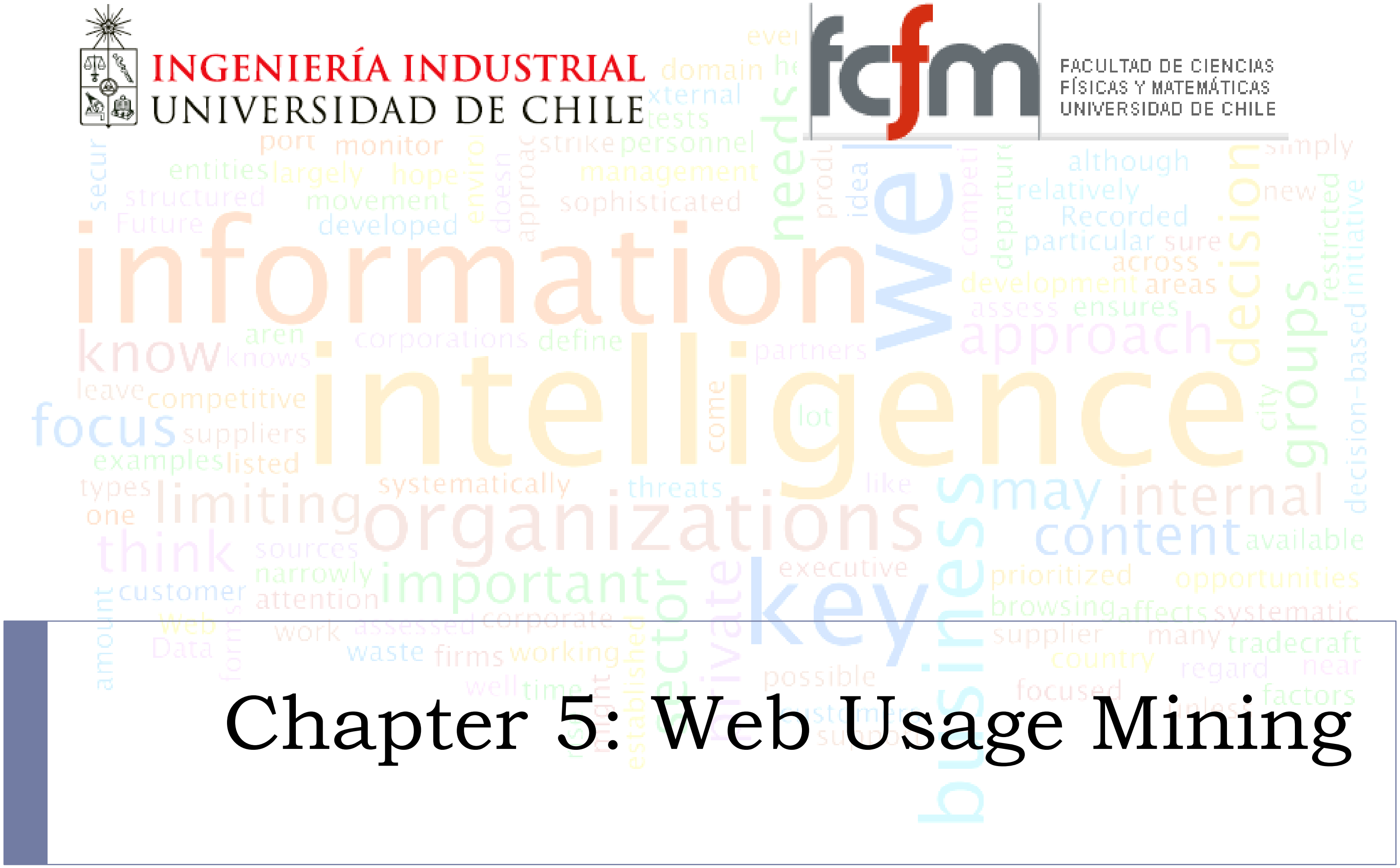
INGENIERÍA INDUSTRIAL
UNIVERSIDAD DE CHILE

fcfm
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

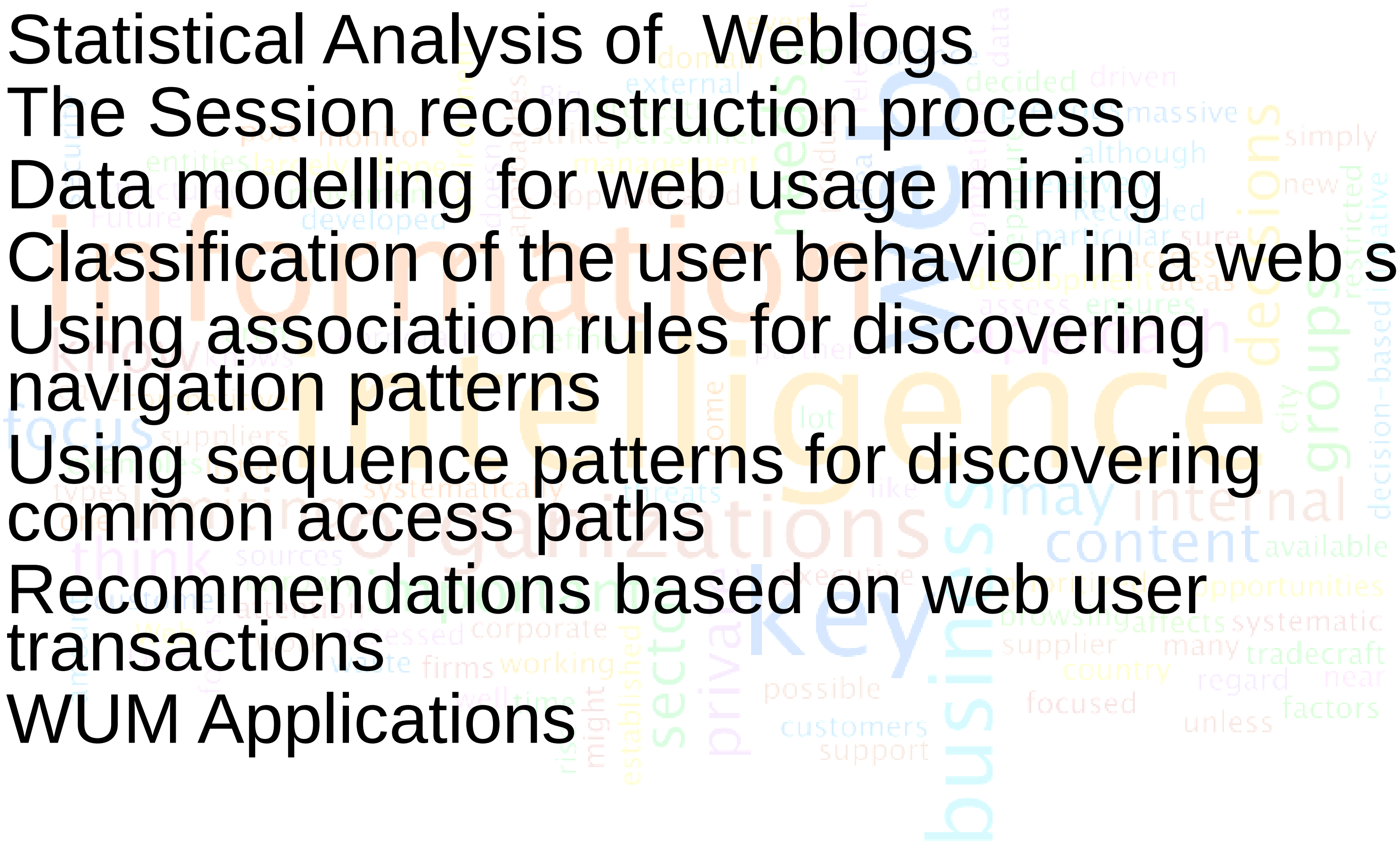


Chapter 5: Web Usage Mining

Professors:	Juan D. Velásquez	http://wi.dii.uchile.cl/
	Gaspar Pizarro V.	@juandvelasquez

Outline

1. Introduction
2. Statistical Analysis of Weblogs
3. The Session reconstruction process
4. Data modelling for web usage mining
5. Classification of the user behavior in a web site
6. Using association rules for discovering navigation patterns
7. Using sequence patterns for discovering common access paths
8. Recommendations based on web user transactions
9. WUM Applications



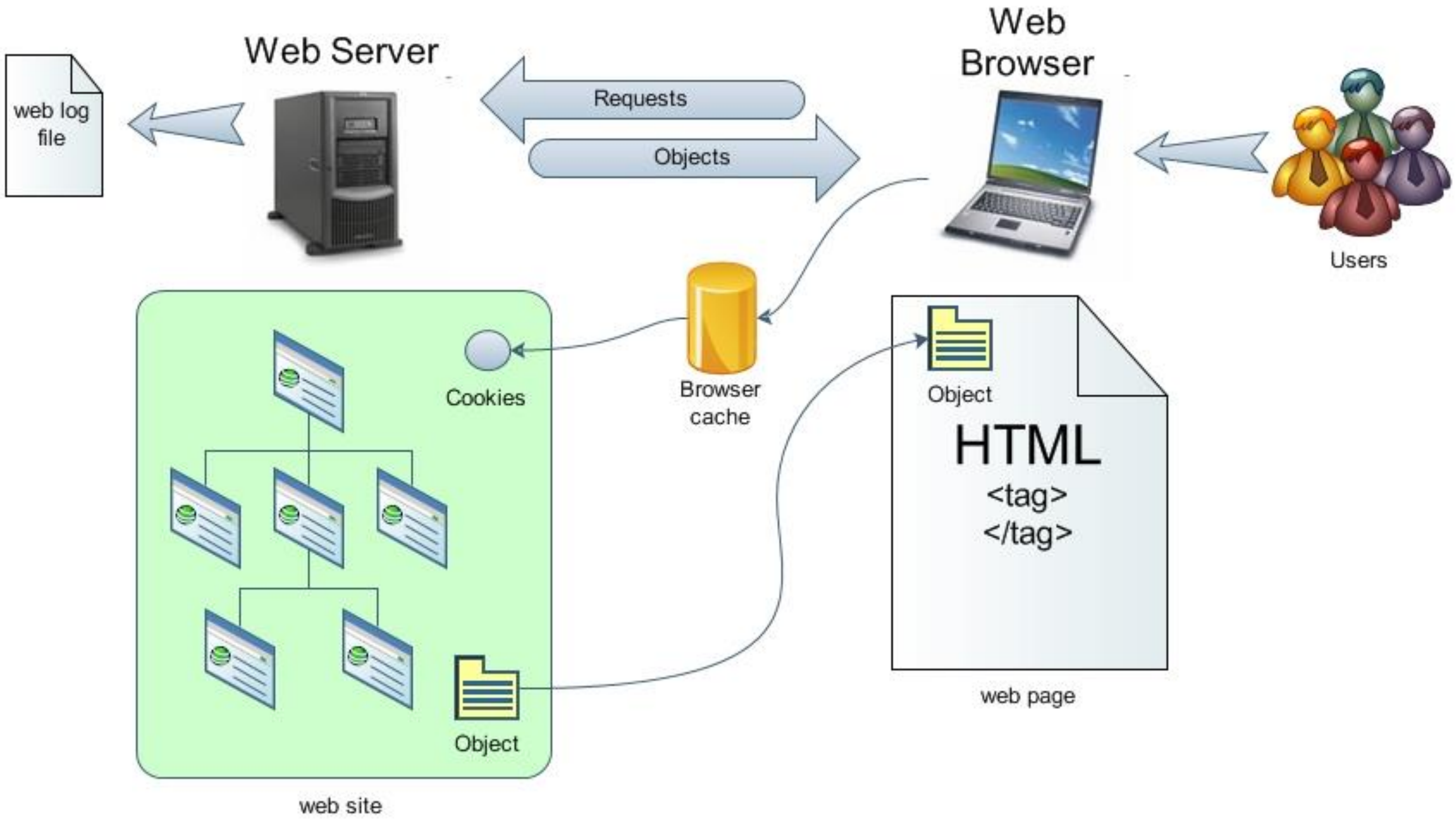
Section 5.1

Introduction

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.**

The Web process



Web server and web browser

- 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 (2)

- The transport operation use the **HTTP protocol.**
- Web pages are written in the **HTML language.**
- A web page contains tags that reference other object to ask to the server or to be download to the user browser.
- **Content** are usually more complex than they appears:
  - Applets
  - Javascripts
  - Dynamics HTML
  - Flash

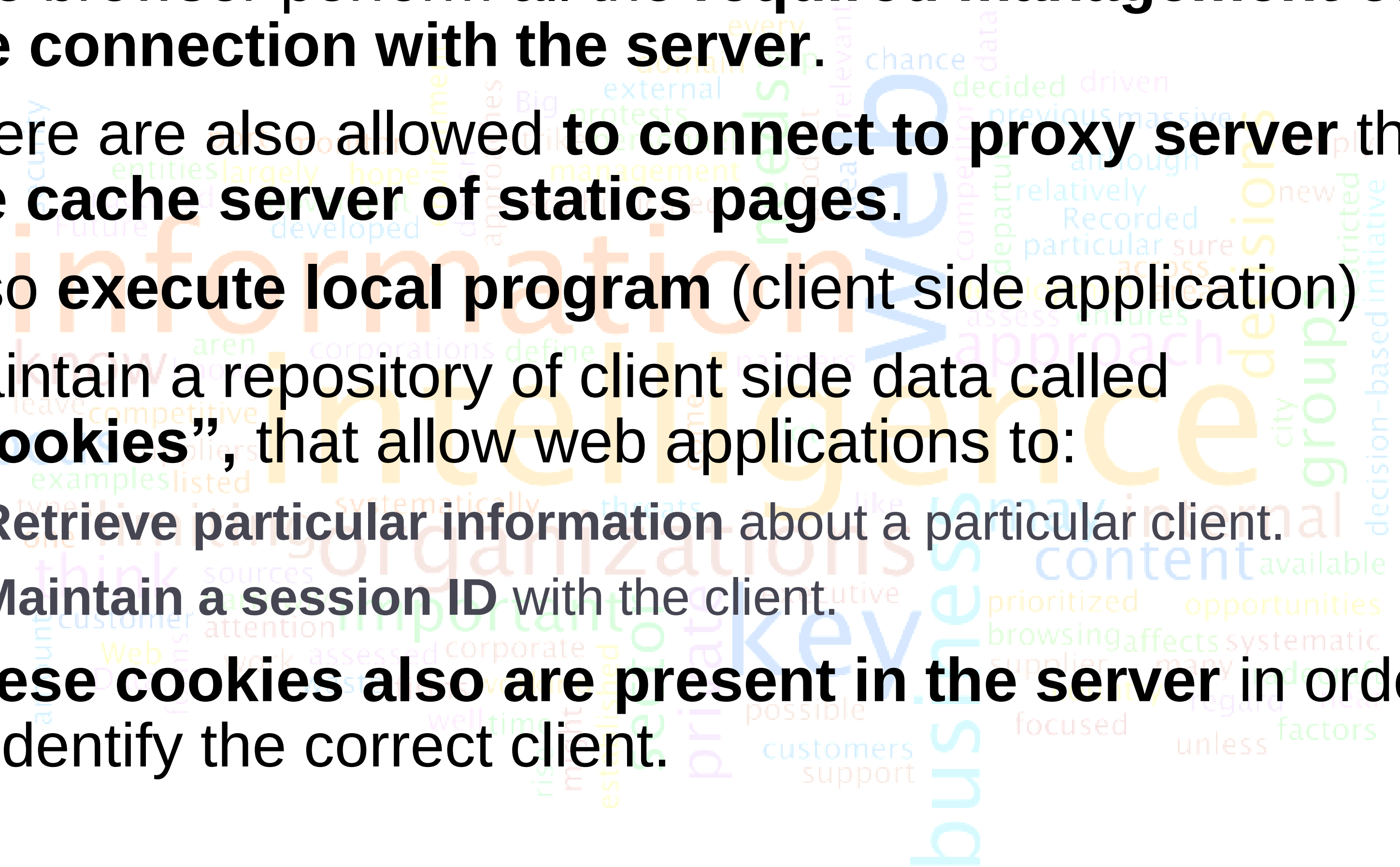
Web server and web browser (3)

- The **web log** registers contain information about the **visitor browsing behaviour**, 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**.

The web Server

- A Server is a program not a machine.
- They usually serve **not only plain web pages** , they also serve **web applications with HTML front-end**.
- Web application architecture: **Multiple layer Model**.
  - **Interface Layer**: Perform html rendering.
  - **Logic Layer**: core business application.
  - **Data Layer**: Storage/Retrieval data process.
- Web server also maintain data logs about **user action** in the web: Some site could have of the order of **Gb/day of logs**.

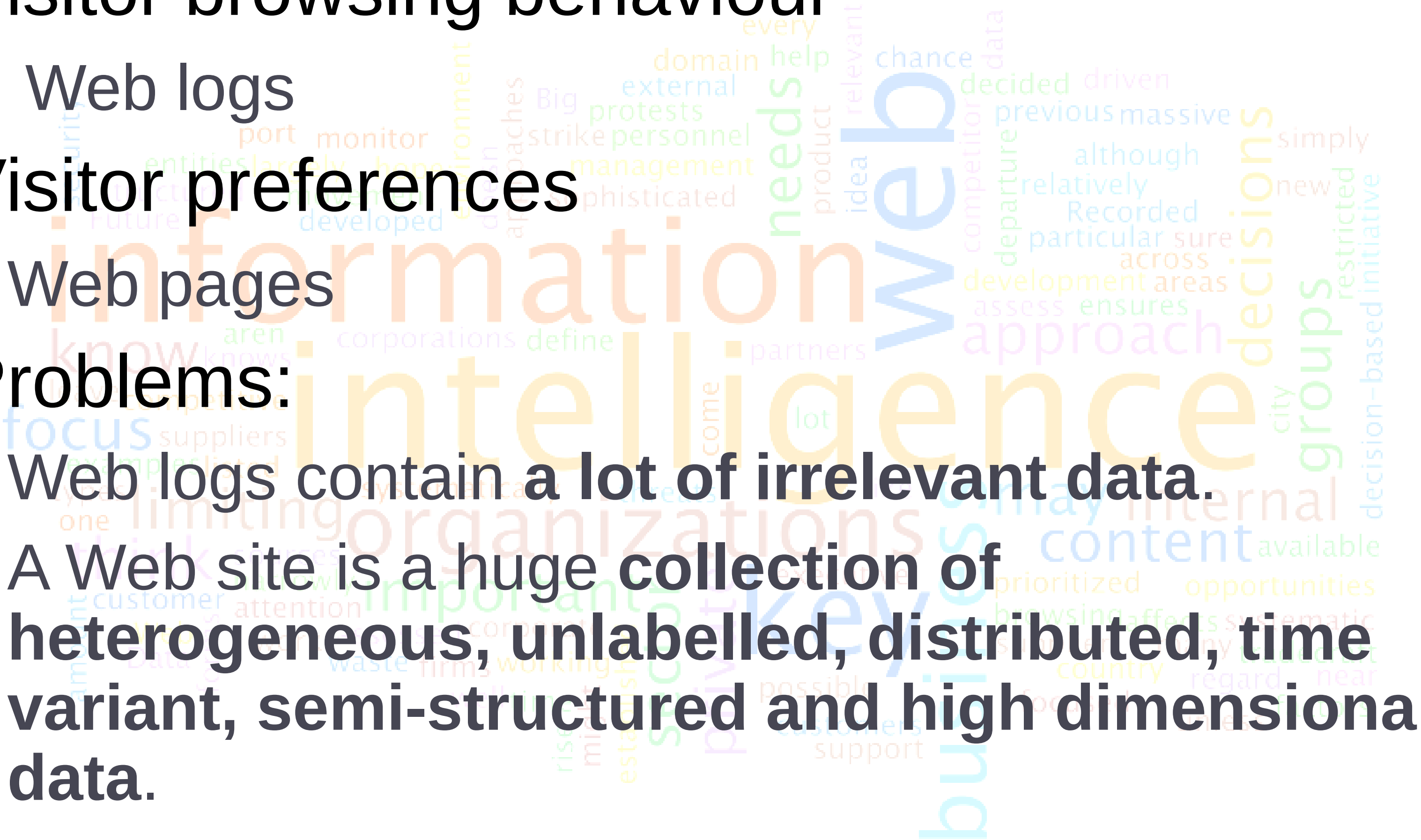
The web Client



- The browser perform all the **required management of the connection with the server**.
- There are also allowed **to connect to proxy server** that are **cache server of statics pages**.
- Also **execute local program** (client side application)
- Maintain a repository of client side data called **“Cookies”**, that allow web applications to:
  - **Retrieve particular information** about a particular client.
  - **Maintain a session ID** with the client.
- **These cookies also are present in the server** in order to identify the correct client.

