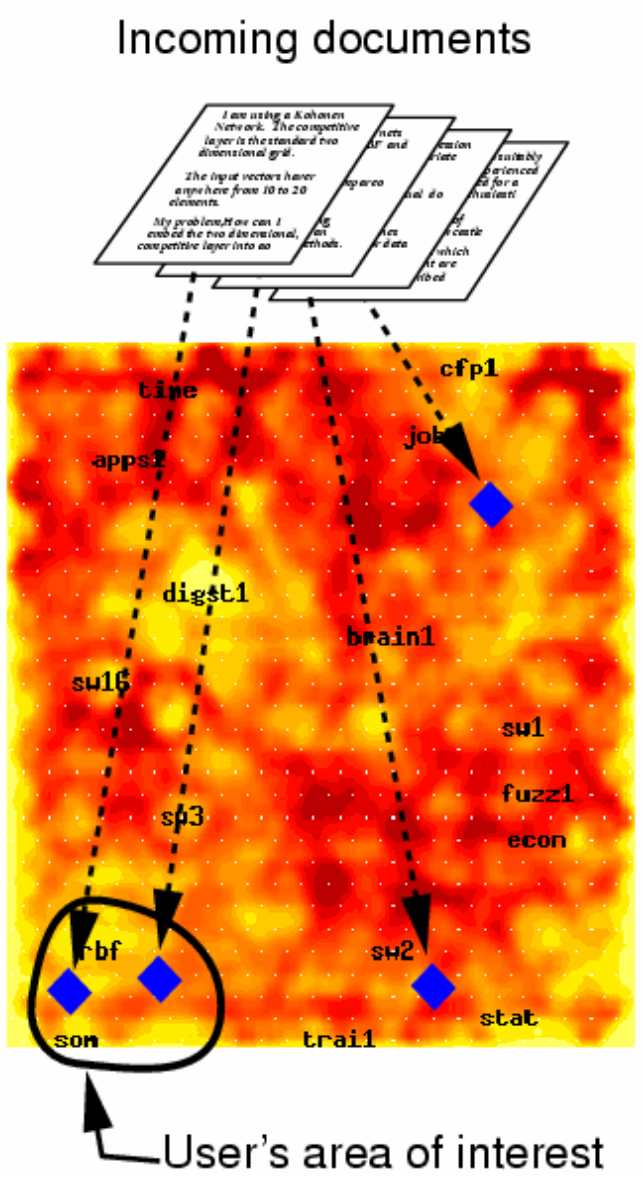
Sample search

- A new document or any document's description can be used for finding related documents.
- The circle on the map denote the location of the most representative document for the question.



How to use the Maps

- As filter that notifies the user of interesting documents.
- As a searching engine.
- A new index method.



Automatic web page text summarization

- The goal is to construct automatically summaries of a natural-language document [Hahn00] .
- In many case the web pages only contain few words and the page could contain non-textual elements (e.g. video, pictures, audio, etc.) [Amitay00] .

- In text summarization research, there are three major approaches [Mani99] :
  - paragraph based.
  - sentence based.
  - Using natural language cues in the text.

Extraction of key-text components from web pages

- The key-text components are parts of an entire document, for instance a paragraph, phrase and inclusive a word, that contain significant information about a particular topic, from the web site user point of view.
- A web site keyword is "*a word or possibly a set of words that make a web page more attractive for an eventual user during his visit to the web site''[Velasquez05b].*
- The assumption is that there exists a correlation between the time that the user spent in a page and his/her interest in its content [Velasquez04b]'.
- Usually, the keywords in a web site have been related with the `` most frequently used words''.
- In [Buyukkokten01] a method to extract keywords from a huge set of web pages is introduced.

WCM summary

- The vector space model is a recurrent method to represent a document as a feature vector.
- Because the set of words used in the construction of the web site could be a lot, it is necessary to apply a stop word cleaning and stemming process.
- A web page content is different to a common document, in fact, a web page contain semi-structured text, i.e., with tags that give additional information about the text component.
- Also a page could contain pictures, sounds, movies, etc.
- Sometime the page text content is a short text or inclusive a set of unconnected words.

Web Usage Mining (WUM)

- Also known as Web log mining
  - mining techniques to discover interesting usage patterns from the secondary data derived from the interactions of the users while surfing the web

WUM: Considerations [Pierrakos03]

- WUM applies traditional data mining methods in order to analyze usage data.
- The sessionization process is necessary to correct the problems detected in the data.
- The goal is to discover patterns in usage data applying different kinds of data mining techniques.
- Applications of WUM can be grouped in two main categories:
  - User modeling in adaptive interfaces, known as personalization.
  - User navigation patterns, in order to improve the web site structure.

WUM: Applications

- Target potential customers for electronic commerce
- Enhance the quality and delivery of Internet information services to the end user
- Improve Web server system performance
- Identify potential prime advertisement locations
- Facilitates personalization/adaptive sites
- Improve site design
- Fraud/intrusion detection
- Predict user's actions (allows prefetching)

Statistics

- Several tools use the conventional statistics for analyzing the user behavior in a web site (http://www.accrue.com/, http://www.netgenesis.com, http://www.webtrends.com).
- Each one of them use graphics interfaces, like histograms, with statistics associate.
- For instance amount of clicks per page during the last month.
- By applying conventional statistics on web logs, one can perform different kinds of analysis [Boving04] .

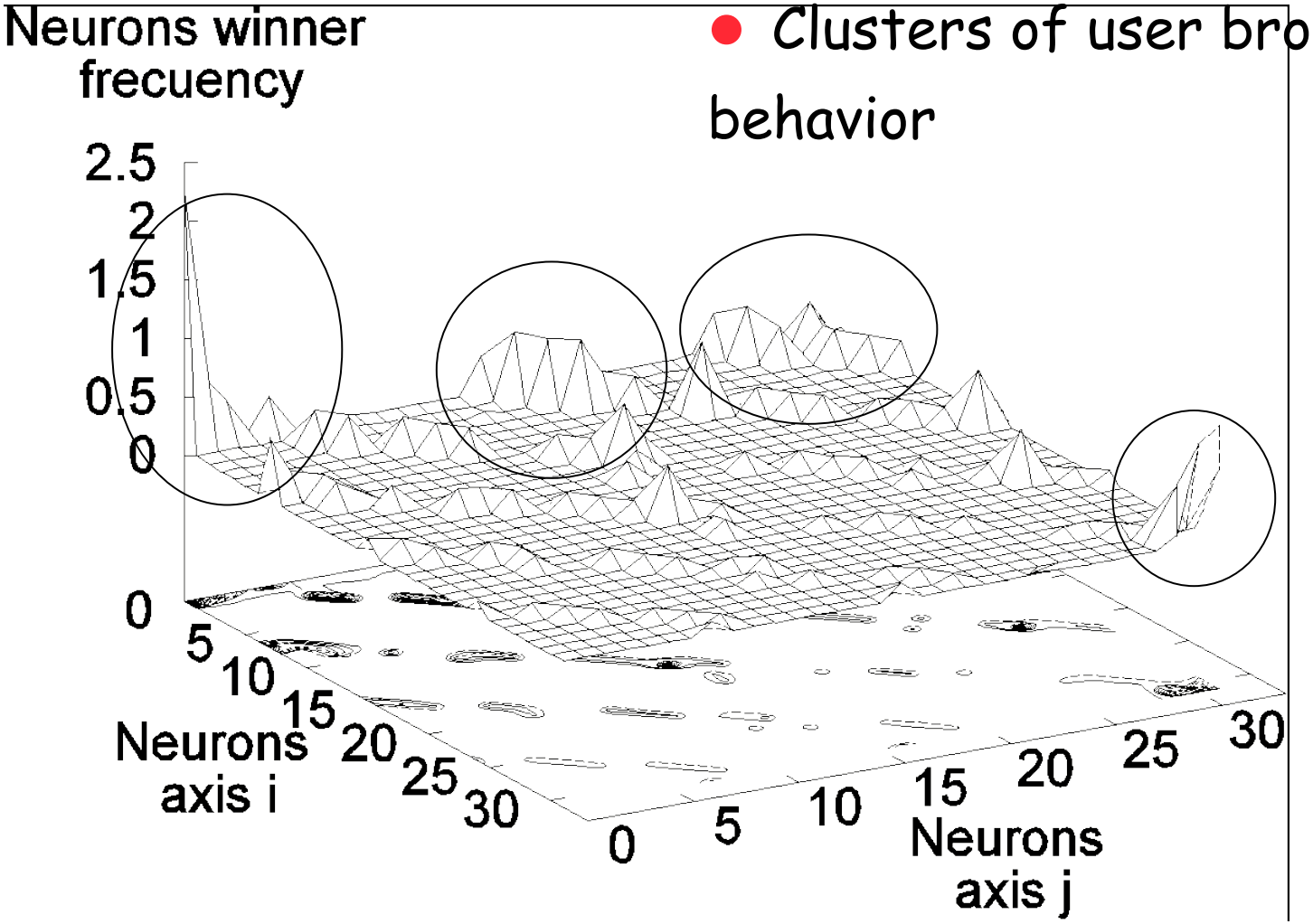
Statistics (2)

- These basic statistics are used to improve the web site structure and content in many institutions, over all the commercial ones [Dellmann04] .
- Their application is simple, there are many web traffic analysis tools and it easy to understand the tools's reports.
- Although the statistic analysis could seem a very simple tool to mining the web logs, it resolve a lot of problems relate with the performance of the web site.

Clustering

- Groups together a set of items having similar characteristics
- User Clusters
  - Discover groups of users exhibiting similar browsing patterns
  - Page recommendation
    - User's partial session is classified into a single cluster
    - The links contained in this cluster are recommended

Clustering (2): Example



- Clusters of user browsing behavior

Clustering (3)

- Grouping users with common browsing behavior and pages with similar content [Srivastava00] .
- An important thing in any clustering technique, is the similarity measure or method used to compare the input vectors for creating the clusters.
- In the WUM case, the similarities are constructed basis on the usage data in the web site, i.e., the set of pages used or visited during the user session.

Clustering (4)

- Because there are several usage vector representations, each similarity must consider the particular user navigation model expressed in the feature vector.
- In [Xiao01] , the usage vector is represented by the set of pages visited during the session.
- In [Joshi00] also the pages visited are consider in the similarity, but the structure of the web site, as well as the URL's are involved.
- In [Hay04] and [Runkler03] , additionally to the usage page data, the sequence of page visited is incorporated in the similarity.

Clustering (5)

- In [Velasquez04a] it is proposes that beyond the page sequence, the time spent per page during the session and the text content in each page must be consider in the similarity measure.
- Several clustering algorithms have been applied to extract patterns from web site usage data [Koutri04] .
- In [Velasquez03] a Self Organizing Feature Map in toroidal topology is used for analyzing the user behavior.

Clustering (6)

- In [Abraham03] , where the ant colonies behavior and their self-organizing capabilities are used to model the user browsing behavior in a web site.
- A conceptual clustering approach [Fisher87, Fisher96] is used in [Perkowitz98] for tackling the index page synthesis problem.
- In [Joshi00, Runkler03] the c-means model [Ball67] is used for creating navigation's clusters.

Classification

- Before to apply a classification process, it is necessary to define the predetermined classes.
- In the web usage mining, mainly the classes describe user's category [Eirinaki03] .
- The decision rule induction have been one of the classification approaches more widely used web usage mining.
- In [Ngu97] HCV, an heuristic attribute-based induction algorithm, is used for classifying the pages visited and keyword used by the user for search tasks.
- The practical result is a set of rules that represent the users' interests.

Classification (2)

- In the decision tree induction approach, the pages visited by the user are consider as positive examples for the induction of Page Interest Estimators (PIE).
- The trees can be constructing by using several algorithms, such as C4.5 [Quinlan93] .
- Another interesting approach is the Naïve Bayesian classifier [Chan00] .
- In order to offer personalized web-based system for web users, in [Yuan04] is implemented a classification technique basis on a Fuzzy Neural Networks (FNNs).

Association Rule Generation

- Discovers the correlations between pages that are most often referenced together in a single server session
- Provide the information
  - What are the set of pages frequently accessed together by Web users?
  - What page will be fetched next?
  - What are paths frequently accessed by Web users?
- Association rule

$$A \longrightarrow B \; [\text{ Support} = 60\%, \text{Confidence} = 80\% \;]$$

Example

"50% of visitors who accessed URLs /diplomas.html and BI/info.html also visited webmining.html"

Association rules

- In WUM, the association rules are focus mainly in the discovery of relations between pages visited by the users [Mobasher01] .
- For instance, an association rule for a MBA program is

  mba/seminar.html ➔ mba/speakers.html
- From a different point of view, in [Schwarzkopf01] a Bayesian network is used for defining taxonomic relations between topics shown in a web site.

Association rules (2)

- Another interesting approach is the utilization of fuzzy association rules for web access path prediction [Wong01] .
- The method applies the case-based reasoning approach on user session extracted from web logs files.
- In this approaches, the time duration is include as an attribute of the web access case.