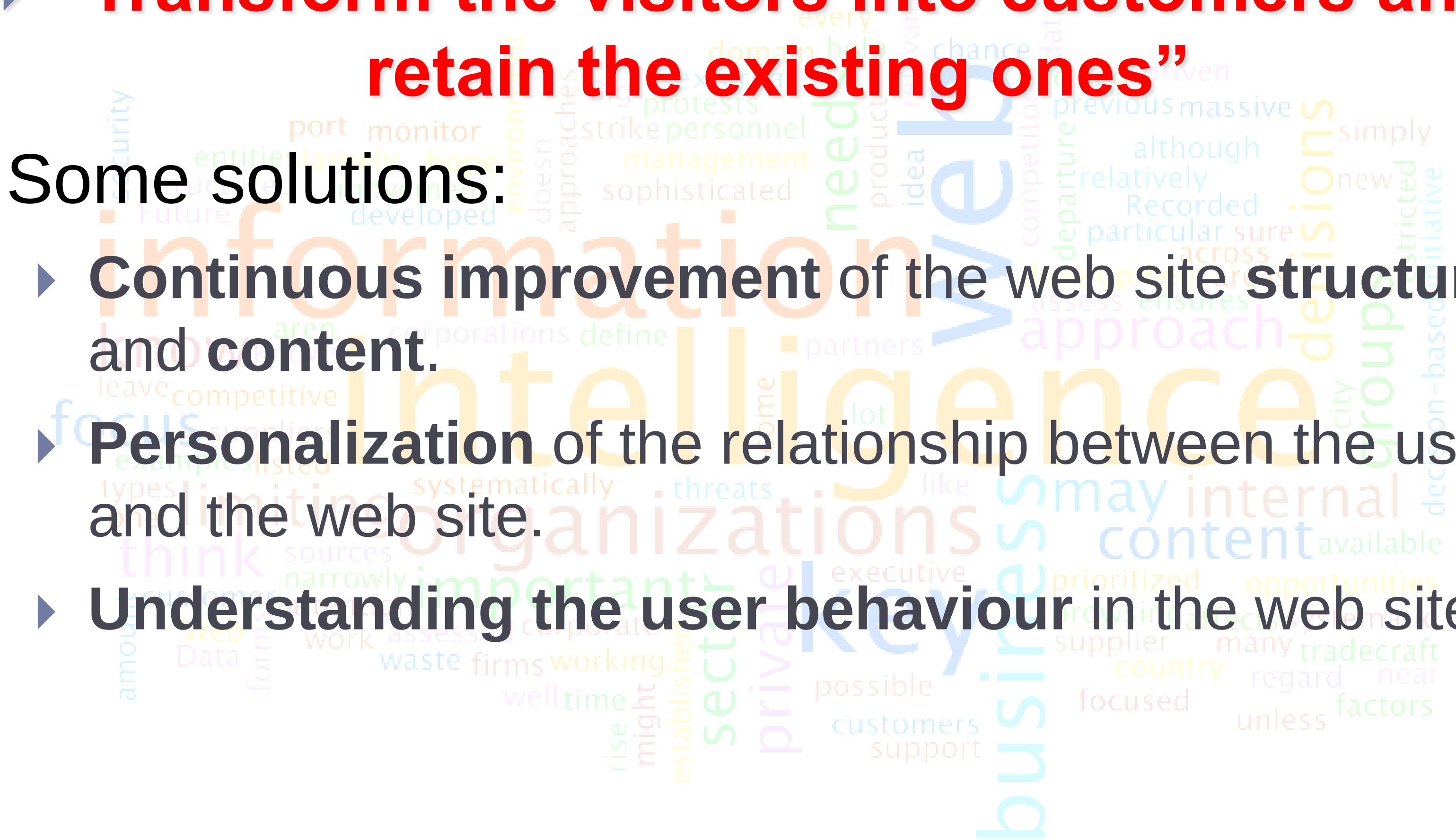
The visitor behavior in a Web site

- Visitor browsing behaviour
  - 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

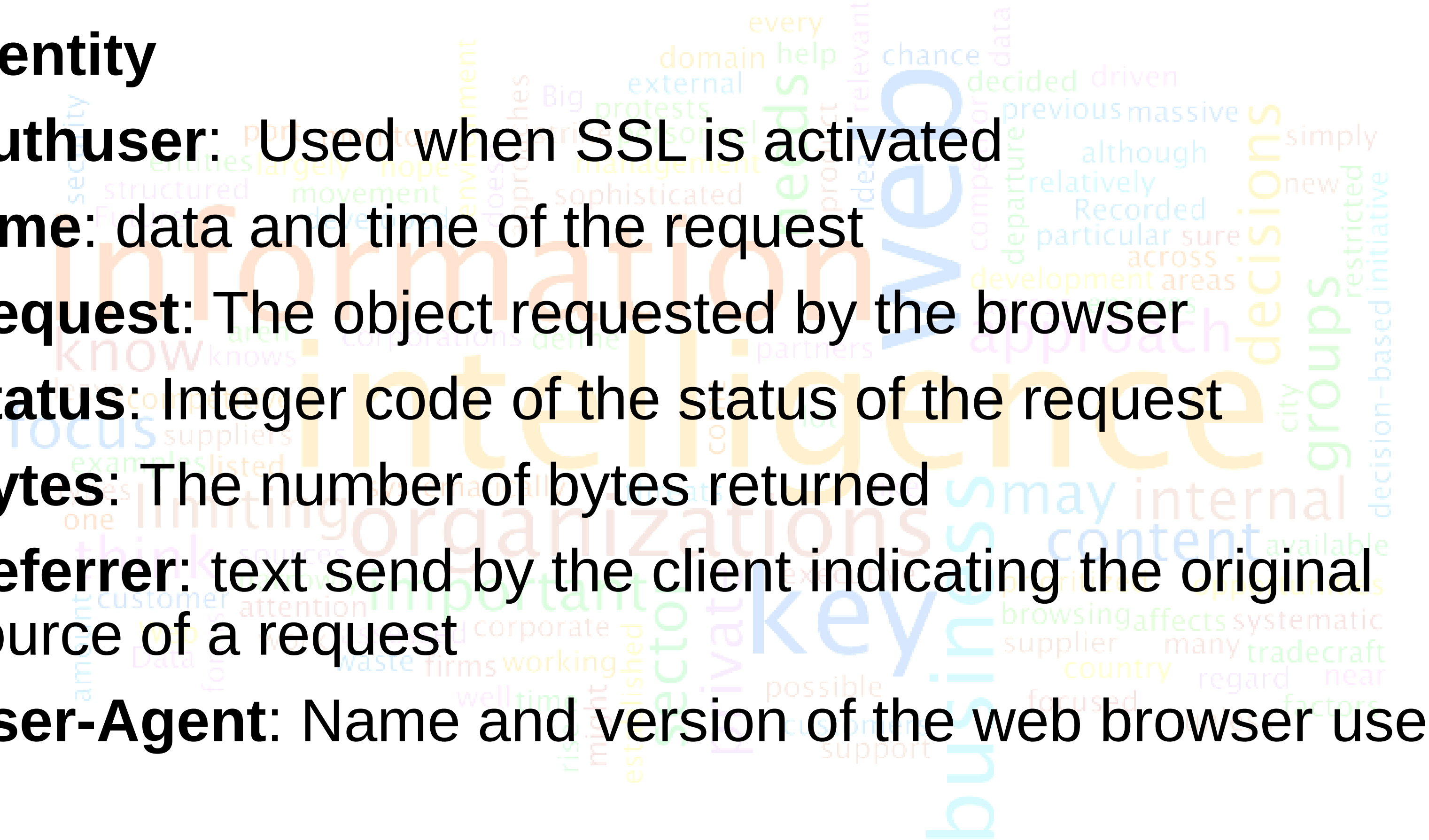
- **“Transform the visitors into customers and retain the existing ones”**
- Some solutions:
  - **Continuous improvement** of the web site **structure** and **content**.
  - **Personalization** of the relationship between the user and the web site.
  - **Understanding the user behaviour** in the web site.



The Server log file

- Usually this repository was used to perform **web server tuning** and other **system administration task**.
- But they acquire some **unexpected VALUE** to the **marketing researcher**.
- **THEY CONTAIN IN A IMPLICIT WAY ALL THE CLIENT BEHAVIOUR:**
  - **WE DON'T NEED A SURVEY, WE ALREADY HAVE THE INFORMATION**
- The Log file write **a line** with this precious information **for each request of a client browser.**

The Server log File: Structure



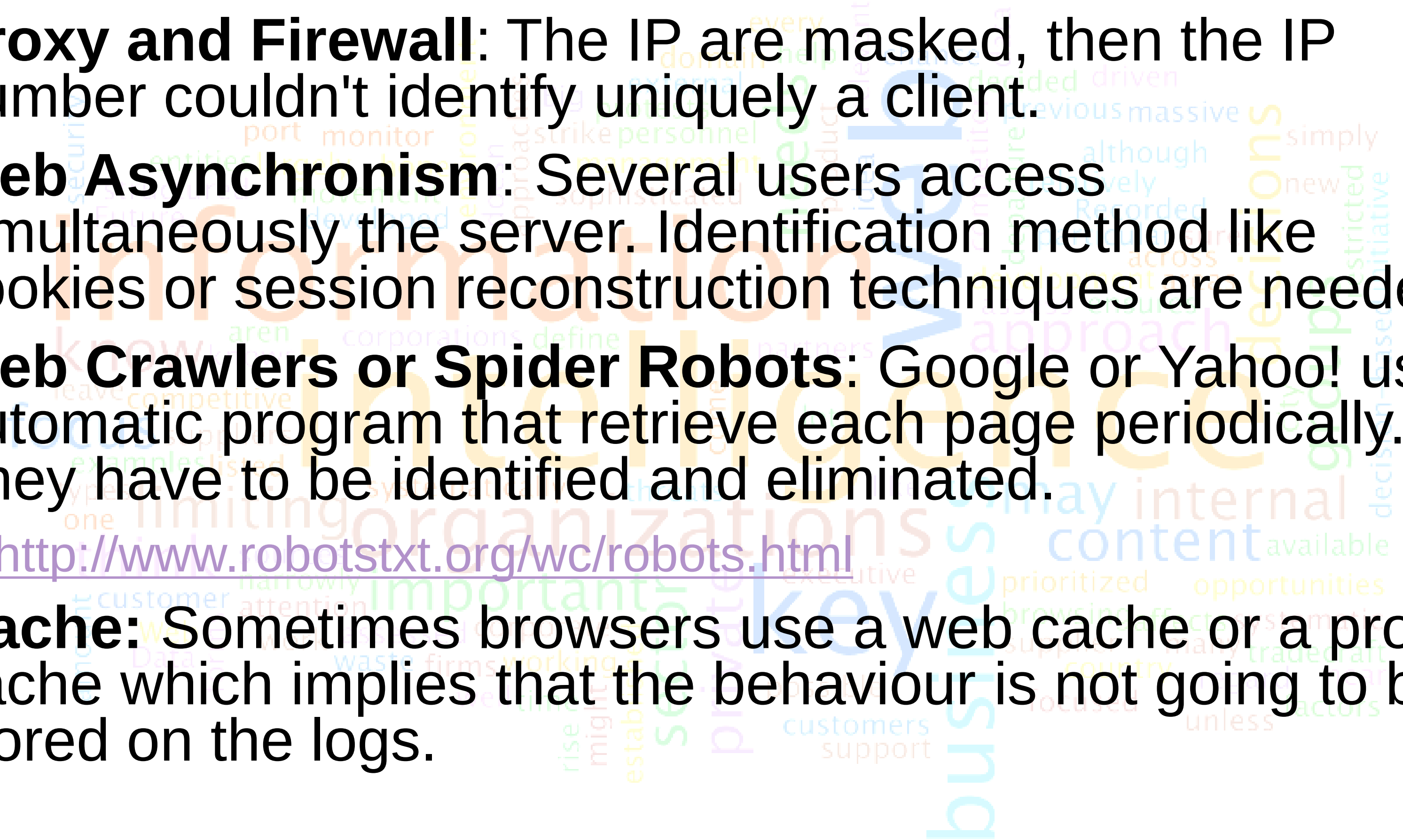
- **IP Address**: Client IP
- **Identity**
- **Authuser**: Used when SSL is activated
- **Time**: data and time of the request
- **Request**: The object requested by the browser
- **Status**: Integer code of the status of the request
- **Bytes**: The number of bytes returned
- **Referrer**: text send by the client indicating the original source of a request
- **User-Agent**: Name and version of the web browser used

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)

The Web Logs: some problems

- **The web log does not store the client id.**
- **Proxy and Firewall**: The IP are masked, then the IP number couldn't identify uniquely a client.
- **Web Asynchronism**: Several users access simultaneously the server. Identification method like cookies or session reconstruction techniques are needed.
- **Web Crawlers or Spider Robots**: Google or Yahoo! use a automatic program that retrieve each page periodically. They have to be identified and eliminated.
  - http://www.robotstxt.org/wc/robots.html
- **Cache:** Sometimes browsers use a web cache or a proxy cache which implies that the behaviour is not going to be stored on the logs.



Section 5.2

Statistical analysis of weblogs

Statistical Methods (1)

- Several tools use **conventional statistical methods** to *analyze user behavior in a web site.*
  - http://www.accrue.com
  - http://www.netgenesis.com
  - http://www.webtrends.com
  - http://www.google.com/analytics/

Statistical Methods (2)

- Each one of them use **graphics interfaces**, like *histograms*, with associated statistics.
- For instance, the amount of clicks per page during the last month can be obtained.
- By applying conventional statistics on web logs, one can perform different kinds of analysis [Boving04] .

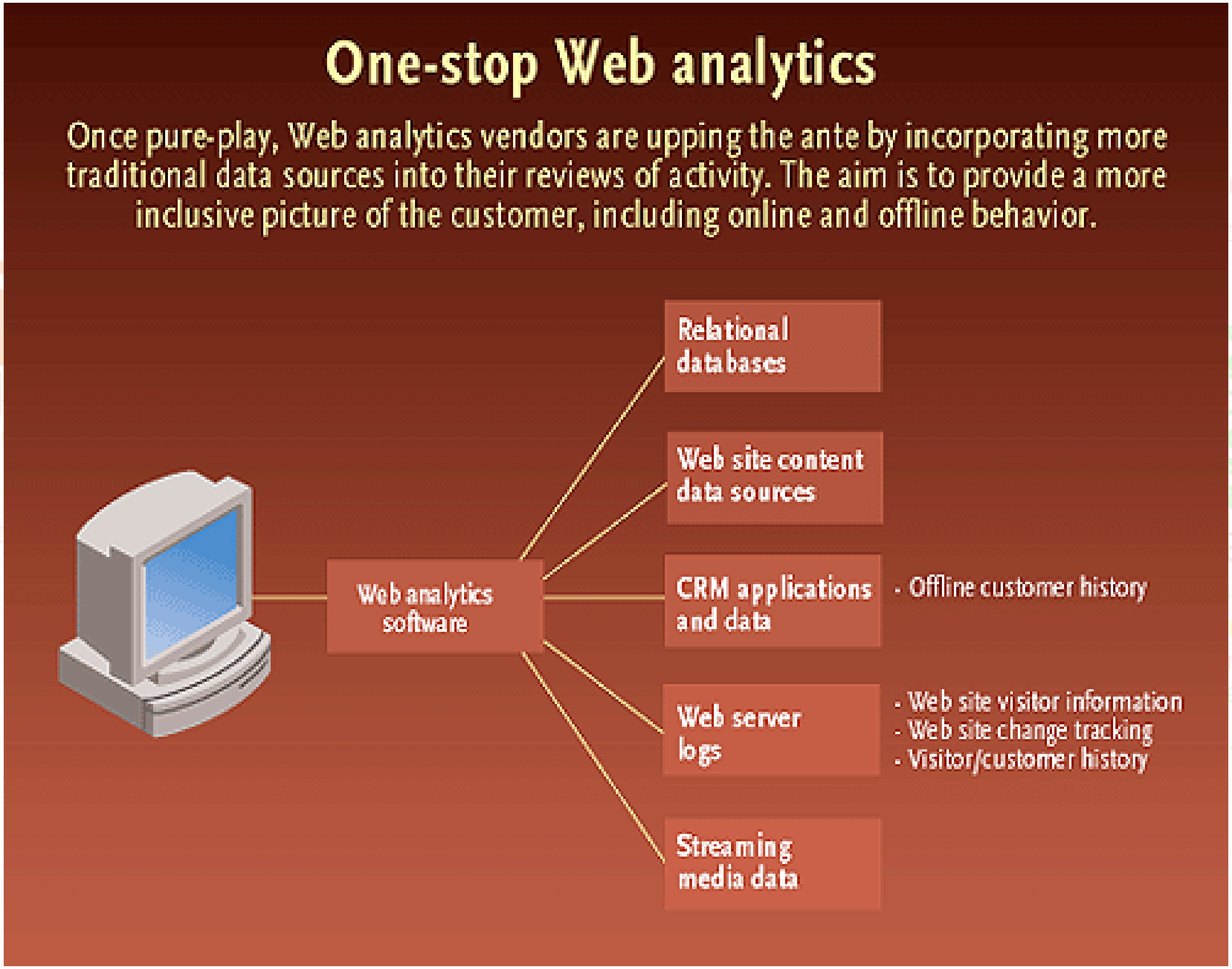
Statistical Methods (3)

- 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 is easy to understand the tools' reports.
- Although the statistical analysis could seem a very simple tool to mine web logs, it solves a lot of problems regarding the **performance of a web site**.

The business: Gain a Competitive Advantage through Web Site Analysis

- Acquire New Customers
- Retain Existing Customers
- Improve Customer Satisfaction
- Sell More Products
- Improve Marketing Initiatives
- Find out Niche Opportunities
- Make Data-Driven Decisions
- Differentiate Web Site from Competitors
- Develop and Enhance Marketing Strategies

One-Stop Web Analytics



One-stop Web analytics

Once pure-play, Web analytics vendors are upping the ante by incorporating more traditional data sources into their reviews of activity. The aim is to provide a more inclusive picture of the customer, including online and offline behavior.

Web analytics software
Relational databases
Web site content data sources
CRM applications and data
- Offline customer history
Web server logs
- Web site visitor information
- Web site change tracking
- Visitor/customer history
Streaming media data

Web Analytics Market

- Market is growing 200 % annually*
- $4 Billion market in 2004*
- Sales of Web Analytics software will surpass all other CRM purchases**

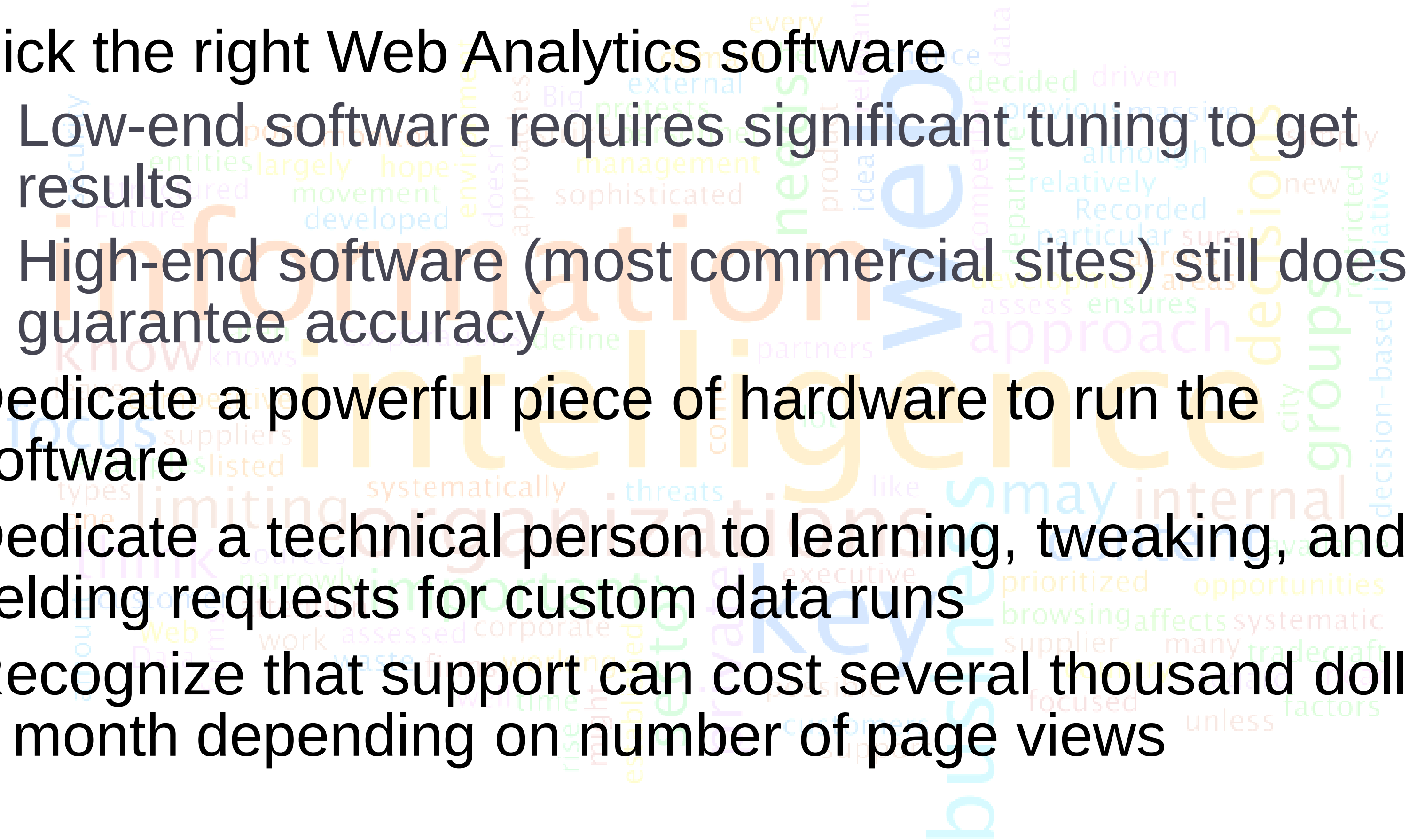
*According to The Aberdeen Group

**According to The Meta Group

Data Analysis

- **Clickstream data analysis** can yield information:
  - Time spent on a particular page
  - Which page receives the most hits
  - Where visitor is referred from (e.g., search engine)
  - “Conversion rate” of visitors (i.e., visitors who buy)
  - Customer buying trends and preferences
  - Determining effective web site design