Sequence patterns

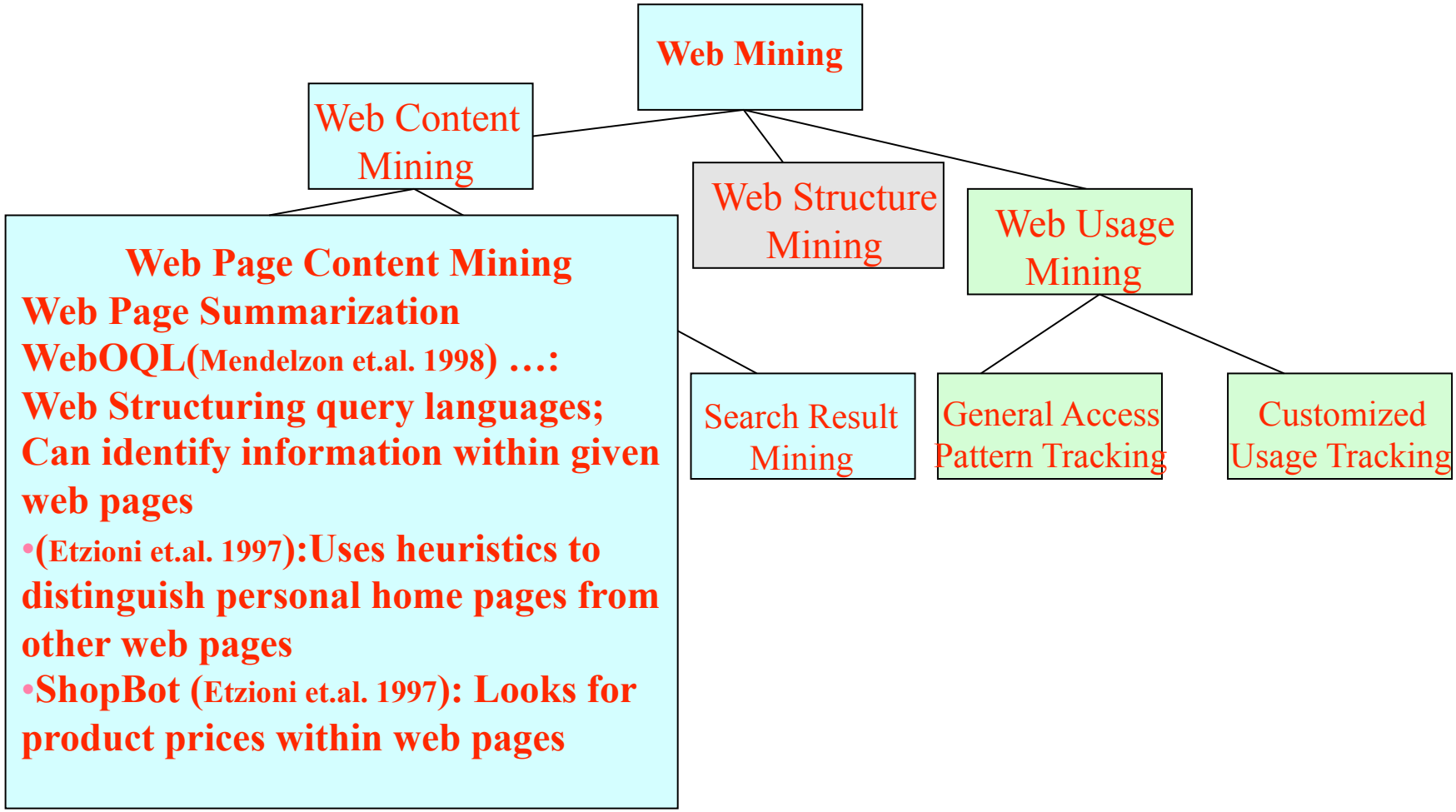
- Discovering frequent subsequences in a set of sequential data.
- Main idea:To find sequential navigation patterns in the user's sessions.
- For instance, the 60% of the user who visit mba/index.html and mba/speakers.html, also in the same session visited mba/seminar.html.
- In sequential patterns, two methods have been used: deterministic and stochastic techniques.
- Deterministic approach [Mortazavi01] .
- Here the user navigation behavior is used for sequential patterns discovery, such us the case of Web Utilization Mining tool [Spiliopoulou99] .

Sequence patterns (2)

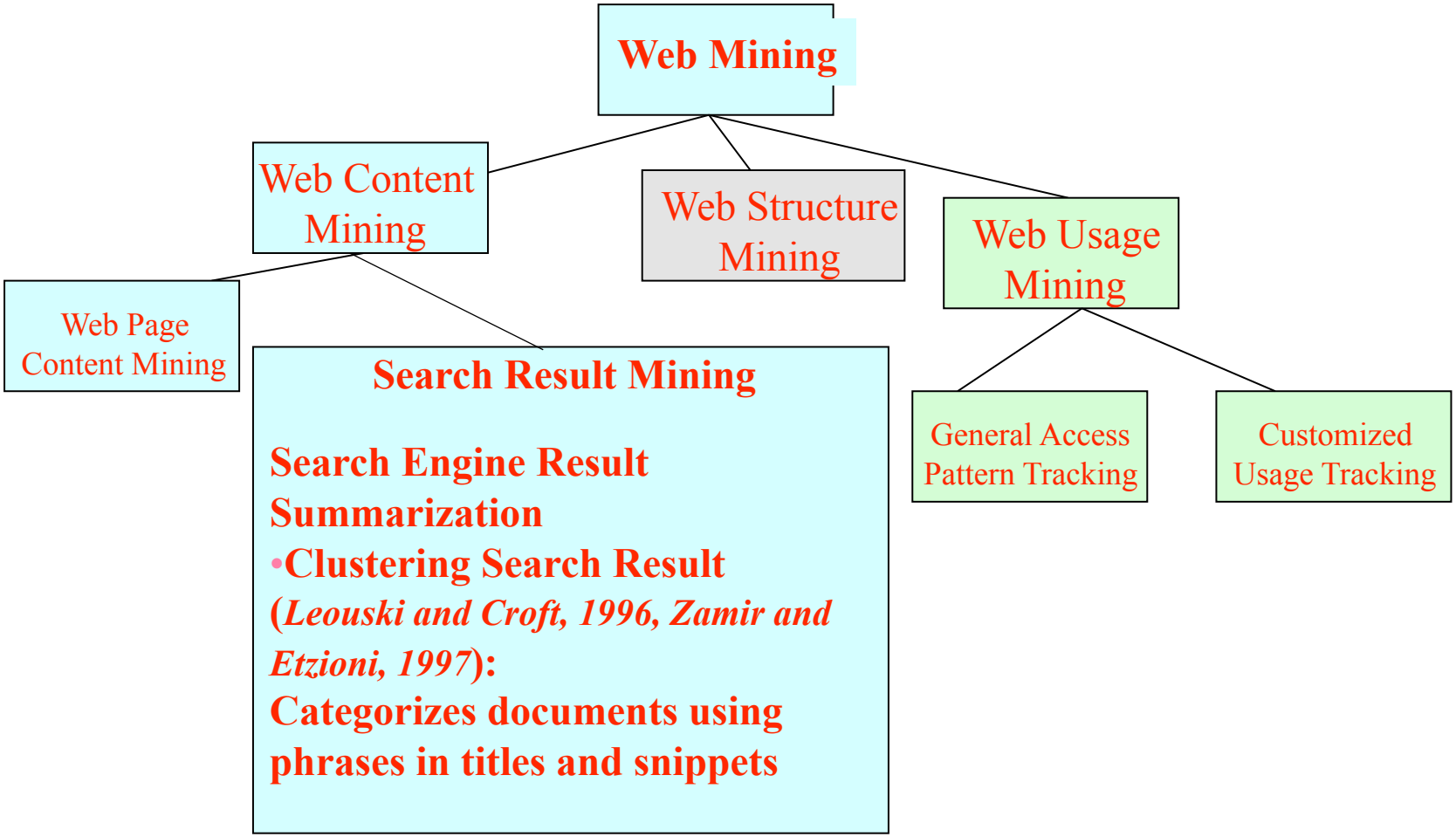
- Another interesting approach to extract sequential patterns is the utilization of clustering techniques [Velasquez05a].
- Stochastic approach. Is the application of Markov Model for sequential pattern discovery task [Bestavros95] .
- Another approach for prediction of the subsequent visits [Borges99] .
- The user session are modeled using hypertext probabilistic grammar.

Web Mining Taxonomy

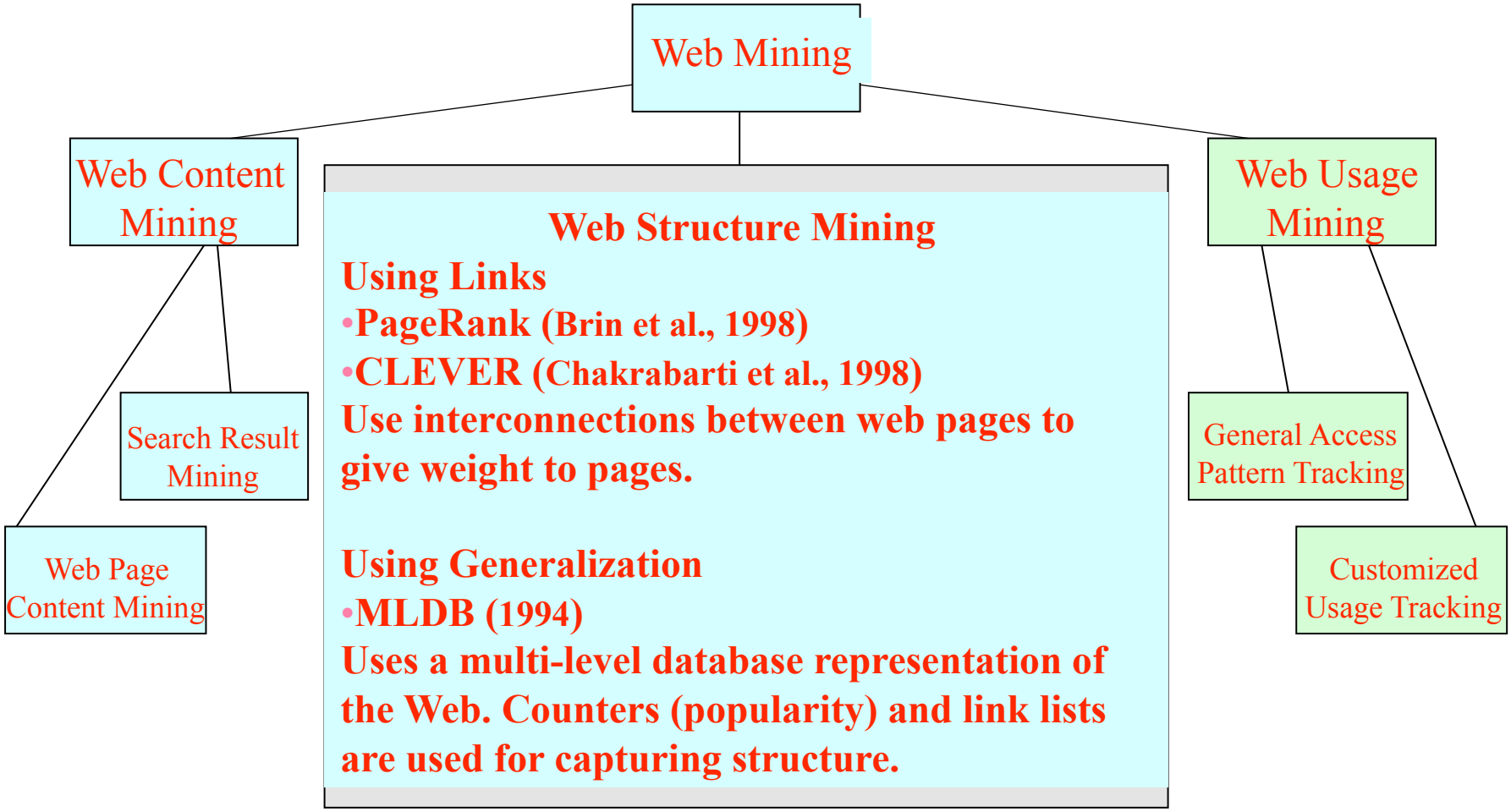
Mining the World Wide Web [Madria99]



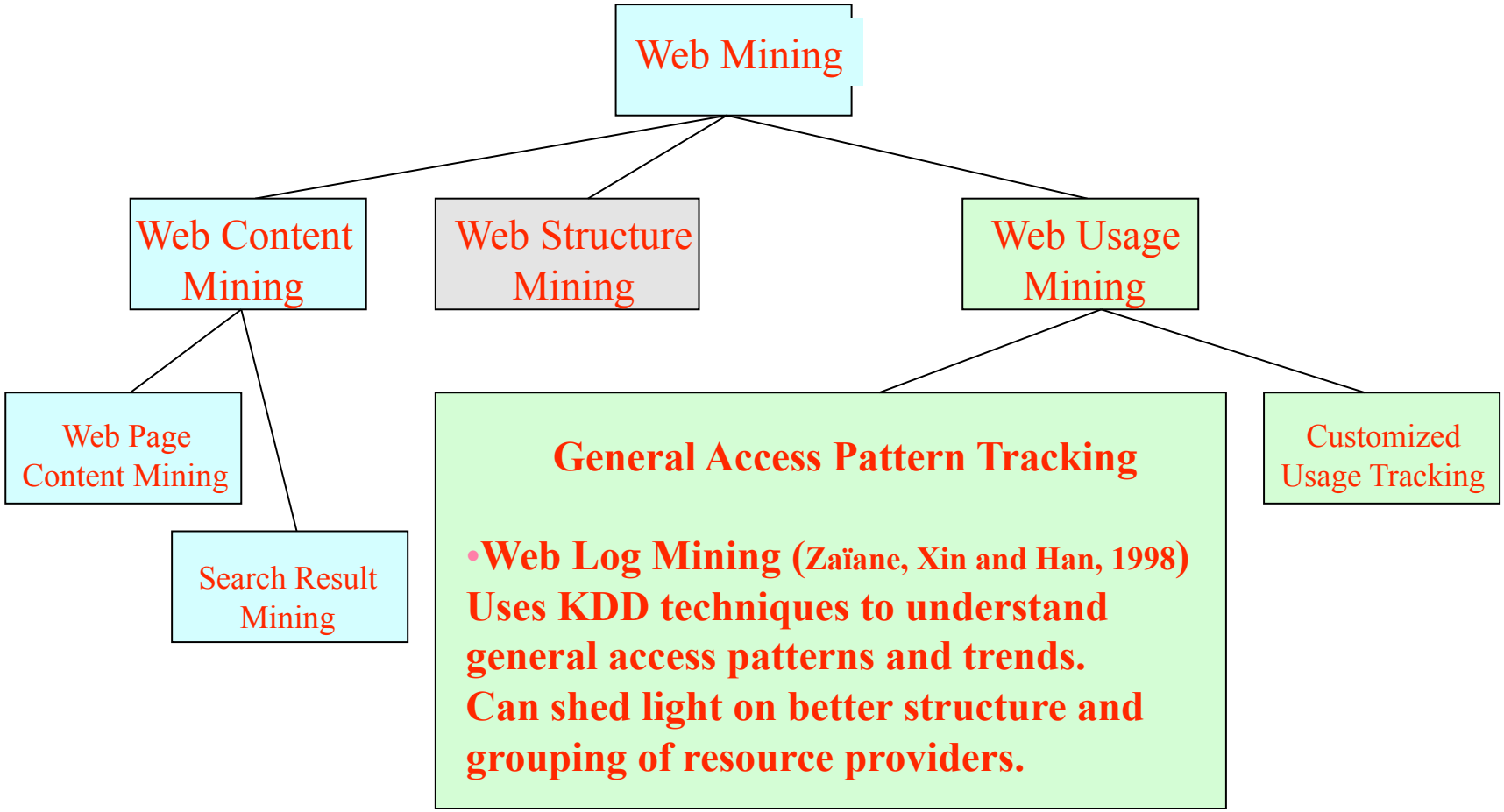
Mining the World Wide Web [Madria99]