Knowing how to pick the right Web Analytics Software



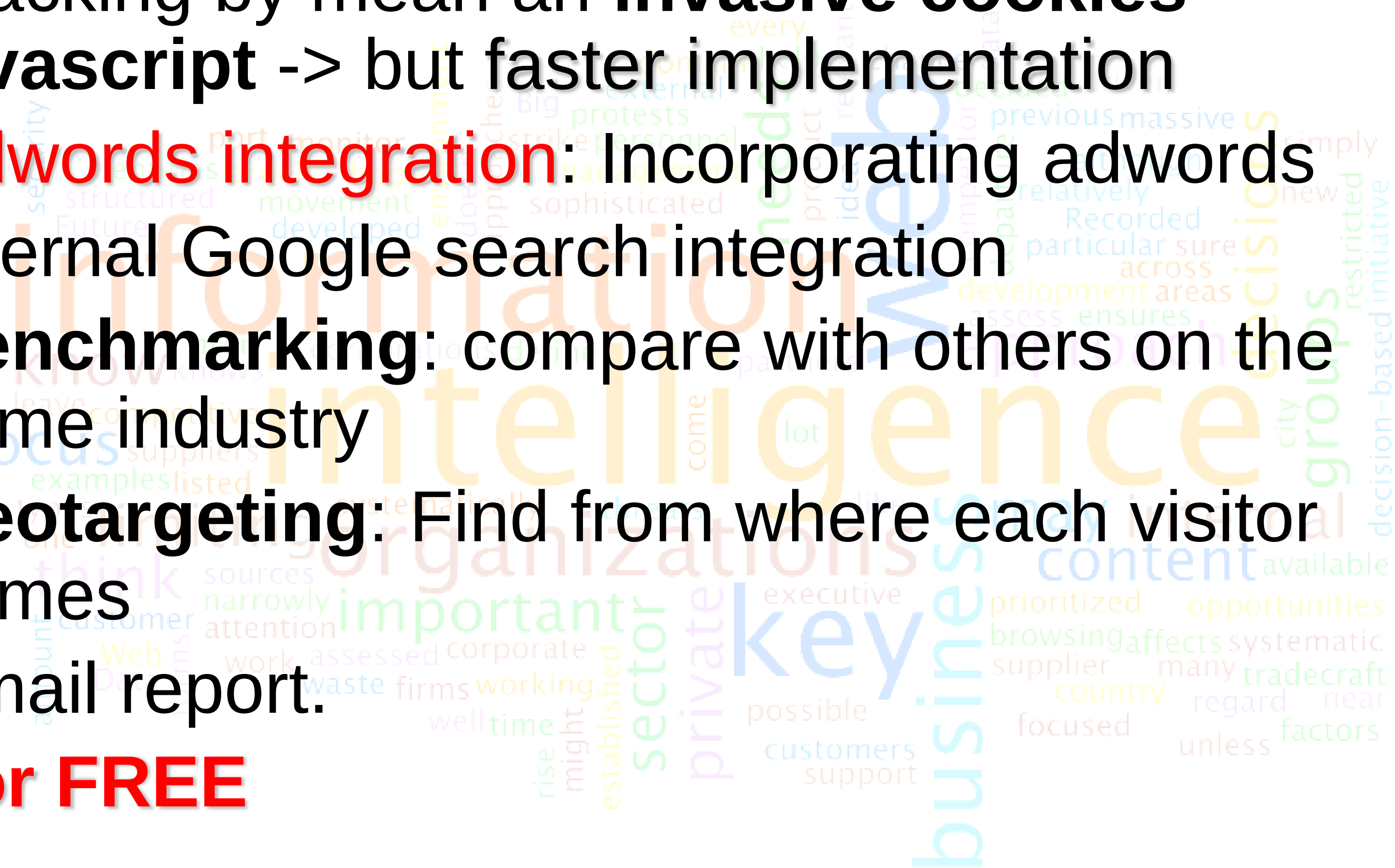
- Decide in advance what data to collect
- Pick the right Web Analytics software
  - Low-end software requires significant tuning to get results
  - High-end software (most commercial sites) still doesn’t guarantee accuracy
- Dedicate a powerful piece of hardware to run the software
- Dedicate a technical person to learning, tweaking, and fielding requests for custom data runs
- Recognize that support can cost several thousand dollars a month depending on number of page views

Low-end to High-end Web Analytics Software (Picking the right Web Analytics Software)

- **Accrue Software** - www.accrue.com
  - Price: $15,000 and up
  - Ideal for big operations
- **NetAcumen** - www.netacumen.com
  - Price: $5,000 a month and up for hosted solution, $150,000 and up for software
  - Ideal for B2B companies
- **NetGenesis** - www.netgenesis.com
  - Price: $160,000 and up
  - Ideal for big businesses (Charles Schwab, GE, etc.)

Google Analytics

- Tracking by mean an **invasive cookies-javascript** -> but faster implementation
- Adwords integration: Incorporating adwords
- Internal Google search integration
- **Benchmarking**: compare with others on the same industry
- **Geotargeting**: Find from where each visitor comes
- Email report.
- **For FREE**



Section 5.3

The session reconstruction process

Session reconstruction

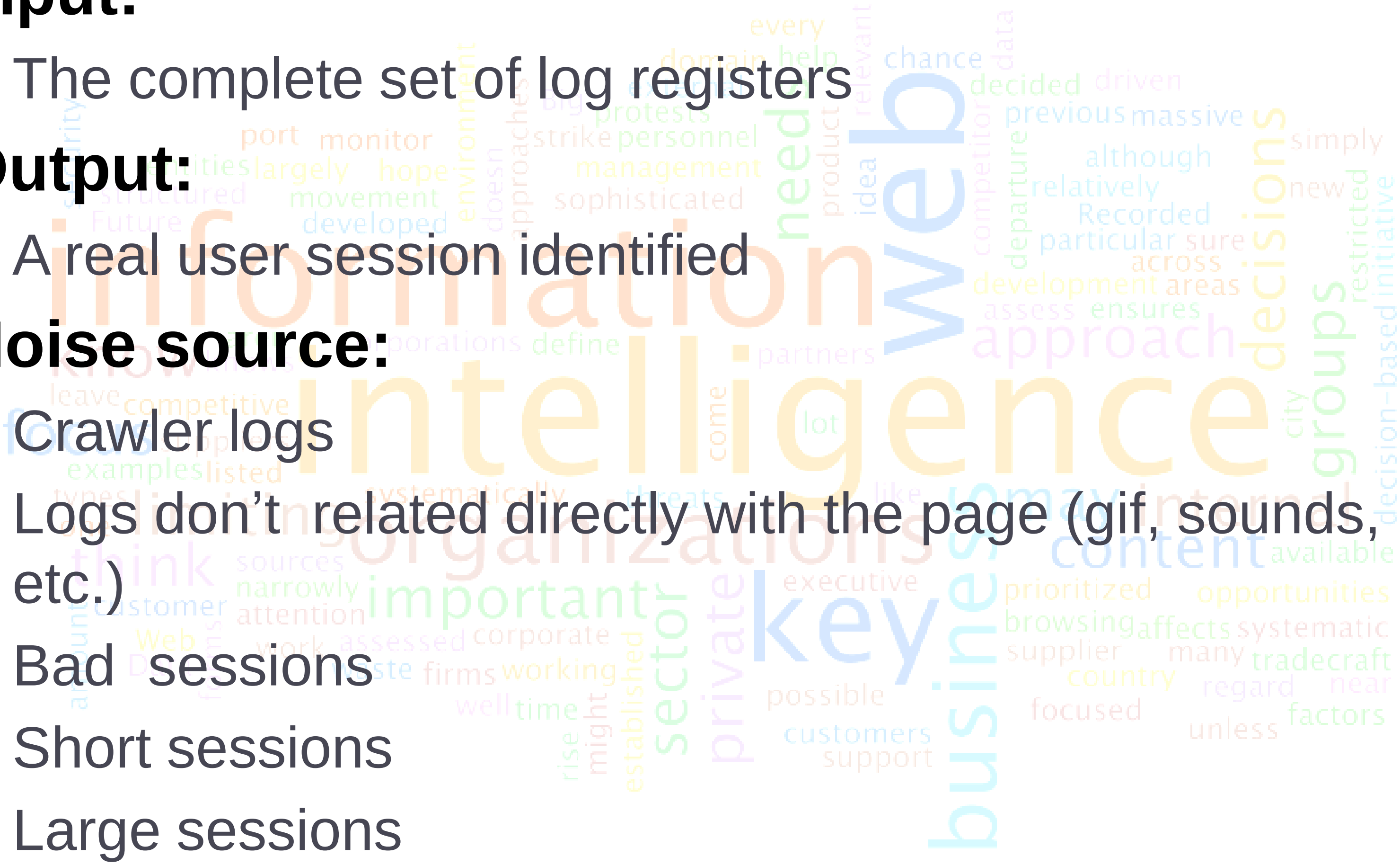
- If we want to understand the user behaviour in a web site, web need to know his/her **real browsing behavior**.
- The **quality of patterns** extracted by using a mining technique **depend on the input data**.
- Elements like **proxies servers**, **dynamic IP**, **missing references** and the **inability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

Session reconstruction process

- **We want to identify the lines in the logs file that belong to a unique valid client.**
- This process is called **“Sesionization”**
- **Usual assumption**:
  - **Each session has a maximum time duration.**
- Strategies:
  - **Proactive strategies**: Identify users with methods like cookies, usually facing privacy problems.
  - **Reactive strategies**: no privacy concerns.
    - **Navigational Oriented Heuristics**: pages visited follows the hyperlink structure. If a page doesn't follows this order is a new session.
    - **Time Oriented Heuristics**: Using usually **30 min for maximum session time**.

Session reconstruction: I/0

- **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

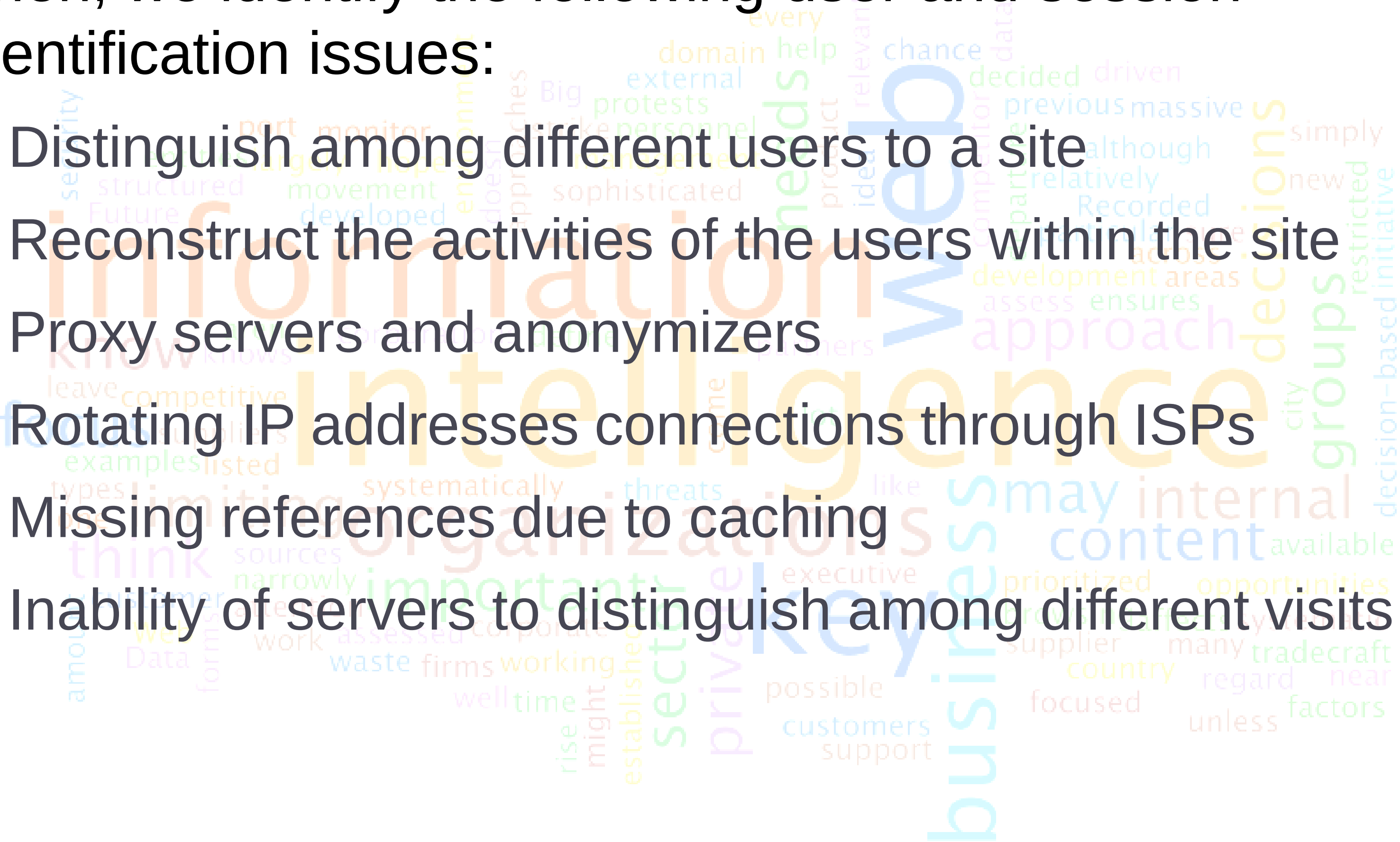


Some problems (1)

- **Single IP address/Multiple Server Sessions**.
  - Proxy server
- **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 behaviour each time.
- **Multiple Agent/Single User**. As before, when a visitor uses different machines that may have **different agents**.

Some problems (2)

- Then, we identify the following 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



Proposing some solutions…

Reactive Sessionization

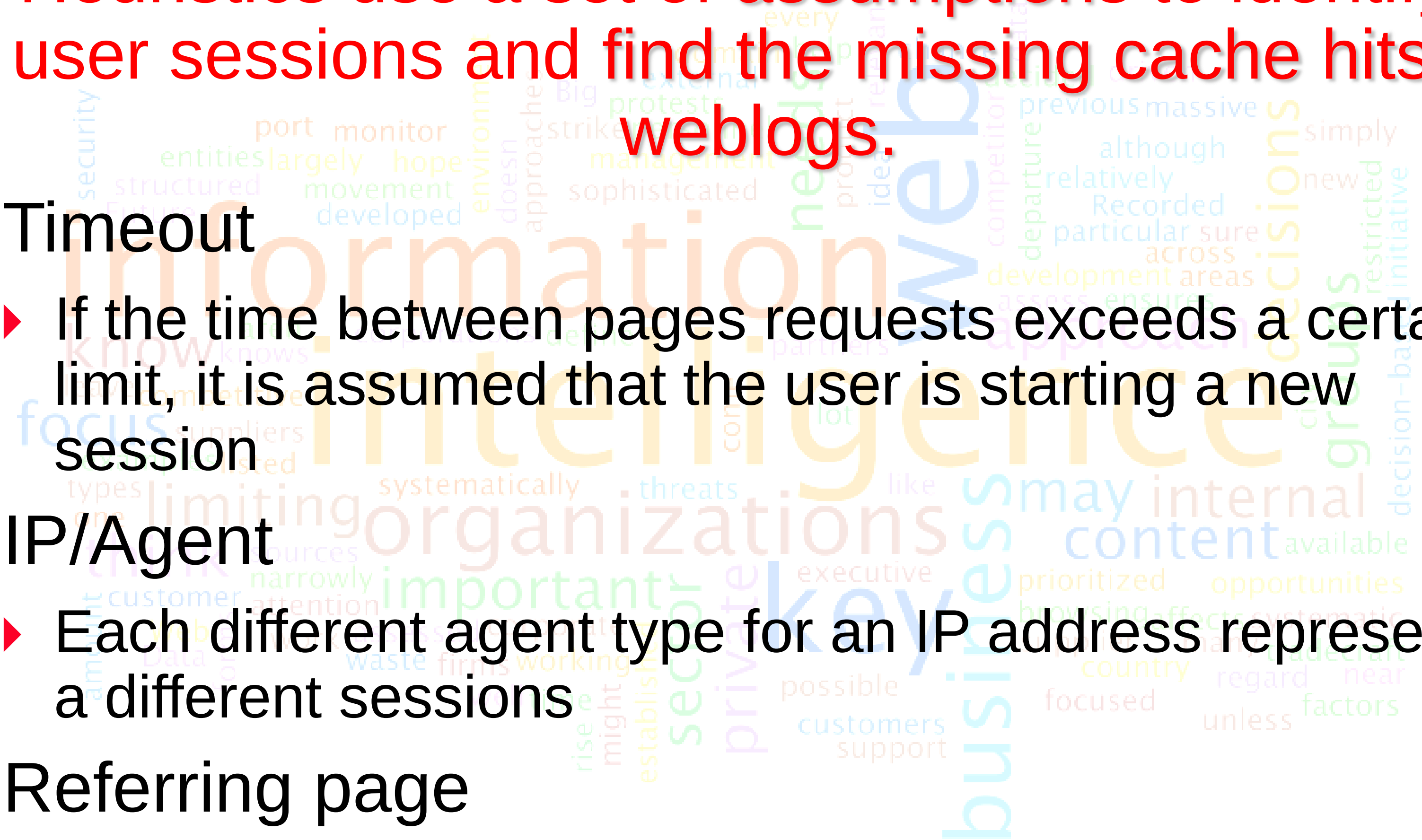
- Sessions are obtained from traditional or basic logs files. Due to the problems that we have identified, in this case real sessions are difficult to be extracted. Usually, heuristics are used.

Proactive Sessionization

- In this case, the movements and actions of users in a Website are directly saved. Several methods to do so exist so far.

Heuristics for reactive sessionization

Heuristics use a set of assumptions to identify user sessions and find the missing cache hits in weblogs.



- 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.

Heuristics for reactive sessionization (2)

- 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

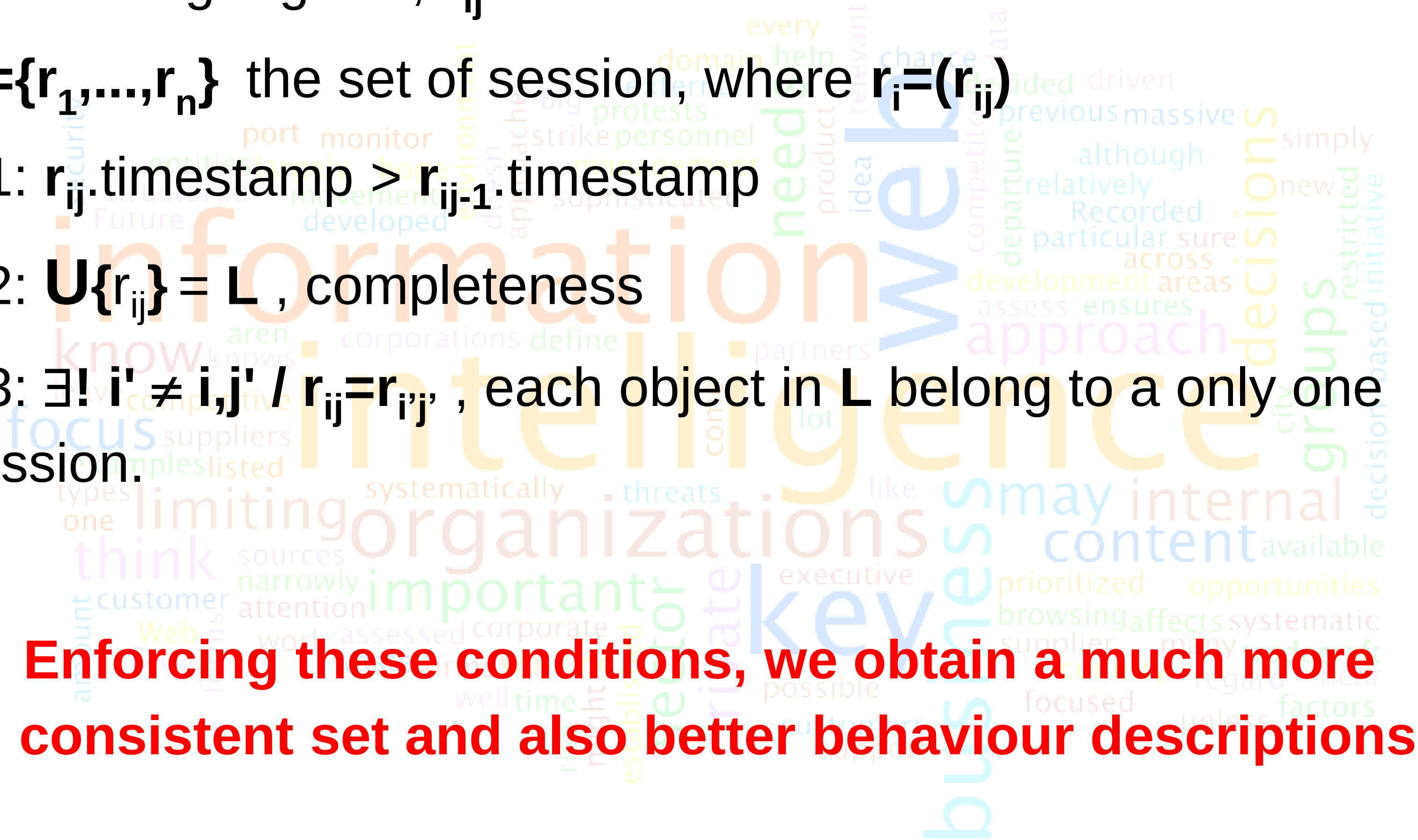
Heuristic-based session reconstruction process

1. **Filtering**: Select **only the relevant log register**, relevant to web pages. Eliminating request for pictures, videos or errors code.
2. **Grouping**: **IP** and **Agents** can be a good selector of client sessions.
3. **Discriminating sessions by time stamps**: Register are selected by **time windows of 30 minutes.**
4. **Identifying irregular session**: **robot or spider** that could be detected looking at the Agent field.
5. **Real session conditioning**

RDBMS could help with the indexing to the efficiency of the calculations.

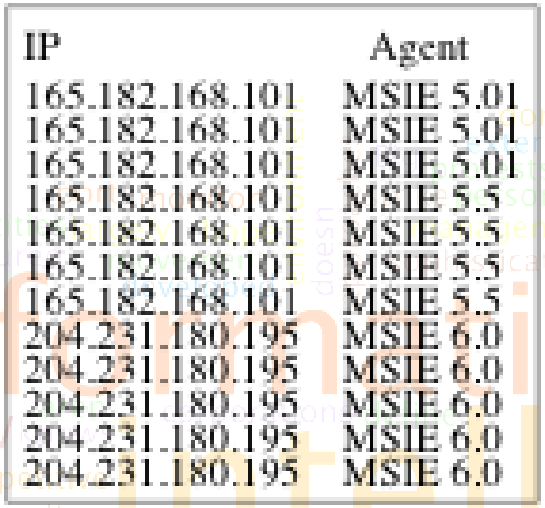
The real session condition

- **L** set of log register, $\mathbf{r_{ij} \in L}$
- $\mathbf{R=\{r_1,\ldots,r_n\}}$ the set of session, where $\mathbf{r_i=(r_{ij})}$
- C1: $\mathbf{r_{ij}}$.timestamp > $\mathbf{r_{ij-1}}$.timestamp
- C2: $\mathbf{U}\{r_{ij}\} = \mathbf{L}$, completeness
- C3: $\mathbf{\exists!\ i' \neq i, j' \,/\, r_{ij}=r_{i'j'}}$, each object in **L** belong to a only one session.

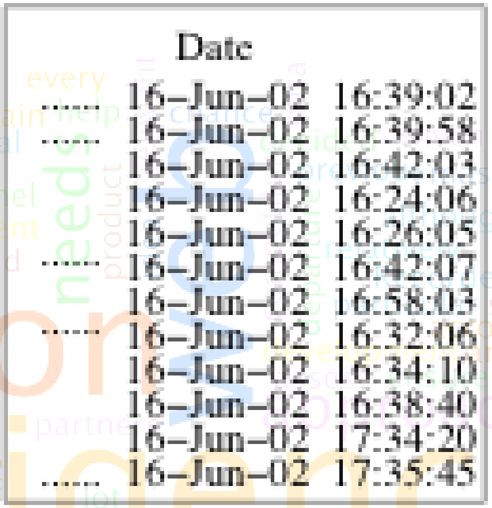


Enforcing these conditions, we obtain a much more consistent set and also better behaviour descriptions.

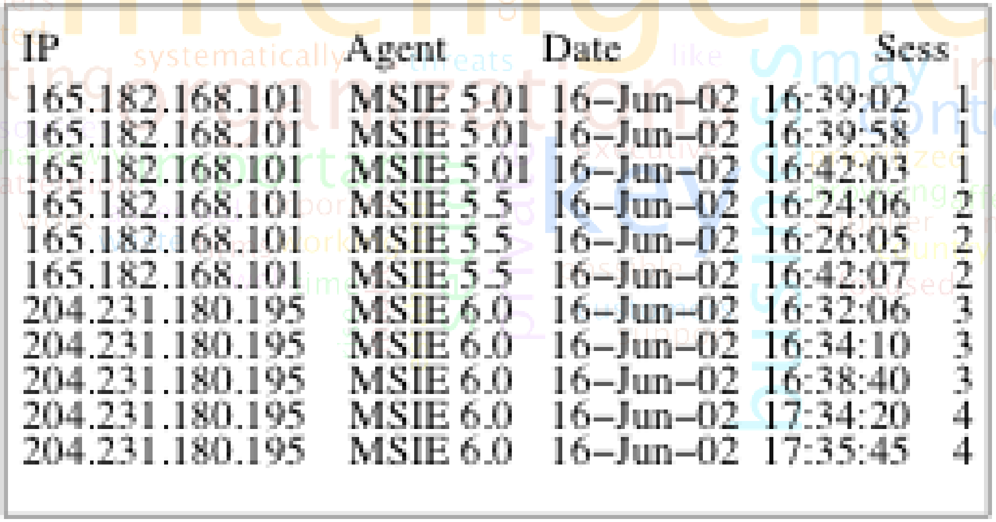
Example of an heuristic-based sessionization process



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

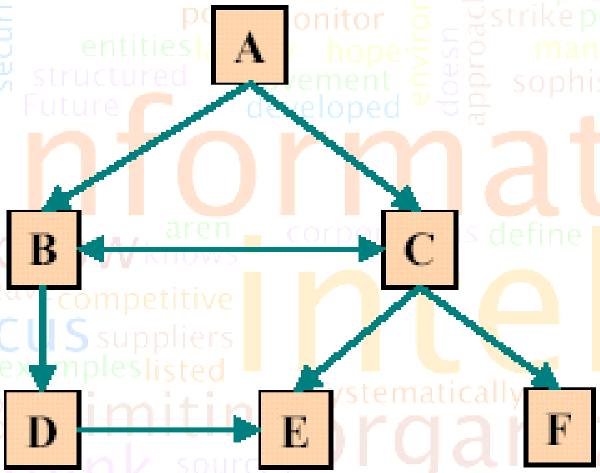


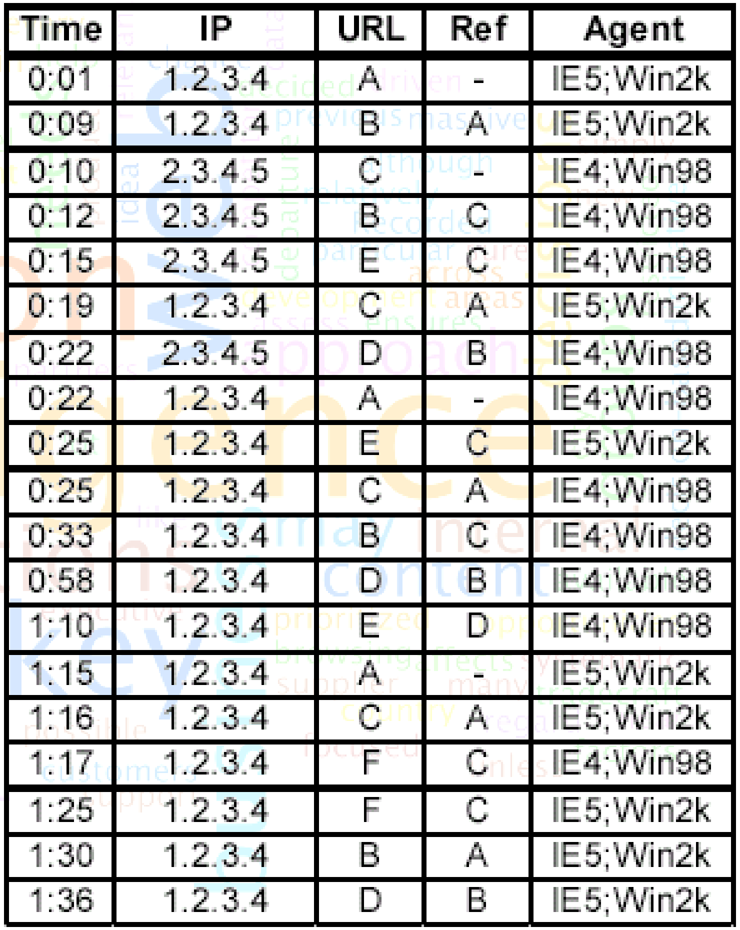
IP	Agent	Date		Sess
165.182.168.101	MSIE 5.01	16–Jun–02	16:39:02	1
165.182.168.101	MSIE 5.01	16–Jun–02	16:39:58	1
165.182.168.101	MSIE 5.01	16–Jun–02	16:42:03	1
165.182.168.101	MSIE 5.5	16–Jun–02	16:24:06	2
165.182.168.101	MSIE 5.5	16–Jun–02	16:26:05	2
165.182.168.101	MSIE 5.5	16–Jun–02	16:42:07	2
204.231.180.195	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

The path completion heuristic

- 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.

The path completion heuristic (2)



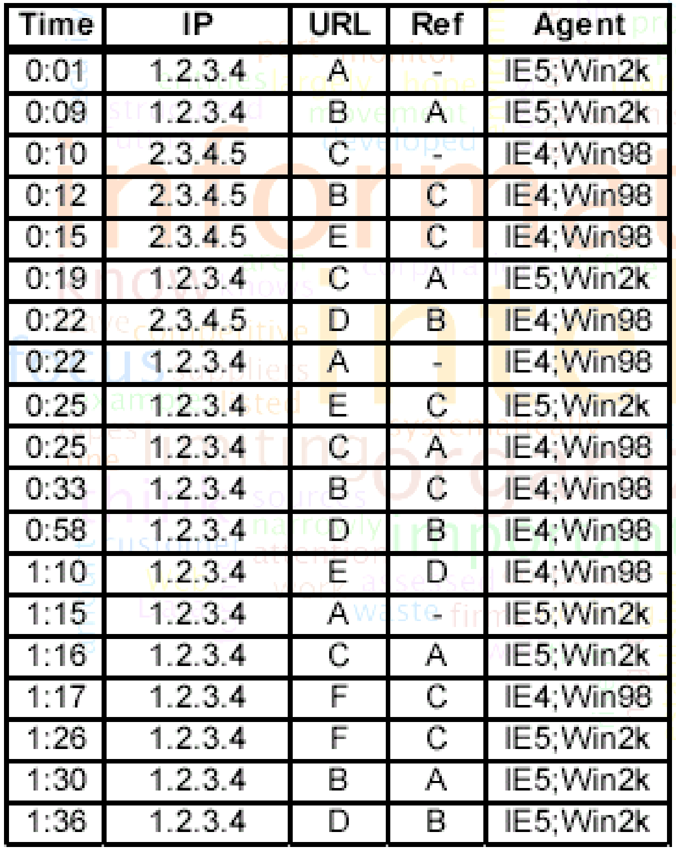


Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:10	2.3.4.5	C	-	IE4;Win98
0:12	2.3.4.5	B	C	IE4;Win98
0:15	2.3.4.5	E	C	IE4;Win98
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE4;Win98
0:22	1.2.3.4	A	-	IE4;Win98
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE4;Win98
0:33	1.2.3.4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

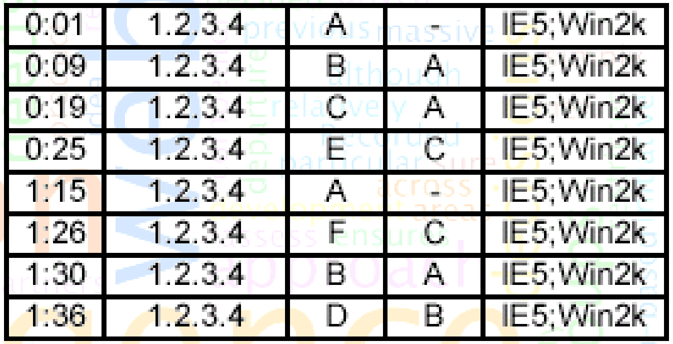
Based on examples in tutorial PKDD-2002 by Berendt et. al.

The path completion heuristic (3)

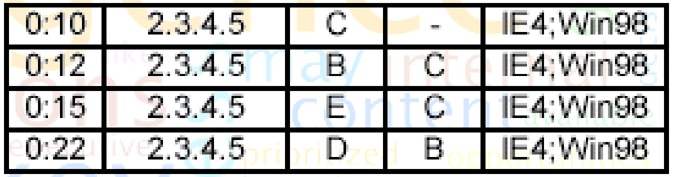
Sort the users (IP+Agent)



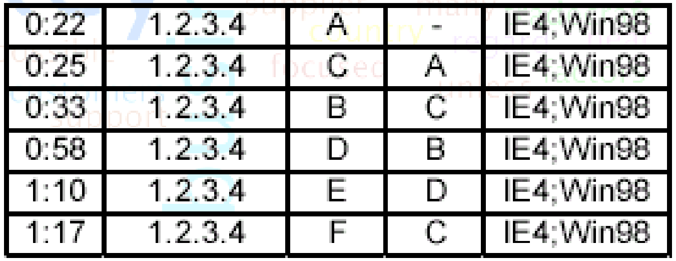
Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:10	2.3.4.5	C	-	IE4;Win98
0:12	2.3.4.5	B	C	IE4;Win98
0:15	2.3.4.5	E	C	IE4;Win98
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE4;Win98
0:22	1.2.3.4	A	-	IE4;Win98
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE4;Win98
0:33	1.2.3.4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



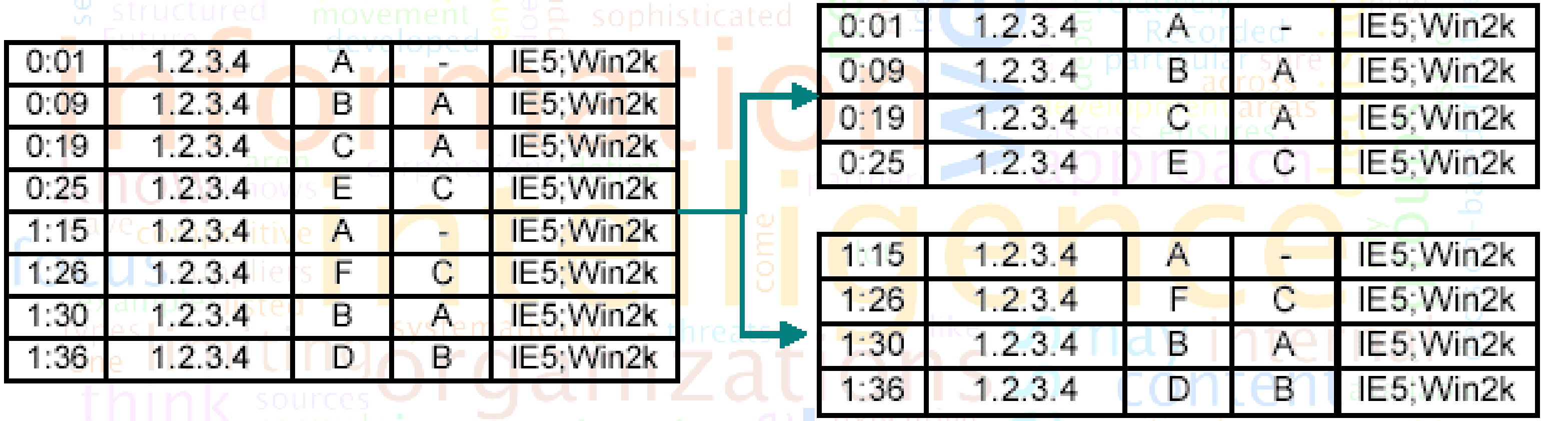
0:10	2.3.4.5	C	-	IE4;Win98
0:12	2.3.4.5	B	C	IE4;Win98
0:15	2.3.4.5	E	C	IE4;Win98
0:22	2.3.4.5	D	B	IE4;Win98



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

The path completion heuristic (4)

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

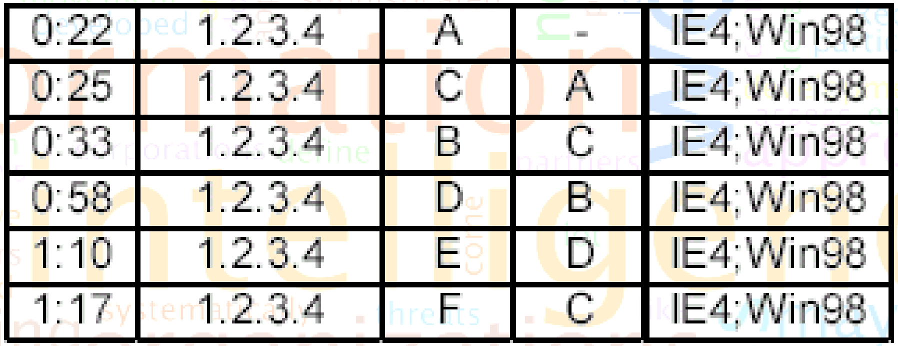
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:19	1.2.3.4	C	A	IE5;Win2k
0:25	1.2.3.4	E	C	IE5;Win2k

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

The path completion heuristic (5)

Sessionize using heuristics (with href)



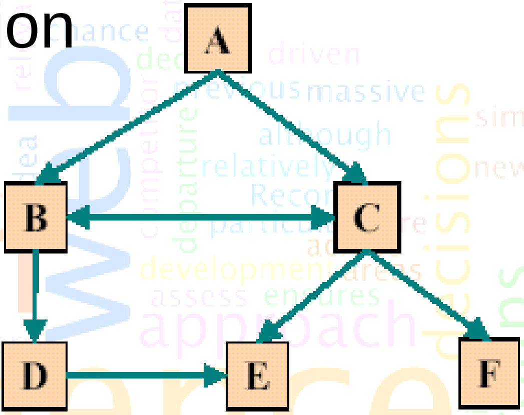
0:22	1.2.3.4	A	-	IE4;Win98
0:25	1.2.3.4	C	A	IE4;Win98
0:33	1.2.3.4	B	C	IE4;Win98
0:58	1.2.3.4	D	B	IE4;Win98
1:10	1.2.3.4	E	D	IE4;Win98
1:17	1.2.3.4	F	C	IE4;Win98

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

The path completion heuristic (6)

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

The path completion heuristic (7)

- What happens 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 behaviour, it is necessary to consider to maintain a register of changes.
- Proposal solution LOGML [Punin WEBKDD'01]

Proactive sessionization methods (1)