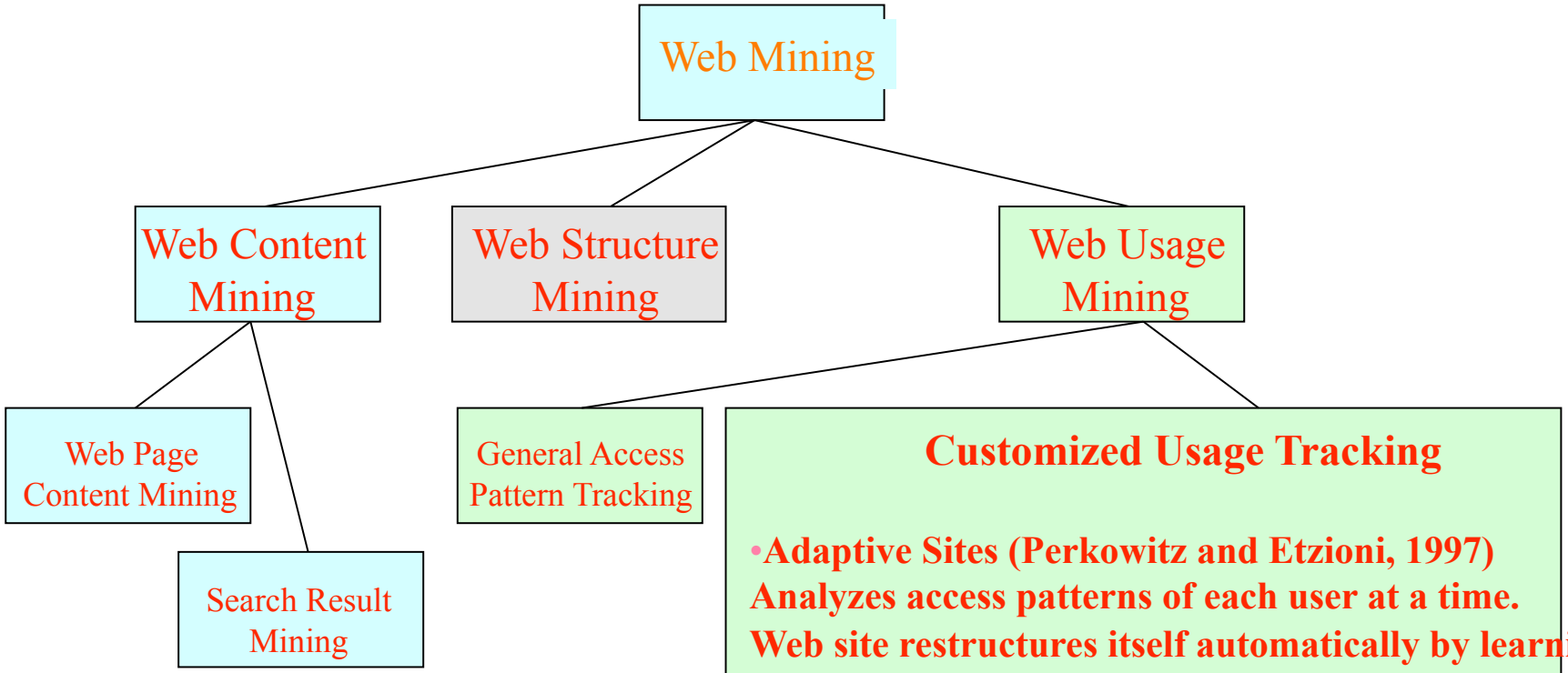
Mining the World Wide Web [Madria99]



Mining the World Wide Web [Madria99]

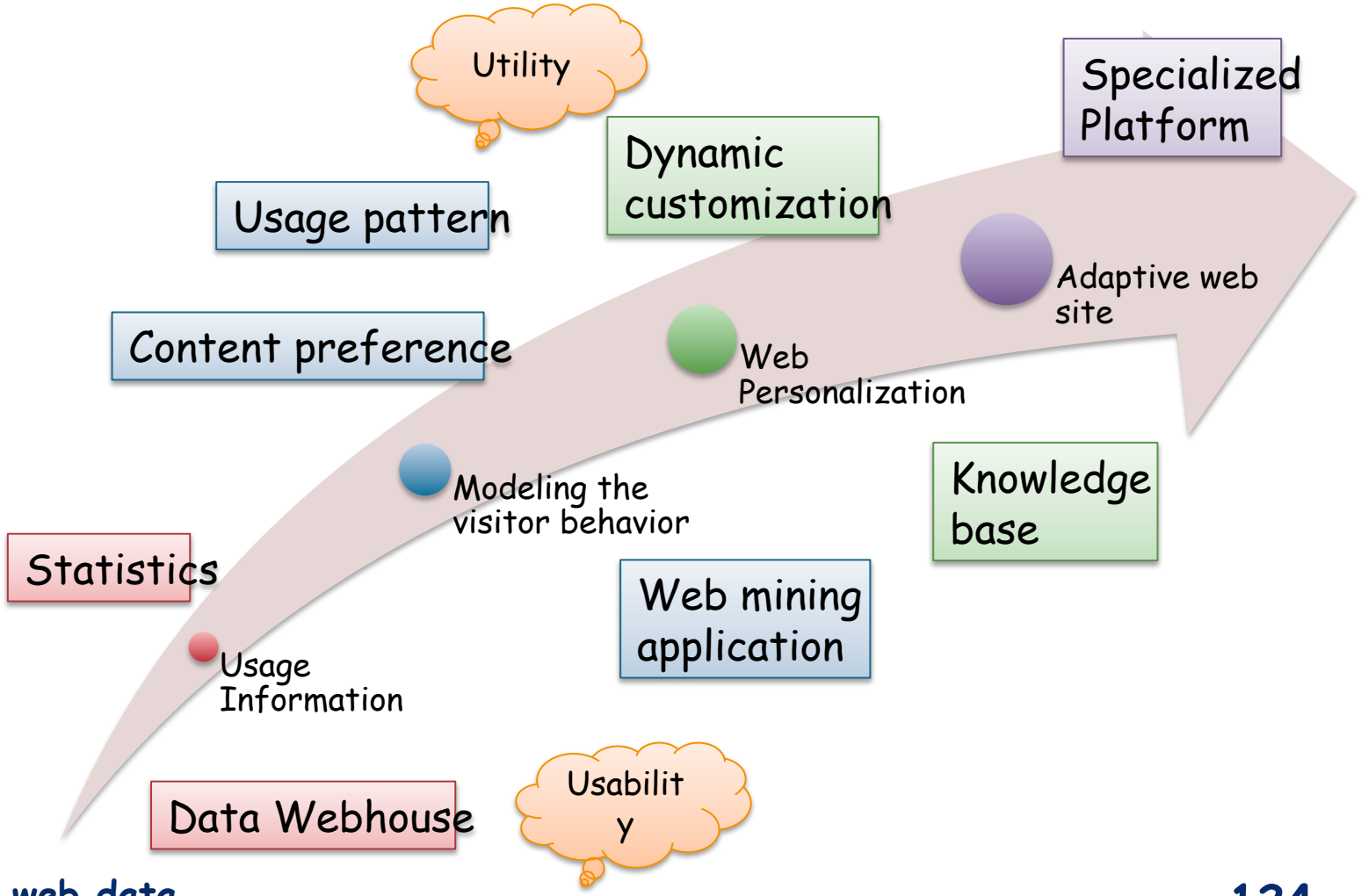


Mining the World Wide Web [Madria99]



5.- Applications

Practical applications



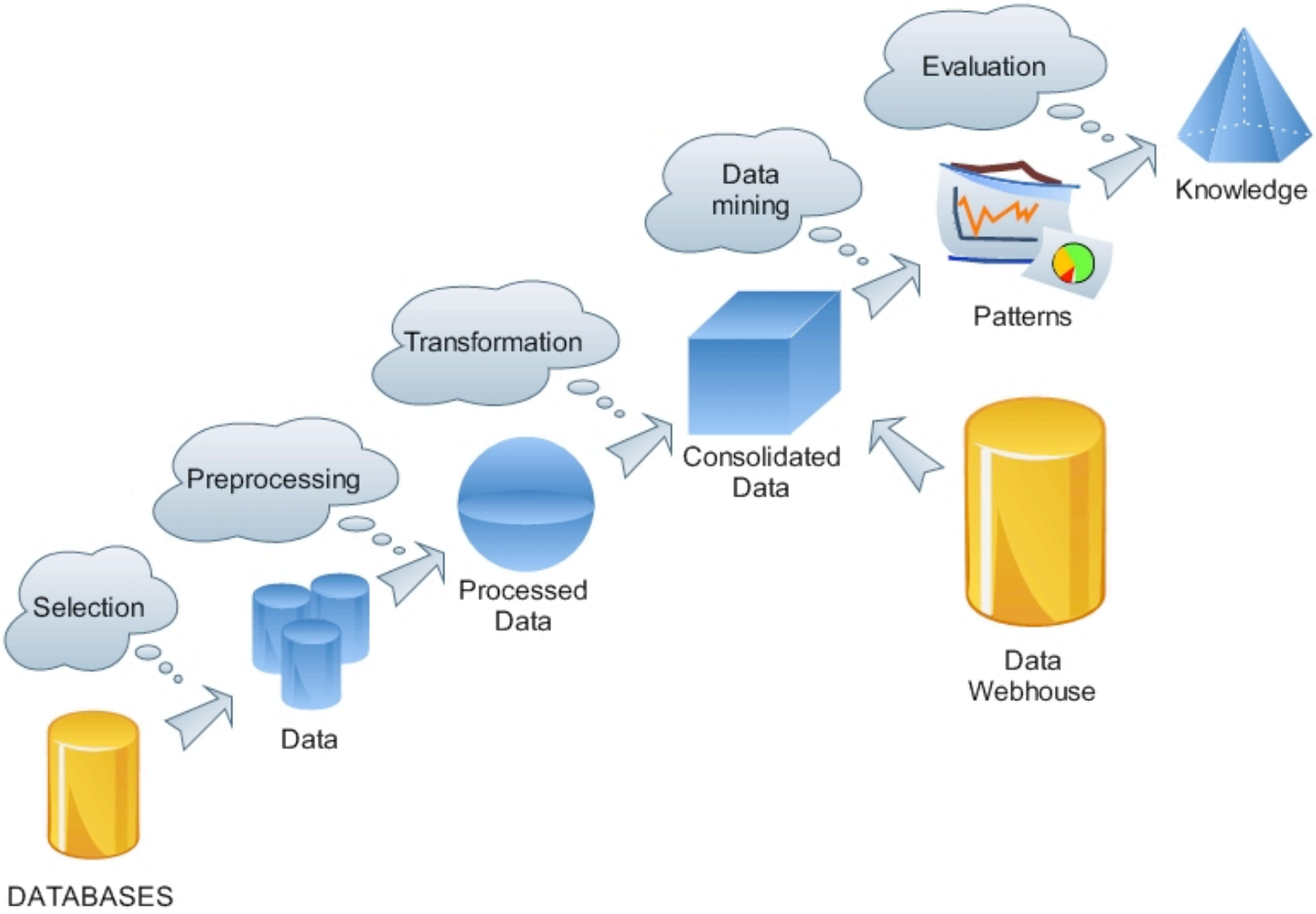
What should we do? [Mombasher01]

1. Data and information about the visitor's behavior in a Web site.
2. Modeling the visitor's behavior.
3. A model for the browsing behavior of the visitor must consider the pages visited, the time spent on each page and the page sequencing.
4. A model for the preferences of the visitor must consider the content of the pages and the time spent on each page in a session.