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

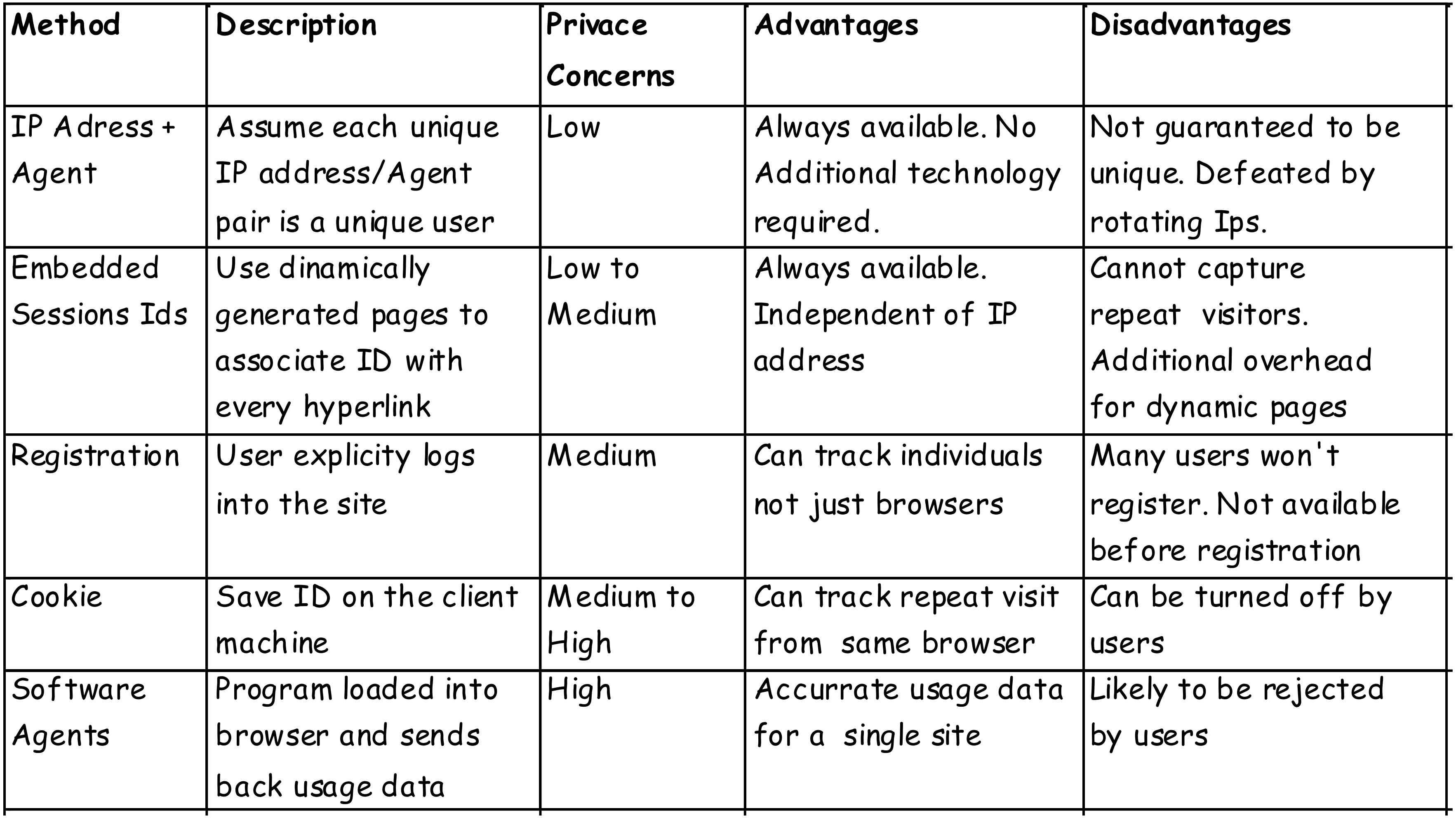
Modified Browser

- 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

In summary



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

Section 5.4

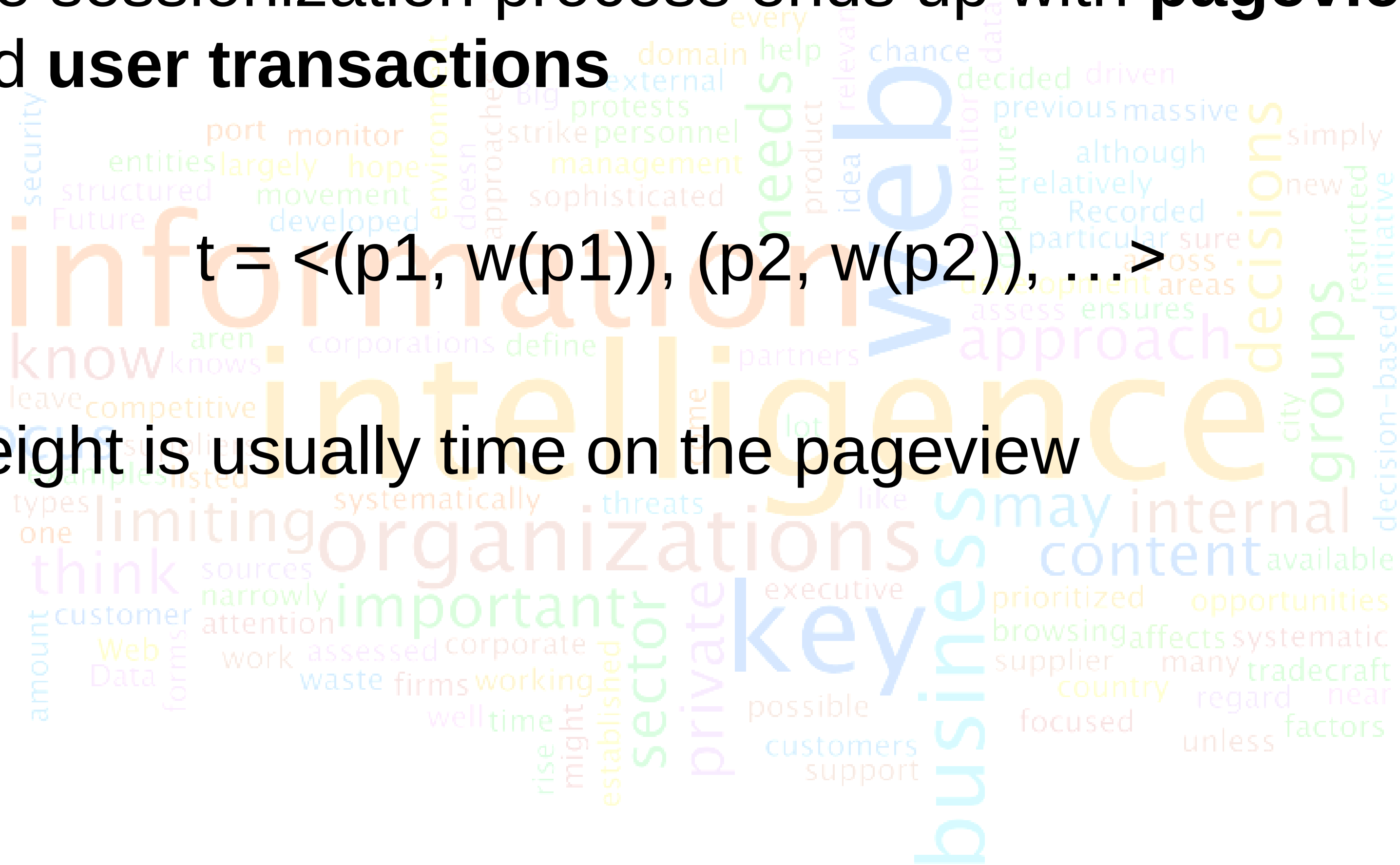
Data modelling for web usage mining

Data modelling for web usage mining

- The sessionization process ends up with **pageviews** and **user transactions**

$$t = <(p1, w(p1)), (p2, w(p2)), \ldots>$$

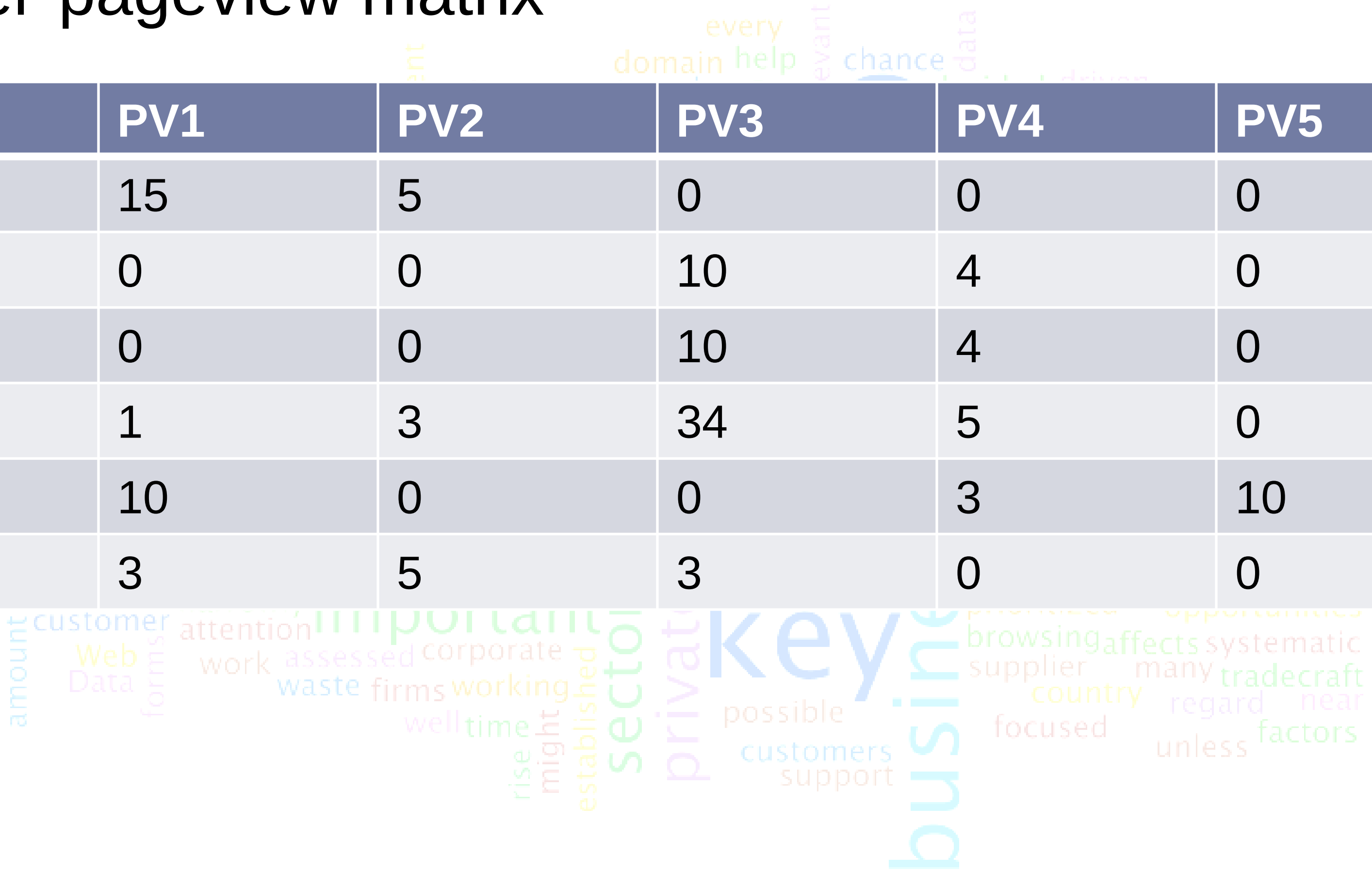
- Weight is usually time on the pageview



Data modelling for web usage mining

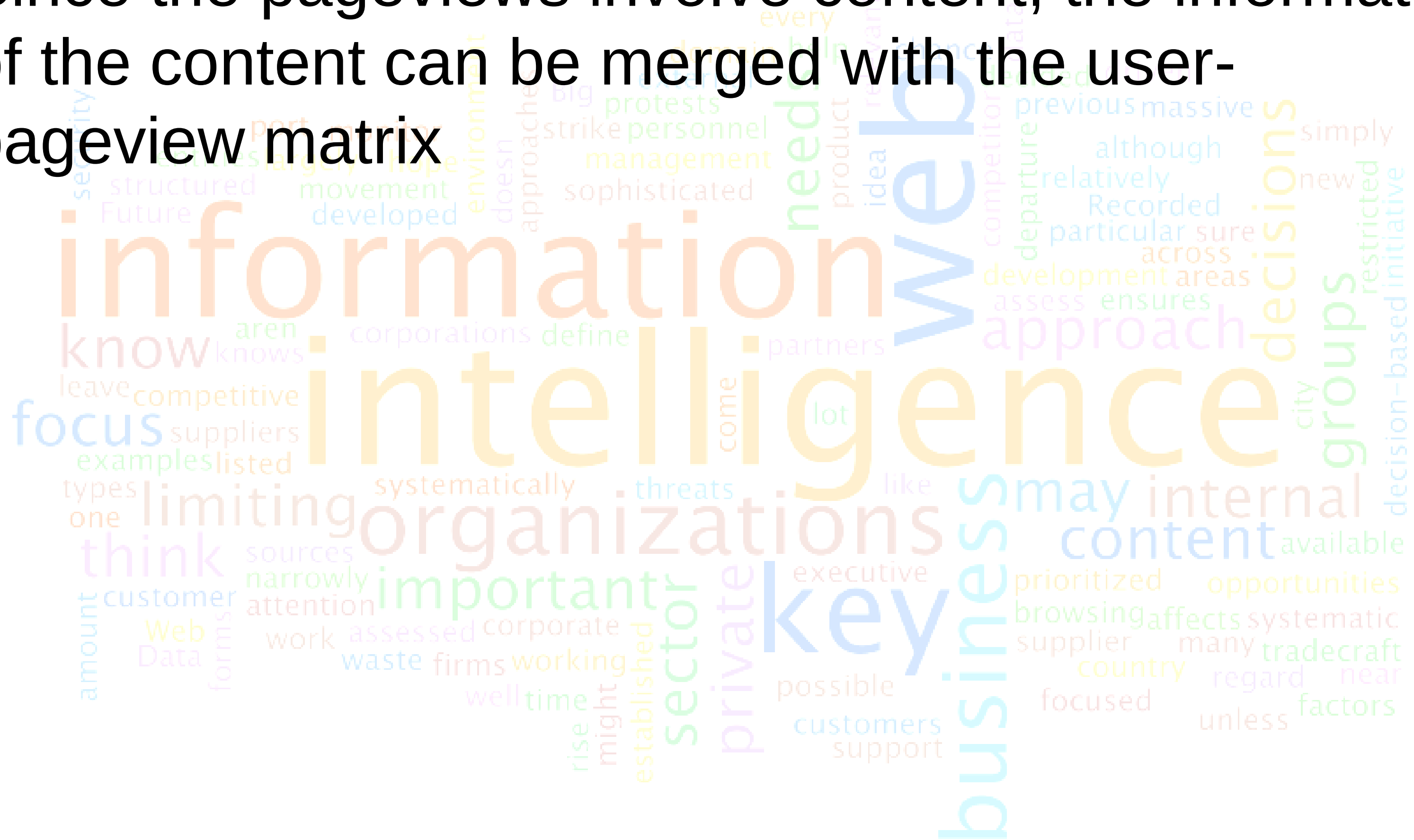
- User-pageview matrix

	PV1	PV2	PV3	PV4	PV5
User1	15	5	0	0	0
User2	0	0	10	4	0
User3	0	0	10	4	0
User4	1	3	34	5	0
User5	10	0	0	3	10
User6	3	5	3	0	0



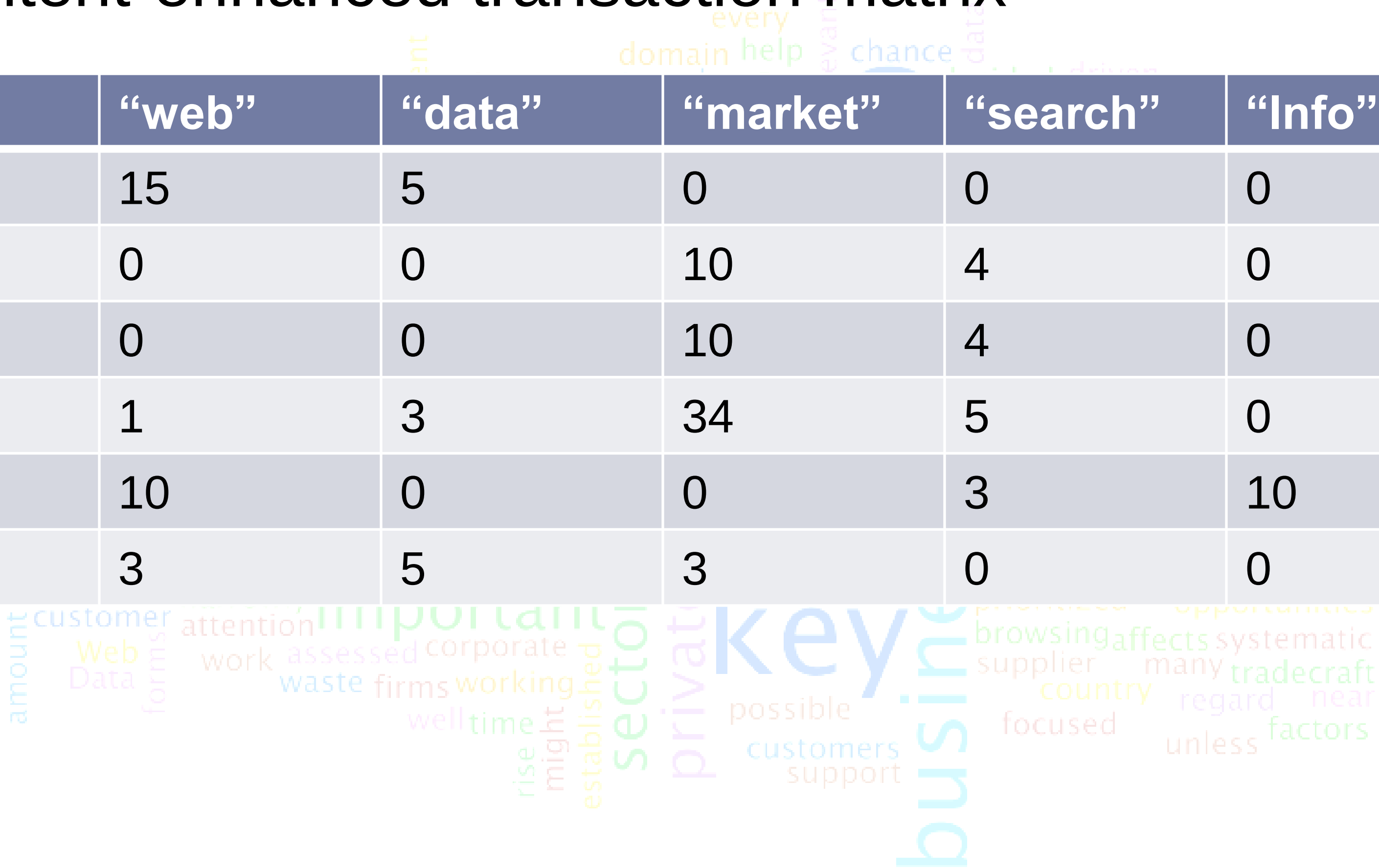
Data modelling for web usage mining

- Since the pageviews involve content, the information of the content can be merged with the user-pageview matrix



Data modelling for web usage mining

- Content-enhanced transaction matrix



	"web"	"data"	"market"	"search"	"Info"
User1	15	5	0	0	0
User2	0	0	10	4	0
User3	0	0	10	4	0
User4	1	3	34	5	0
User5	10	0	0	3	10
User6	3	5	3	0	0

Section 5.5

Classification of the user behavior in a Web site

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 in web usage mining.

Classification (2)

- 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 (3)

- 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).

Section 5.6

Using association rules for discovering navigation patterns

Association Rule Generation

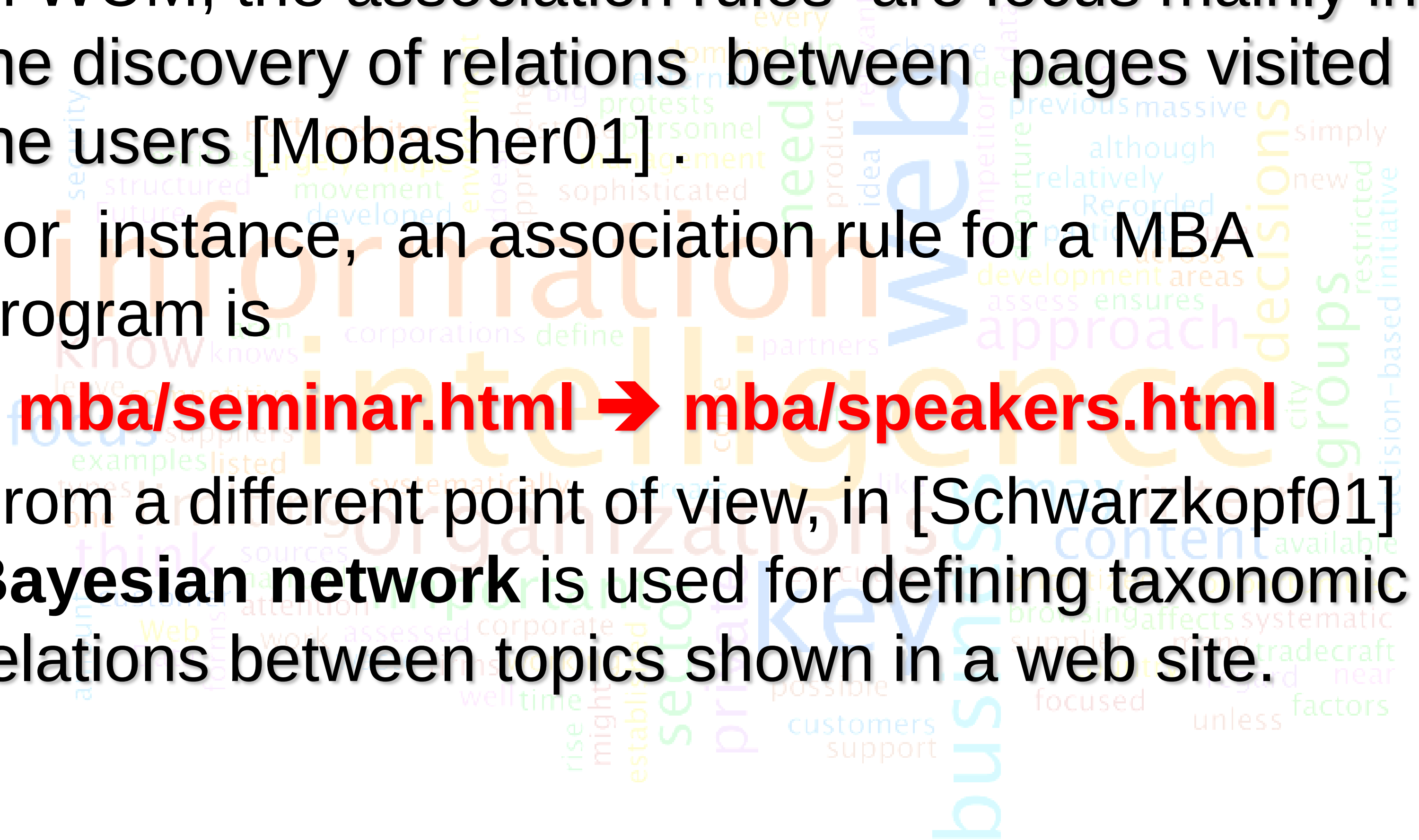
- Discovers the correlations between pages that are most often referenced together in a single server session
- Provide the information
  - What are the set of pages frequently accessed together by Web users?
  - What page will be fetched next?
  - What are paths frequently accessed by Web users?
- Association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