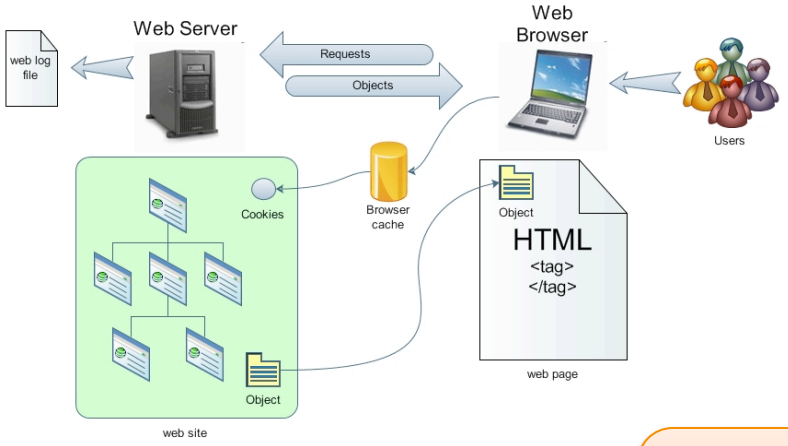
Applications

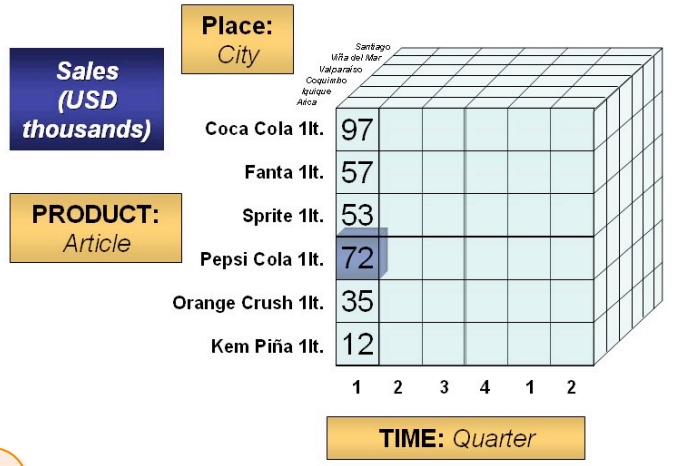
A DATA WEBHOUSE

The KDD process



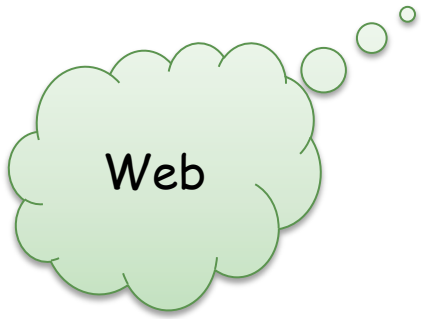
The Web Warehouse

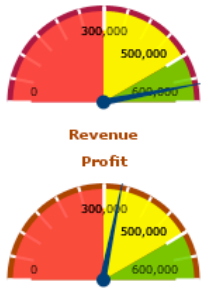


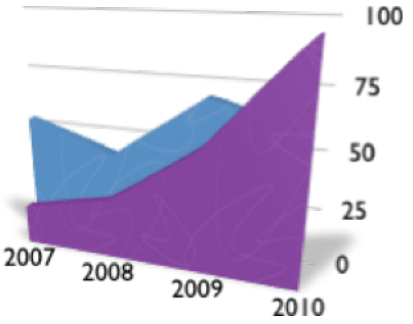


Web Warehouse

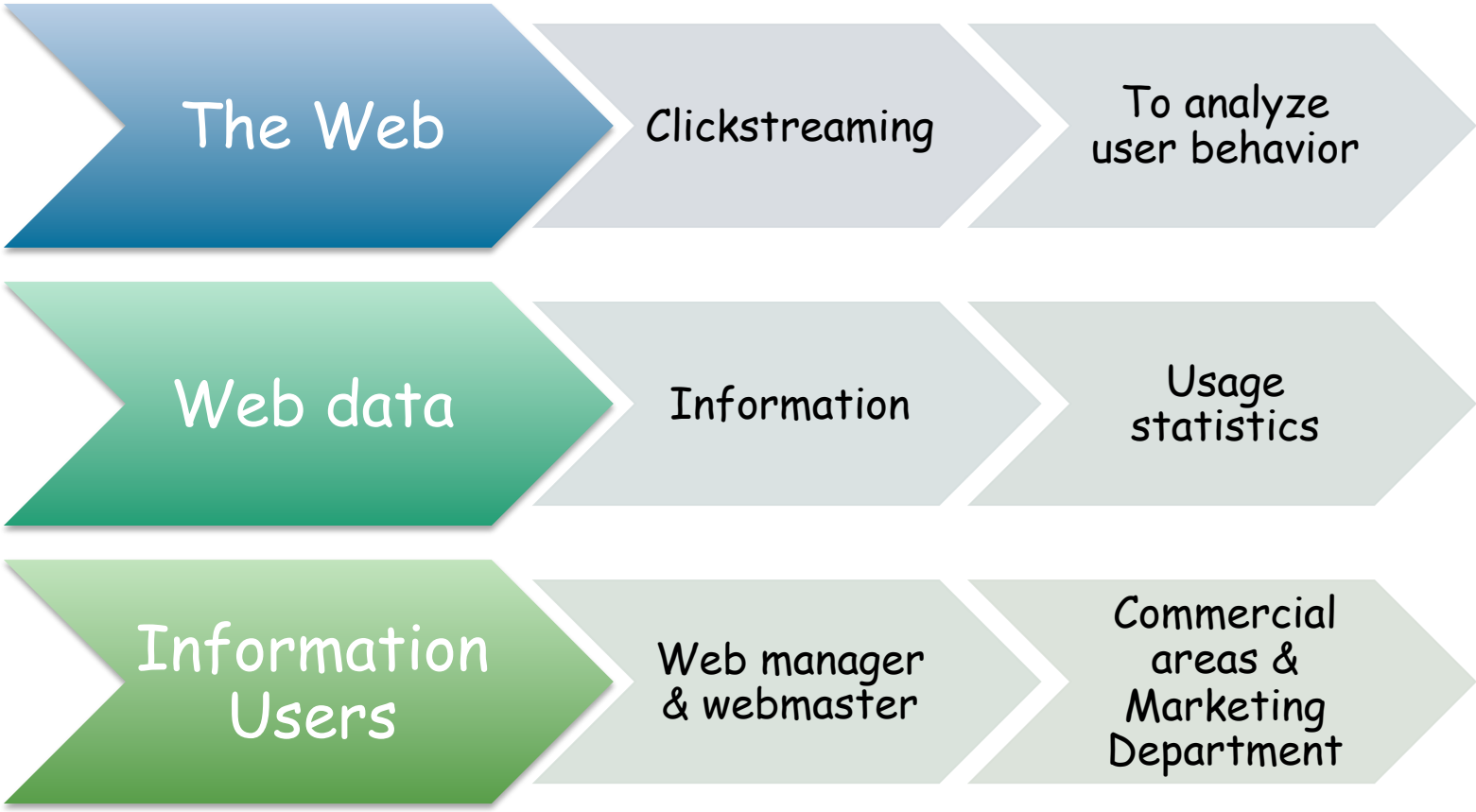
Introduced By Ralph Kimball





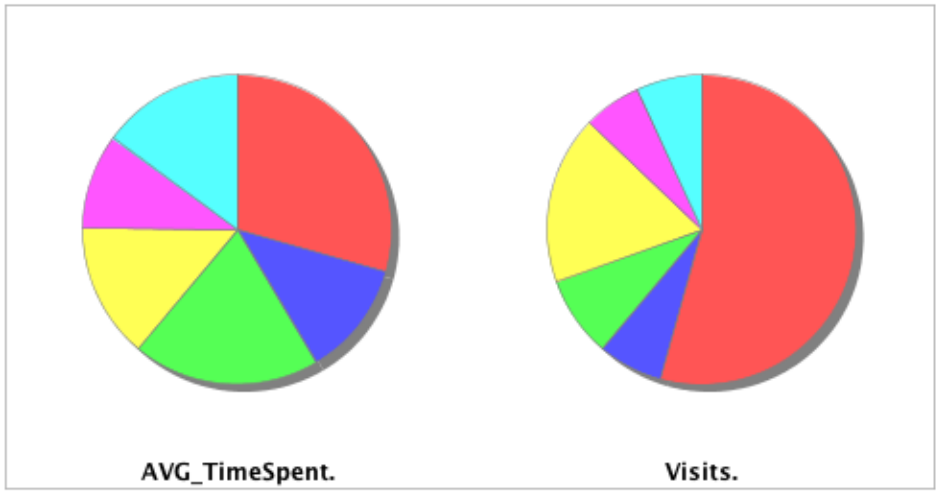


Web Warehousing



Technological Architecture

About web users



Slicer:

Chile. Germany. Great Britain. Japan. Korea. United States.

	Measures	
Location	AVG_TimeSpent	Visits
Chile	93,188	473
Germany	38,407	59
Great Britain	62	73
Japan	44,701	154
Korea	31,509	53
United States	47,407	59

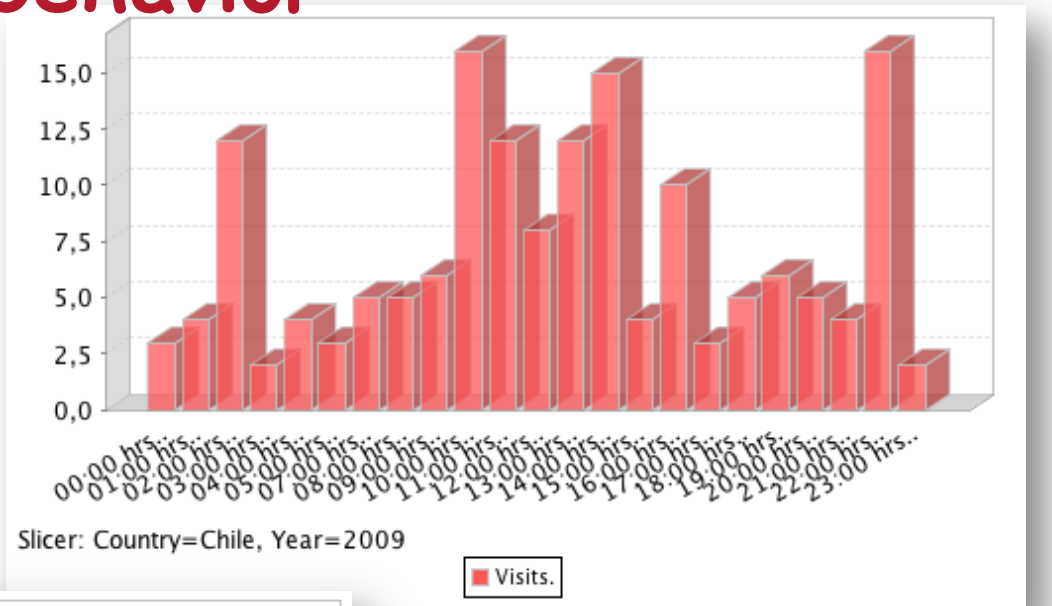
	Measures	
Browser	AVG_TimeSpent	Visits
-All Browsers	67,54	981
-Linux	66,252	143