  A → B [Support = 60%, Confidence = 80%]
- **Example:** “80% of visitors who accessed URLs /diplomas.html and BI/info.html also visited webmining.html”

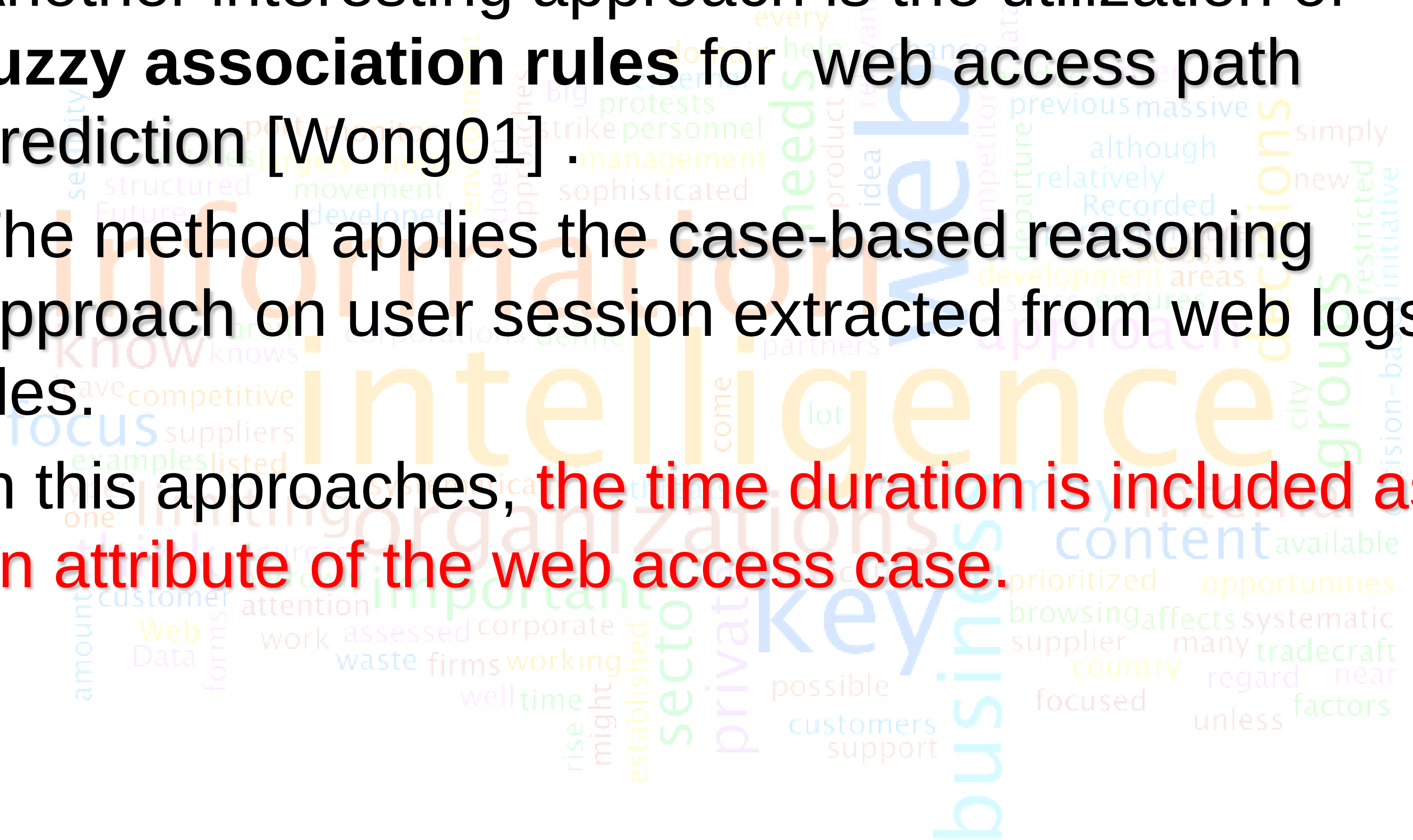
Association rules

- 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)

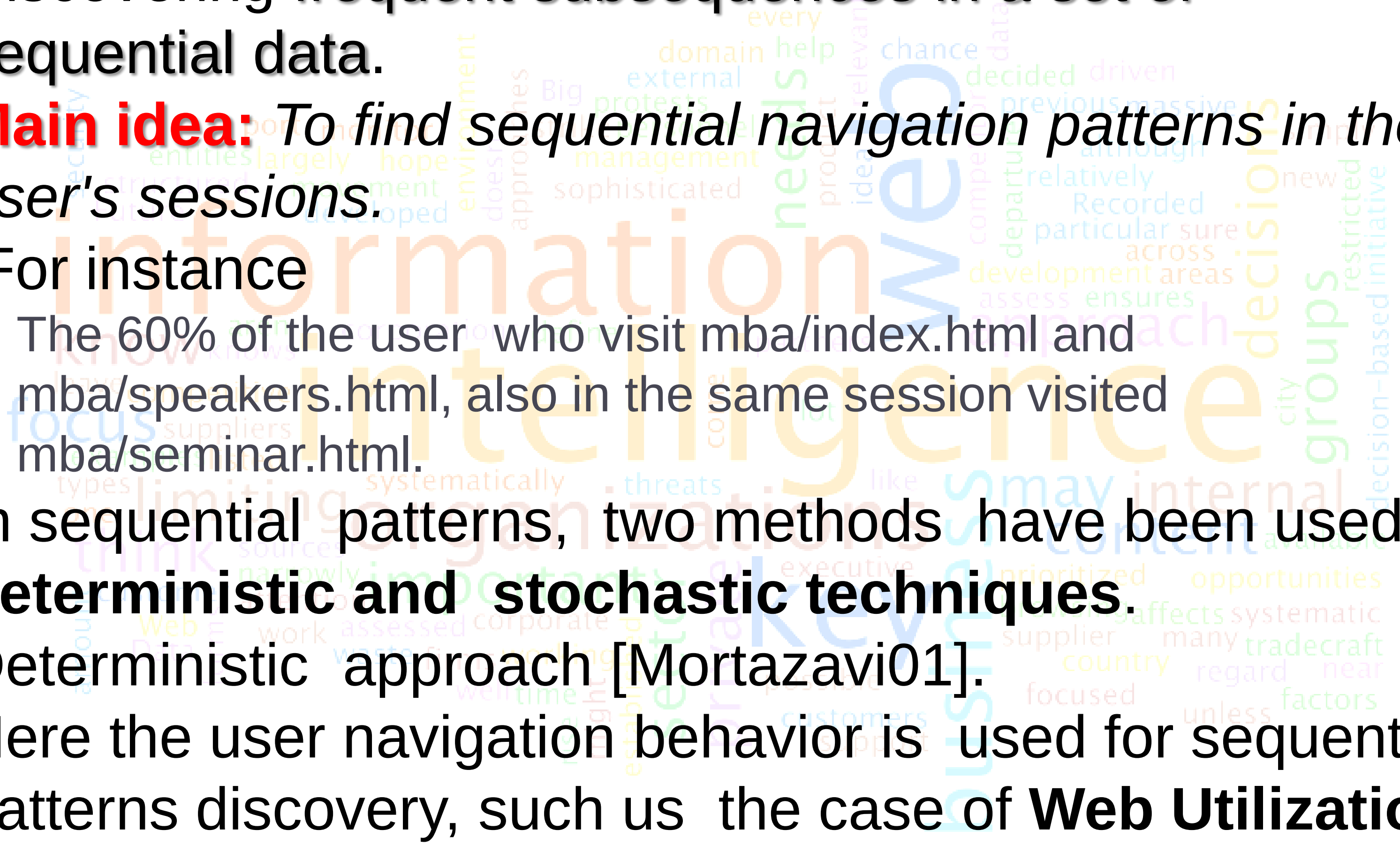


- Another interesting approach is the utilization of **fuzzy association rules** for web access path prediction [Wong01] .
- The method applies the case-based reasoning approach on user session extracted from web logs files.
- In this approaches, the time duration is included as an attribute of the web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

Section 5.7

Using sequence patterns for discovering common access paths

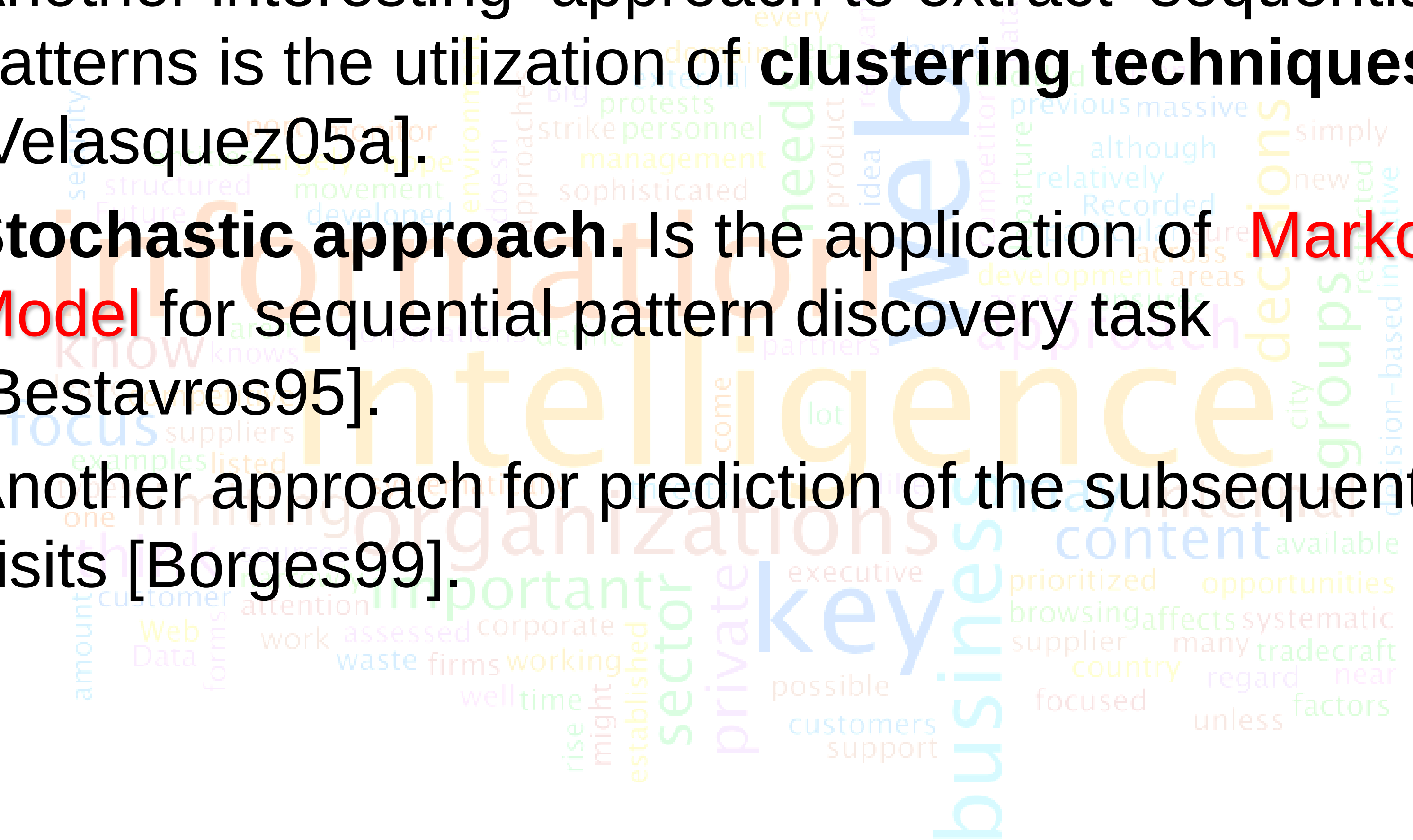
Sequence patterns



- 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].

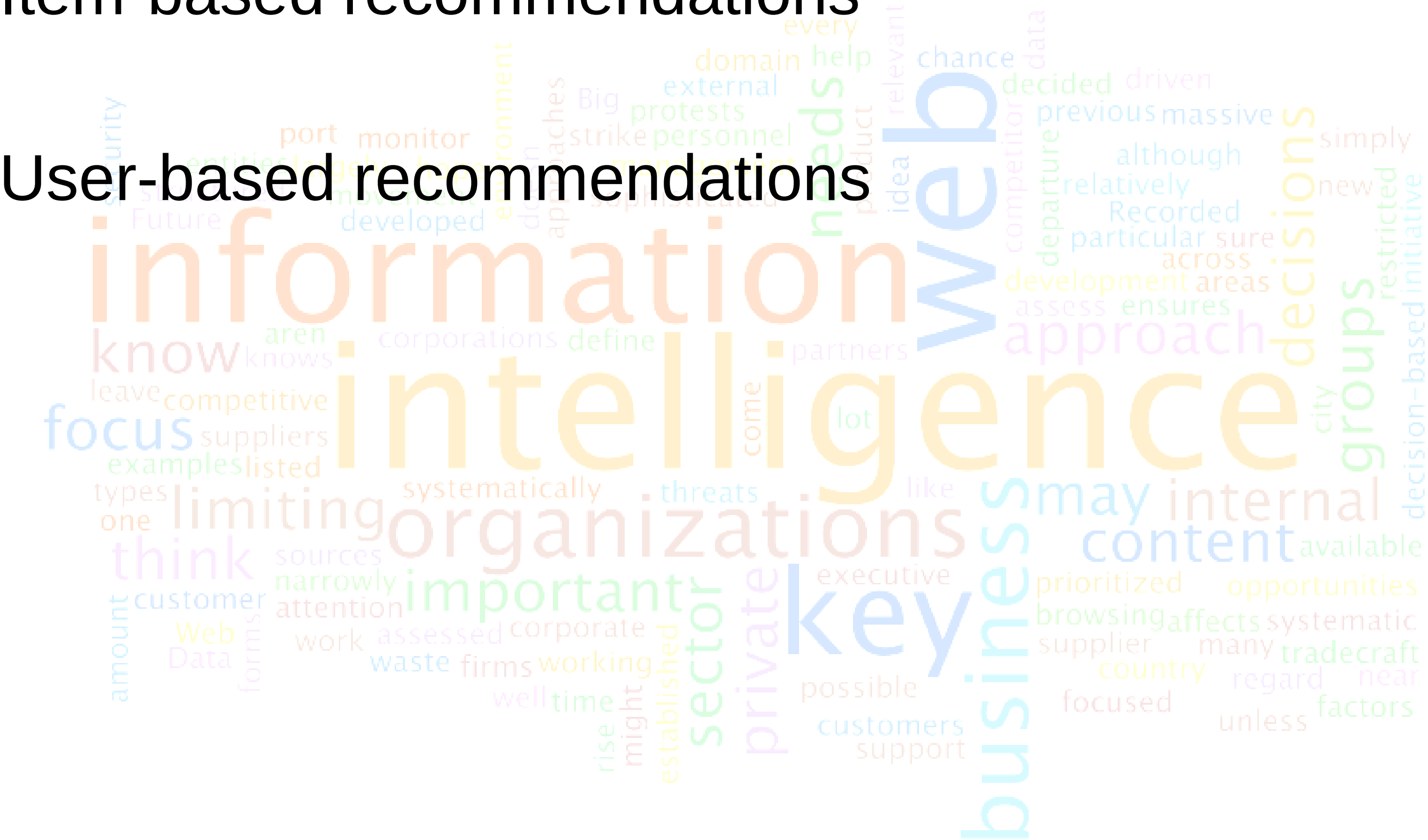


Section 5.8

Recommendations based on web user transactions

What page to go next?

- Item-based recommendations
- User-based recommendations



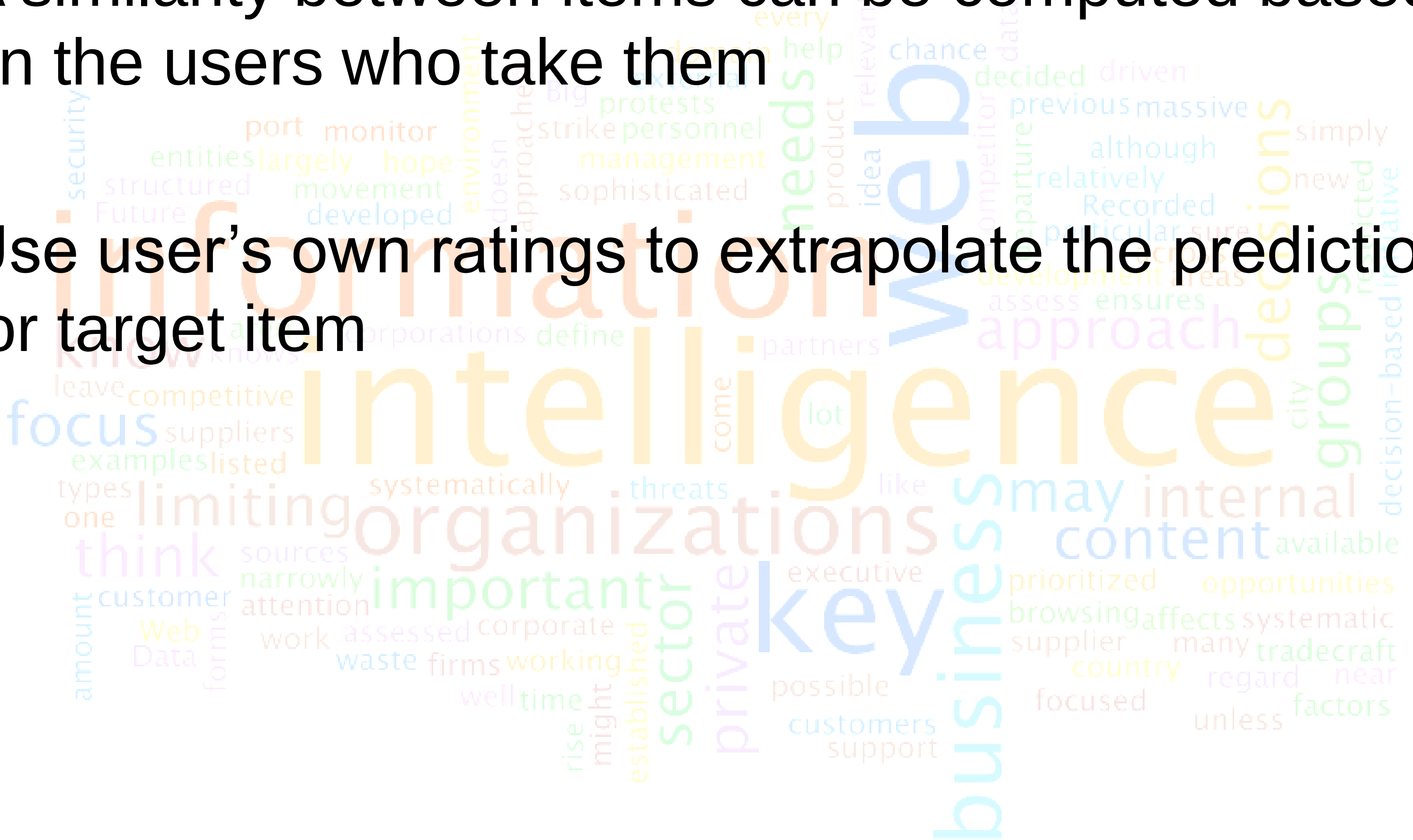
User-based recommendations

- For a user, the system finds their neighbors
- Similarity between users is computed based on the ratings they have made to common items
- Rating to unknown items are computed based on the similarity between the users and the new user and their ratings on the unknown item
- Does not scale well

Item-based recommendations

- A similarity between items can be computed based on the users who take them

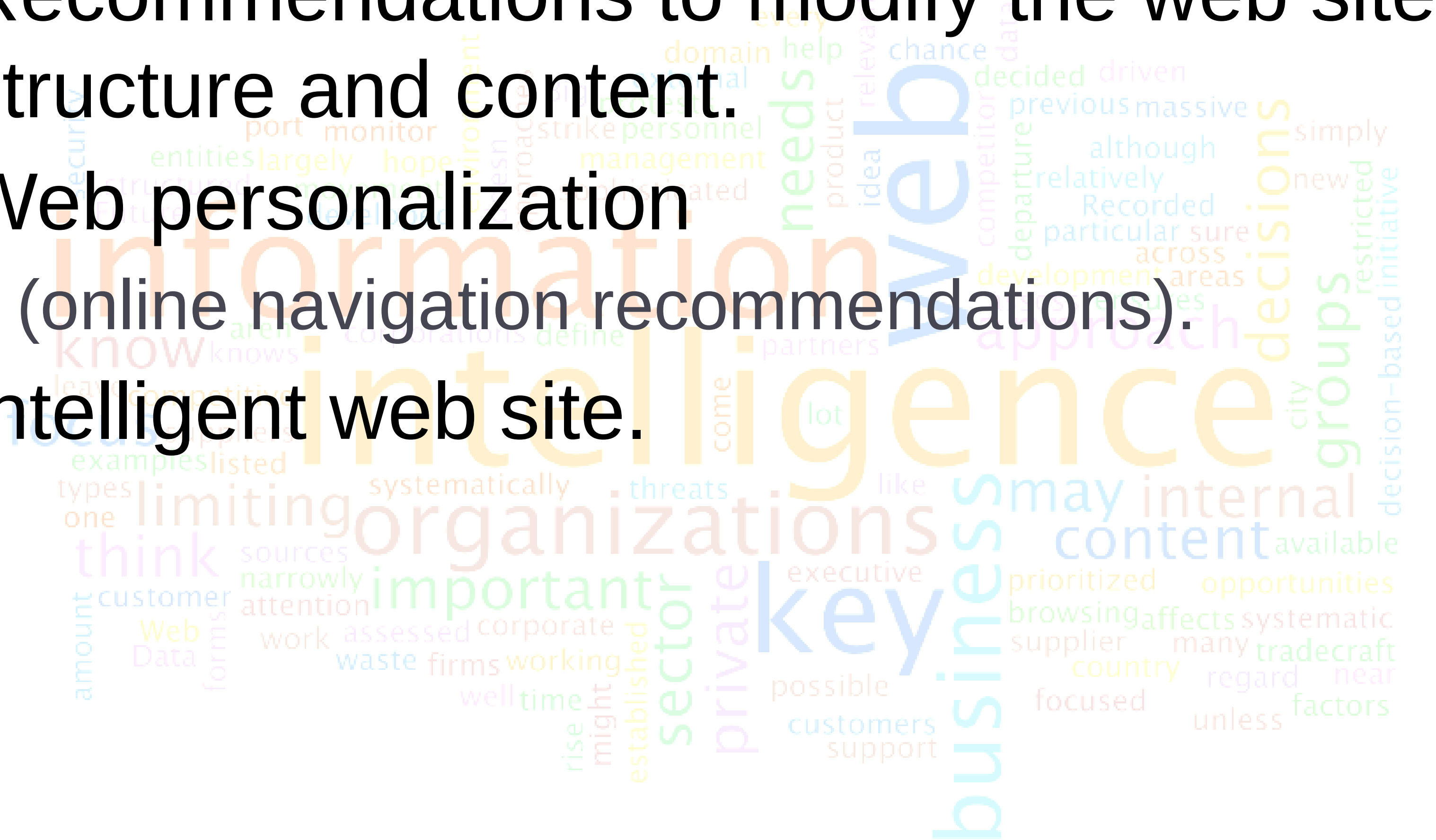
- Use user's own ratings to extrapolate the prediction for target item



Section 5.9

WUM Applications

Improving the relationship between the web site and the user



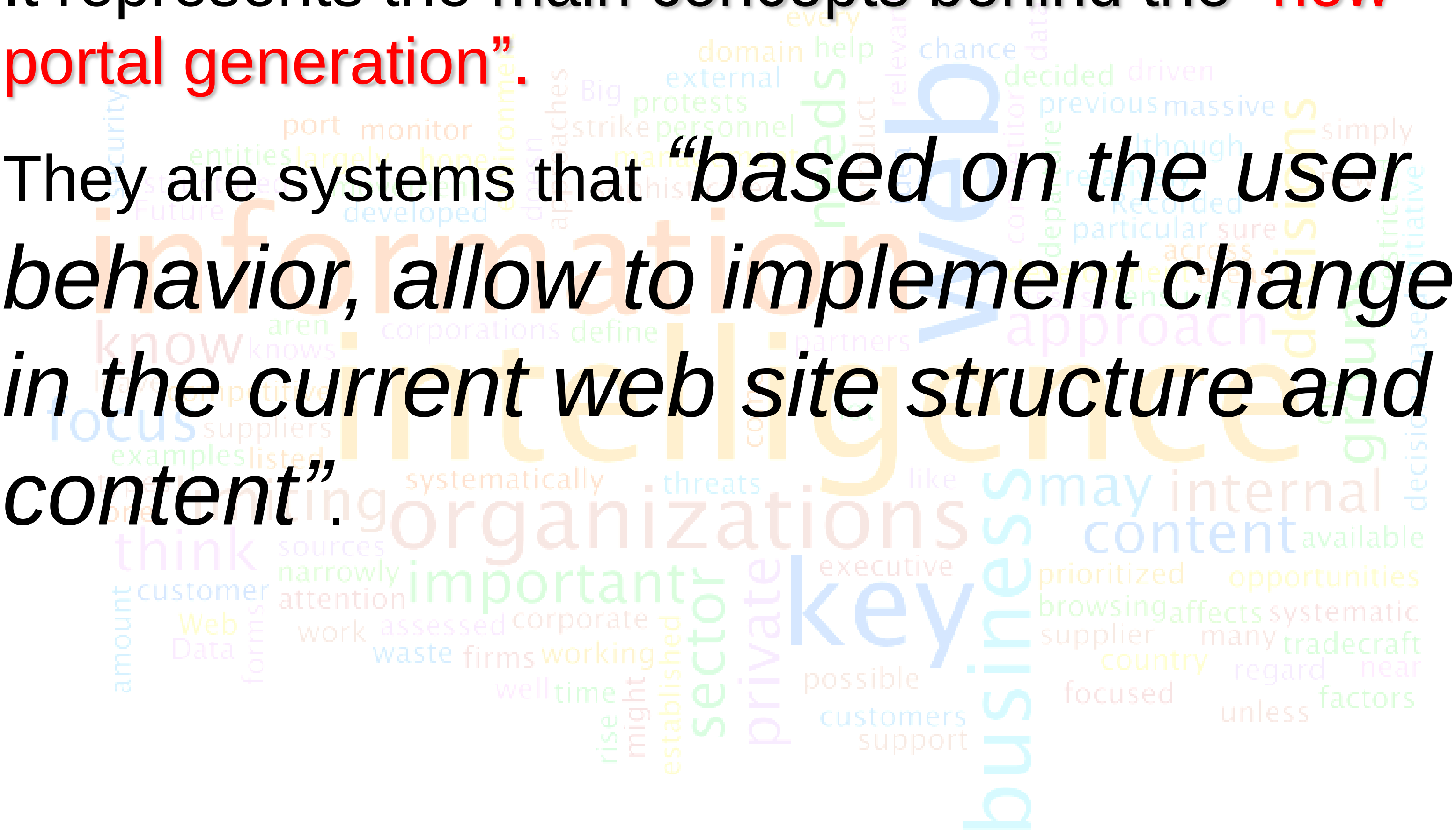
- Recommendations to modify the web site structure and content.
- Web personalization
  - (online navigation recommendations).
- Intelligent web site.

Web personalization [Lu03,Mombasher00]

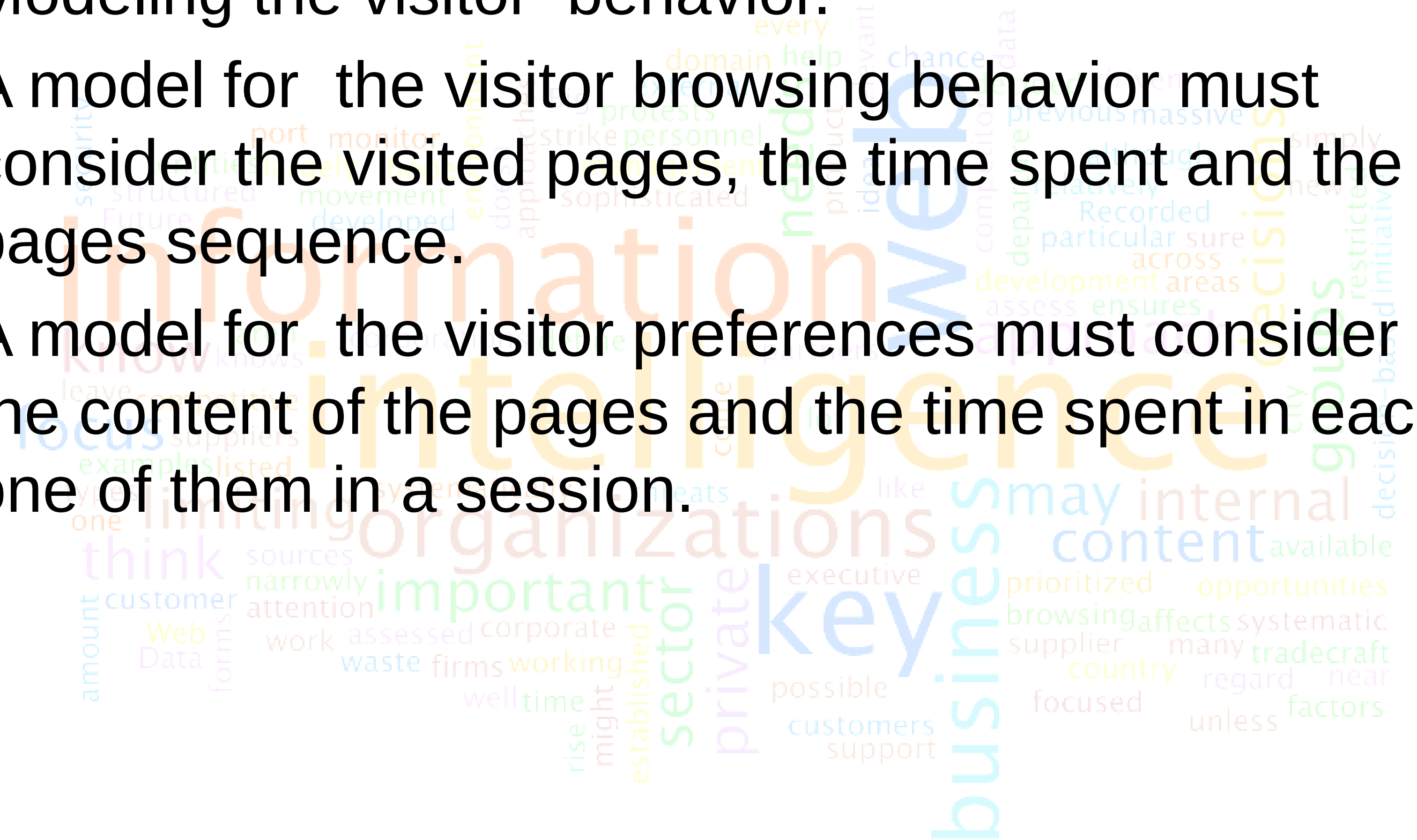
- It is the process where the web server and the related applications, dynamically, customize the content for a particular user, based on information about **his/her behavior in a website.**
- This is different to another related concept called "**customization**" where the visitor interacts with the web server using an interface to create her/his own web site, e.g., "MyBanking Page".

Adaptive Web site [Coenen00, Perkowitz98, Velasquez04b]

- It represents the main concepts behind the "new portal generation".
- They are systems that *"based on the user behavior, allow to implement changes in the current web site structure and content"*.

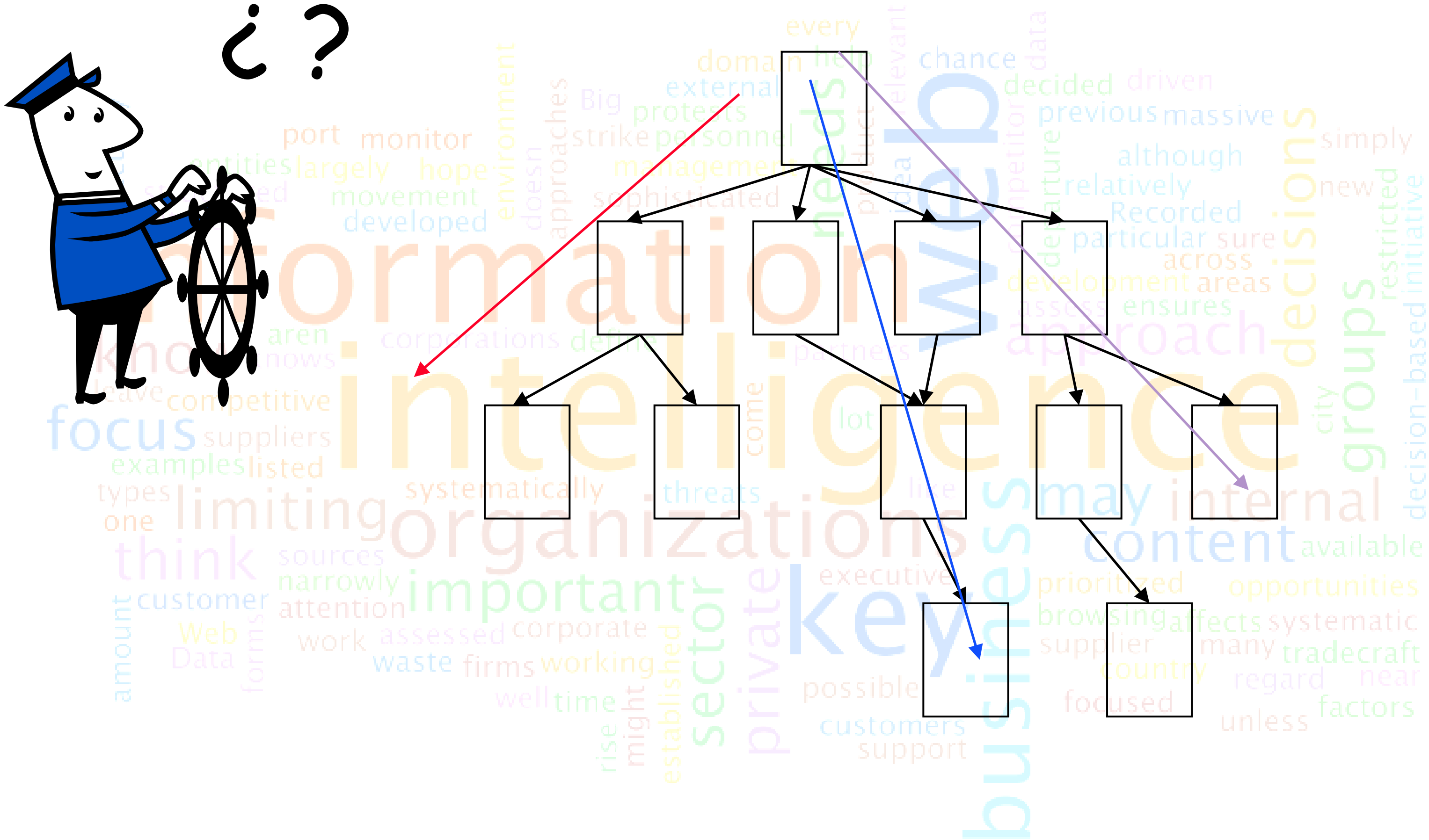


What we should do? [Mombasher01]



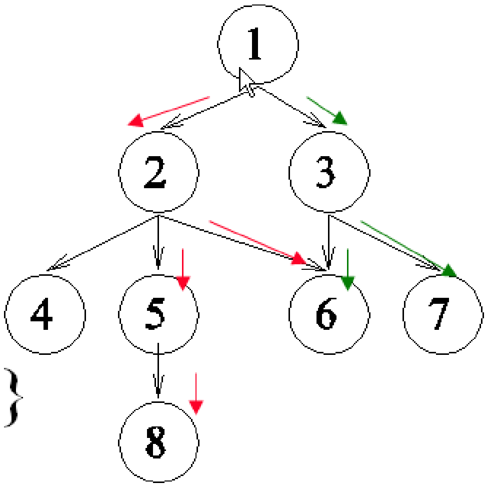
- Modeling the visitor behavior.
- A model for the visitor browsing behavior must consider the visited pages, the time spent and the pages sequence.
- A model for the visitor preferences must consider the content of the pages and the time spent in each one of them in a session.

Visitor browsing behavior



Comparing the sequences [Runkler03, Velasquez04a]

- The sequence of a navigation can be represented by a graph. Each page is identified by an identification number.



S_1= (1,2,6,5,8)

S_2=(1,3,6,7)

$$G_1 = \{1 \to 2, 2 \to 6, 2 \to 5, 5 \to 8\}$$

$$G_2 = \{1 \to 3, 3 \to 6, 3 \to 7\}$$

$$E(G_1) = \{1,2,6,5,8\}$$

$$E(G_2) = \{1,3,6,7\}$$

- We need to know how similar or different are both sequence representations!!

Notion of similarity

$$dG(G_1, G_2) = 2\frac{\|E(G_1) \cap E(G_2)\|}{\|E(G_1) + E(G_2)\|} = \begin{cases} 0 & if \quad E(G_1) \cap E(G_2) = \phi \\ 1 & if \quad E(G_1) \equiv E(G_2) \end{cases}$$

Comparing sequences: An example

$S_1 = "12648"$

$S_2 = "1367"$

$S_3 = "12856"$

$S_4 = "1367"$

$L(S_1, S_2) = 3$

$L(S_3, S_4) = 4$

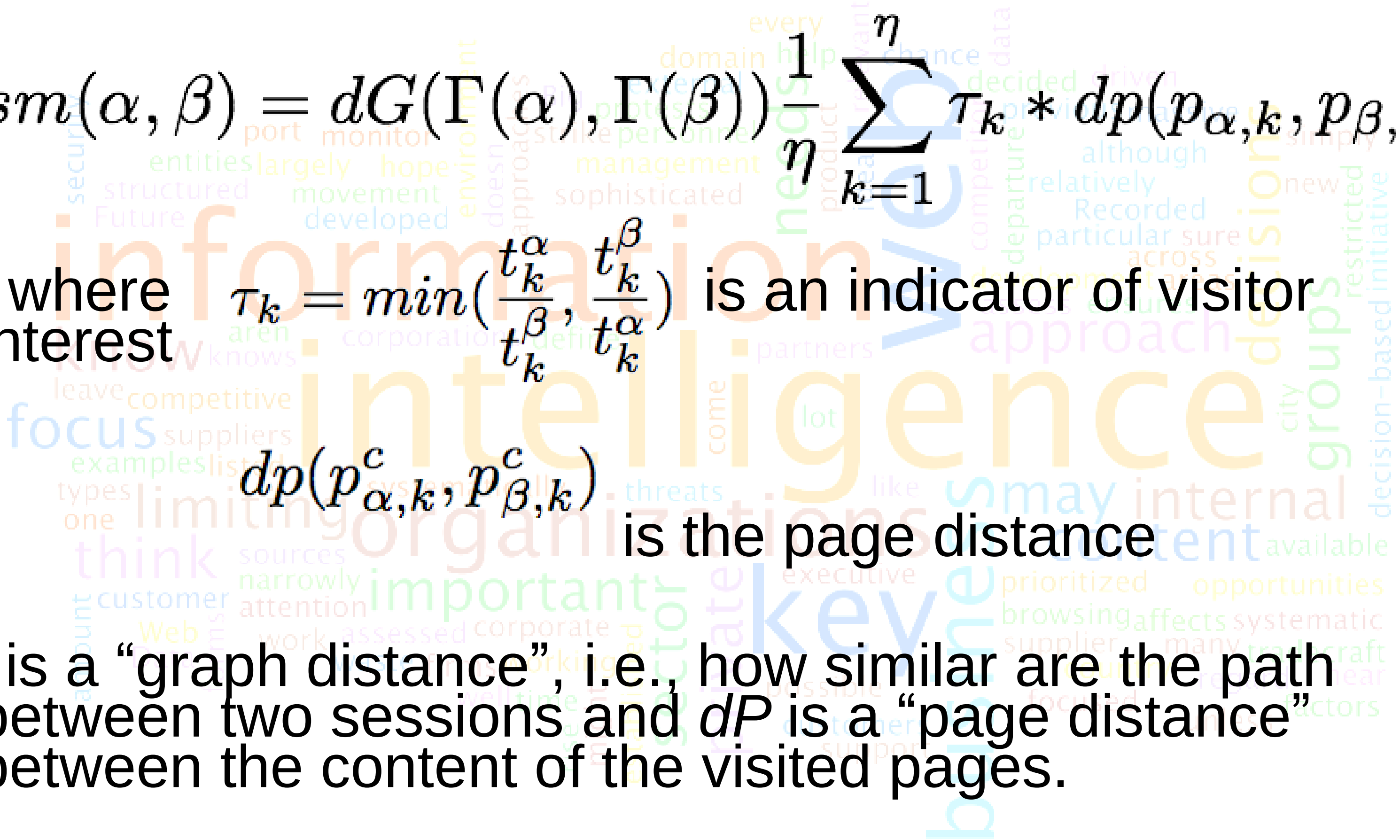
$$L(S_1, S_2) = \begin{cases} 0 & S_1 \equiv S_2 \\ \left]0, \max\{\| E(G_1) \|, \| E(G_2) \|\}\right] & else \end{cases}$$

$$dG(G_1, G_2) = 1 - 2 \frac{L(S_1, S_2)}{\| E(G_1) \| + \| E(G_2) \|}$$

$$= 1 - 2 \tfrac{3}{5+4} = 0,\overline{3}$$

Comparing browsing behavior

- Then the similarity measure is:



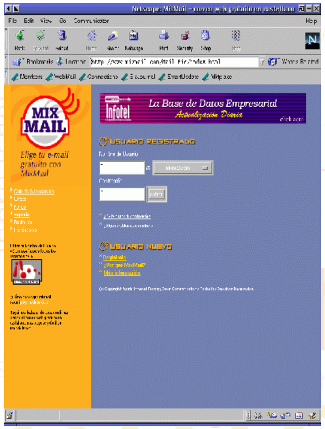
$$sm(\alpha,\beta) = dG(\Gamma(\alpha),\Gamma(\beta))\frac{1}{\eta}\sum_{k=1}^{\eta}\tau_k * dp(p_{\alpha,k},p_{\beta,k})$$

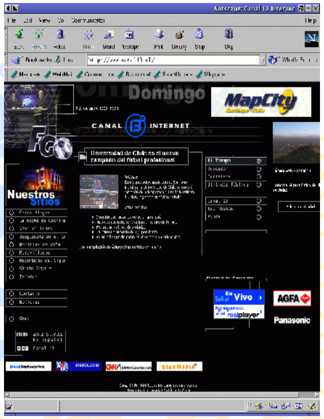
where $\tau_k = min(\frac{t_k^\alpha}{t_k^\beta},\frac{t_k^\beta}{t_k^\alpha})$ is an indicator of visitor interest

$dp(p^c_{\alpha,k},p^c_{\beta,k})$ is the page distance

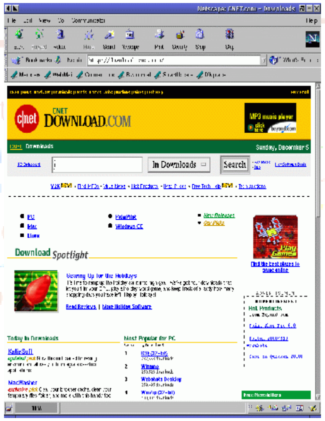
dG is a “graph distance”, i.e., how similar are the path between two sessions and *dP* is a “page distance” between the content of the visited pages.