Firefox 3.0	62,424	66
Firefox 3.5	69,532	77
-Mac OSX	62,283	106
Firefox 3.5	64,425	40
Safari 531-9	60,985	66
-Windows	68,553	732
Firefox 3.0	70,316	79
Firefox 3.5	74,619	105
Google Chrome 2.0	78,024	85
Internet Explorer 6.0	64,824	108
Internet Explorer 7.0	67,797	177
Internet Explorer 8.0	59,479	140
Opera 9.64	74,5	38

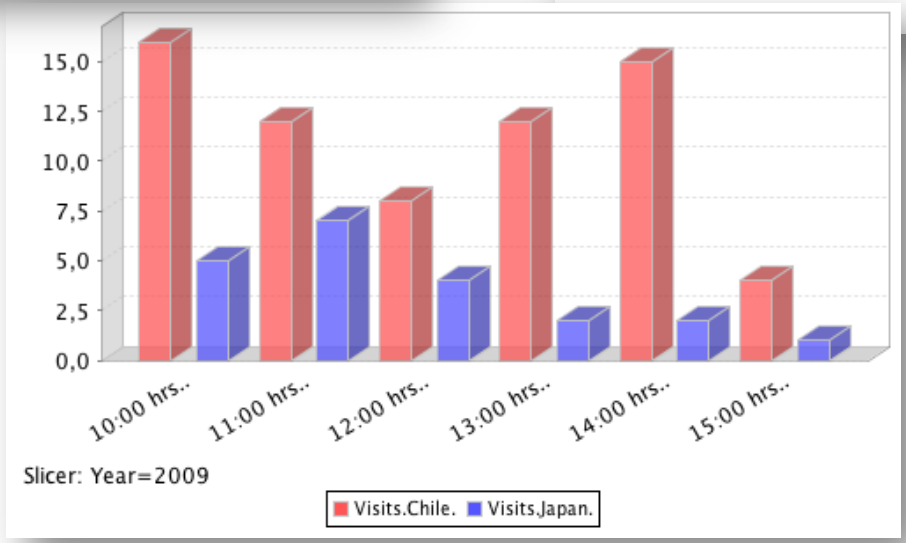
Machine		Measures	
Screen	Resolution	Visits	AVG_TimeSpent
Dual	1280x1024	61	70,885
Normal	1024x768	218	72,339
	1280x1024	197	64,655
	800x600	366	65,904
Wide	1440x900	139	66,942

About web user behavior

Timetable	Measures
Hour	AVG_TimeSpent
09:00 hrs.	67,333
10:00 hrs.	69,467
11:00 hrs.	54,321
12:00 hrs.	55,714
13:00 hrs.	67,3
14:00 hrs.	71,233
15:00 hrs.	68,636
16:00 hrs.	66,476
17:00 hrs.	54,5

Slicer: [(All)=All Locations] [Year=2009]





[http://www.dii.uchile.cl]