What is she/he looking for?



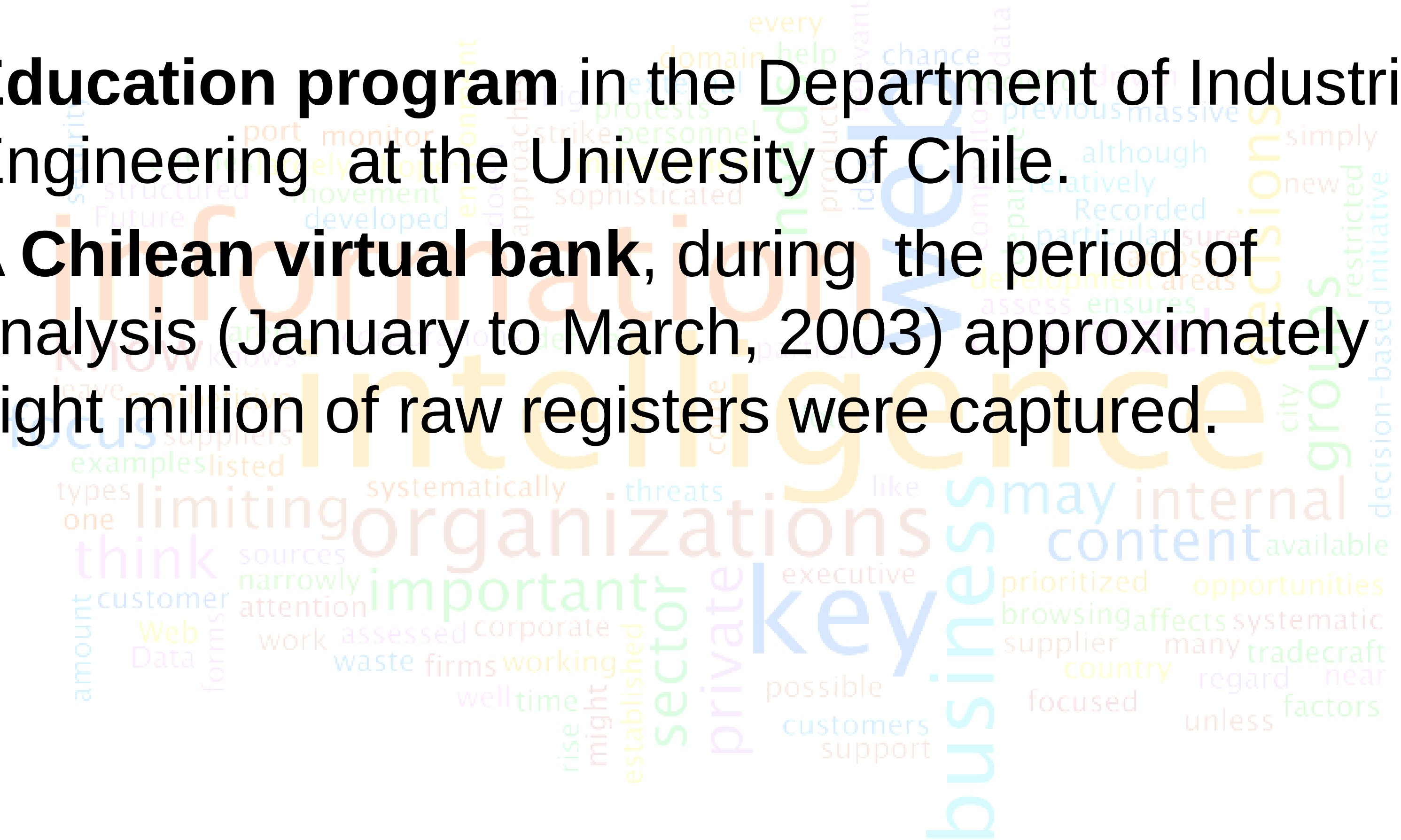


- Free text.
- Movies
- Pictures
- Sounds



Web site selection

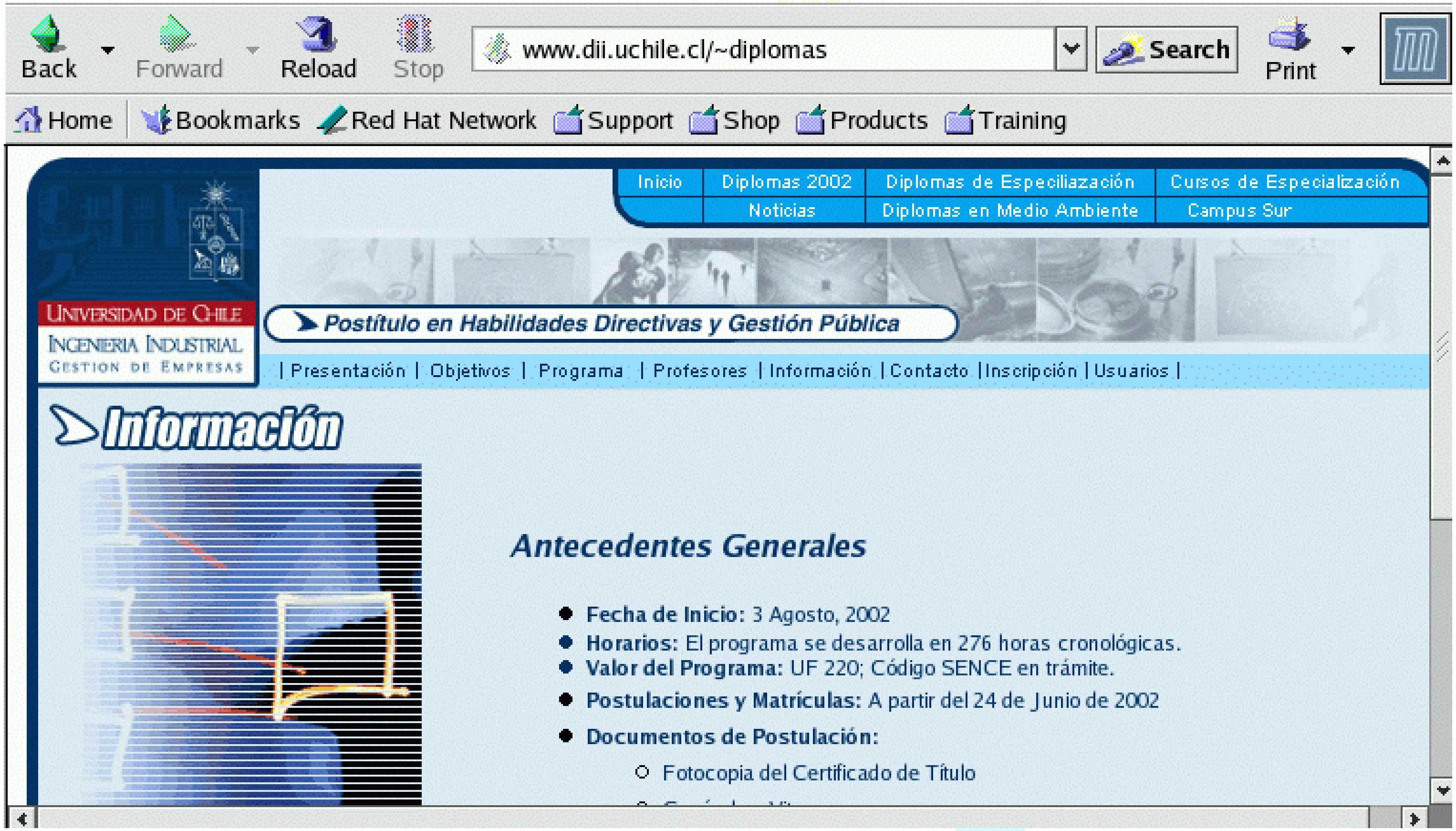
- **Education program** in the Department of Industrial Engineering at the University of Chile.
- A **Chilean virtual bank**, during the period of analysis (January to March, 2003) approximately eight million of raw registers were captured.



Educational web site

- The web site characteristics are:
  - 142 static web pages written in Spanish.
  - Approximately 24,000 web logs registers were considered, corresponding to the period from August to October 2002.
- 4113 visitor behavior vectors were identified.
- Pages=122, different words identified=6234 (R=6234 and Q=122).

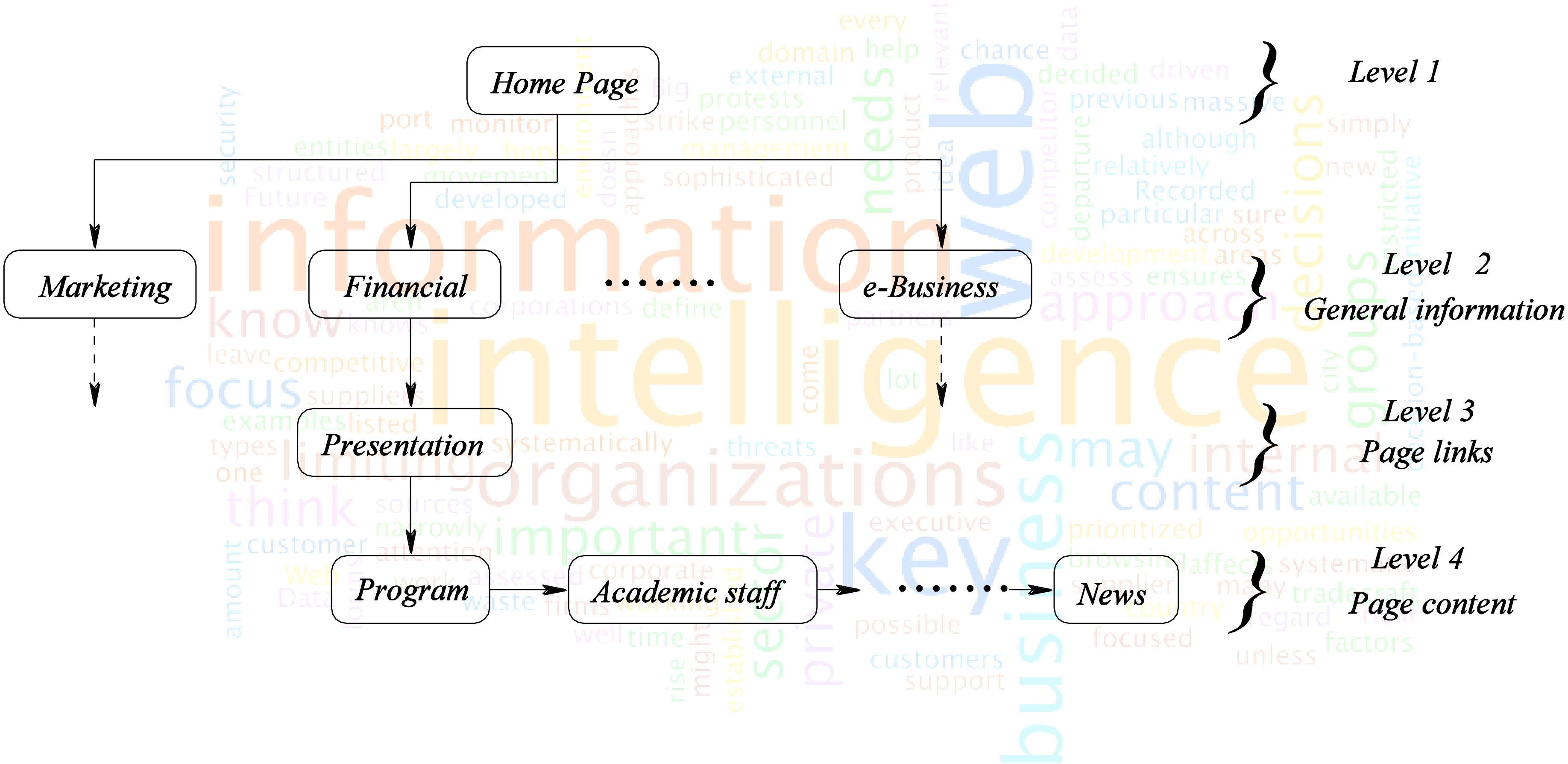
Educational web site: old page template


Back
Forward
Reload
Stop
www.dii.uchile.cl/~diplomas
Search
Print
Home
Bookmarks
Red Hat Network
Support
Shop
Products
Training
Inicio
Diplomas 2002
Diplomas de Especilaización
Cursos de Especialización
Noticias
Diplomas en Medio Ambiente
Campus Sur
Universidad de Chile
Ingeniería Industrial
Gestión de Empresas
Postítulo en Habilidades Directivas y Gestión Pública
| Presentación | Objetivos | Programa | Profesores | Información | Contacto | Inscripción | Usuarios |
Información

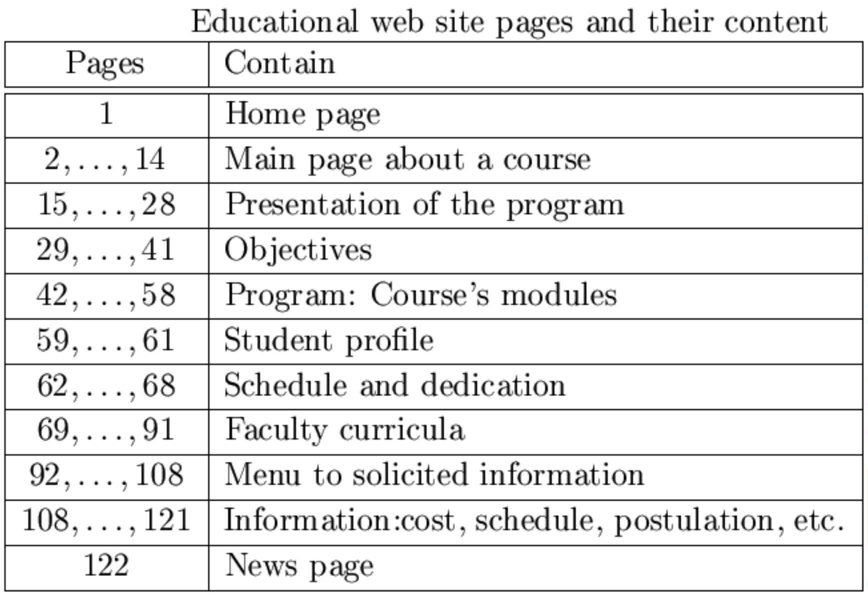
Antecedentes Generales

- **Fecha de Inicio:** 3 Agosto, 2002
- **Horarios:** El programa se desarrolla en 276 horas cronológicas.
- **Valor del Programa:** UF 220; Código SENCE en trámite.
- **Postulaciones y Matrículas:** A partir del 24 de Junio de 2002
- **Documentos de Postulación:**
  - Fotocopia del Certificado de Título

Educational web site: old site layout

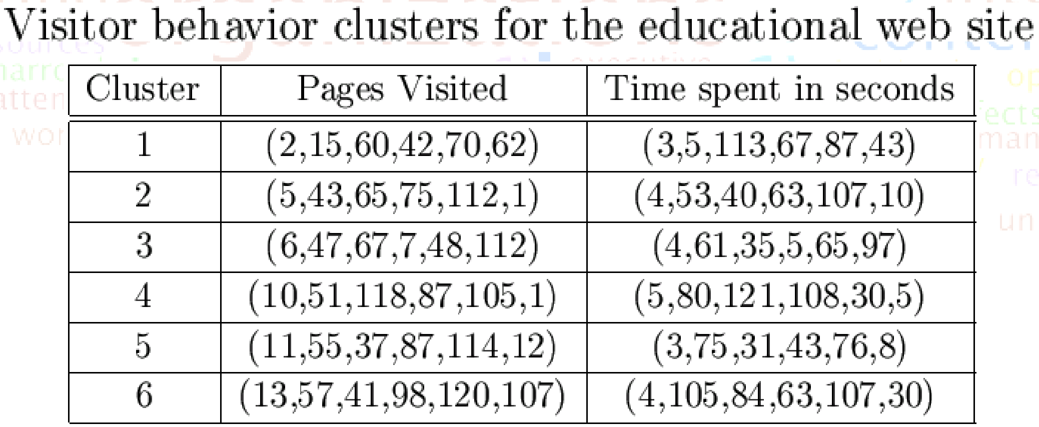


Educational web site: results



Educational web site pages and their content

Pages	Contain
1	Home page
2, . . . , 14	Main page about a course
15, . . . , 28	Presentation of the program
29, . . . , 41	Objectives
42, . . . , 58	Program: Course's modules
59, . . . , 61	Student profile
62, . . . , 68	Schedule and dedication
69, . . . , 91	Faculty curricula
92, . . . , 108	Menu to solicited information
108, . . . , 121	Information:cost, schedule, postulation, etc.
122	News page

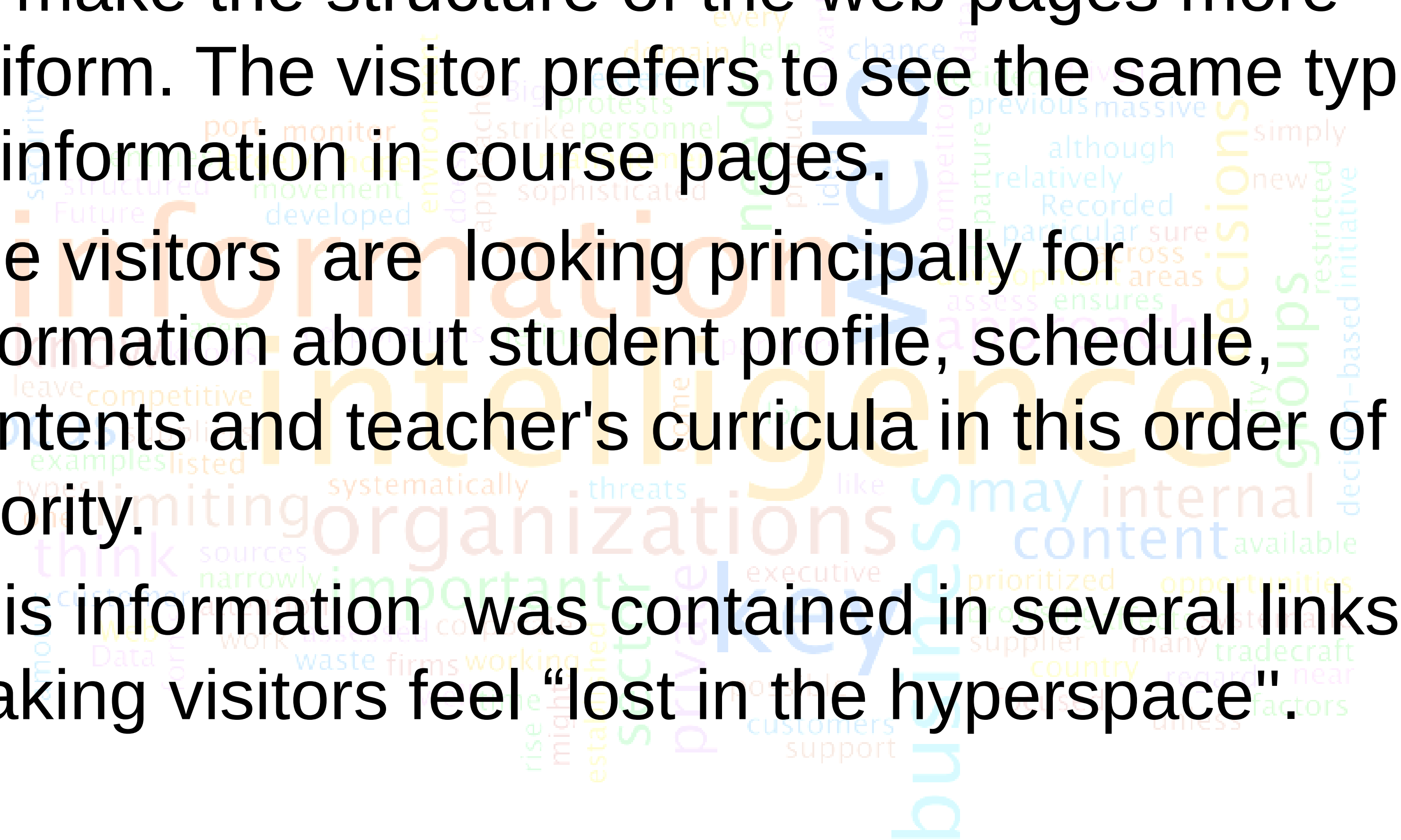


Visitor behavior clusters for the educational web site

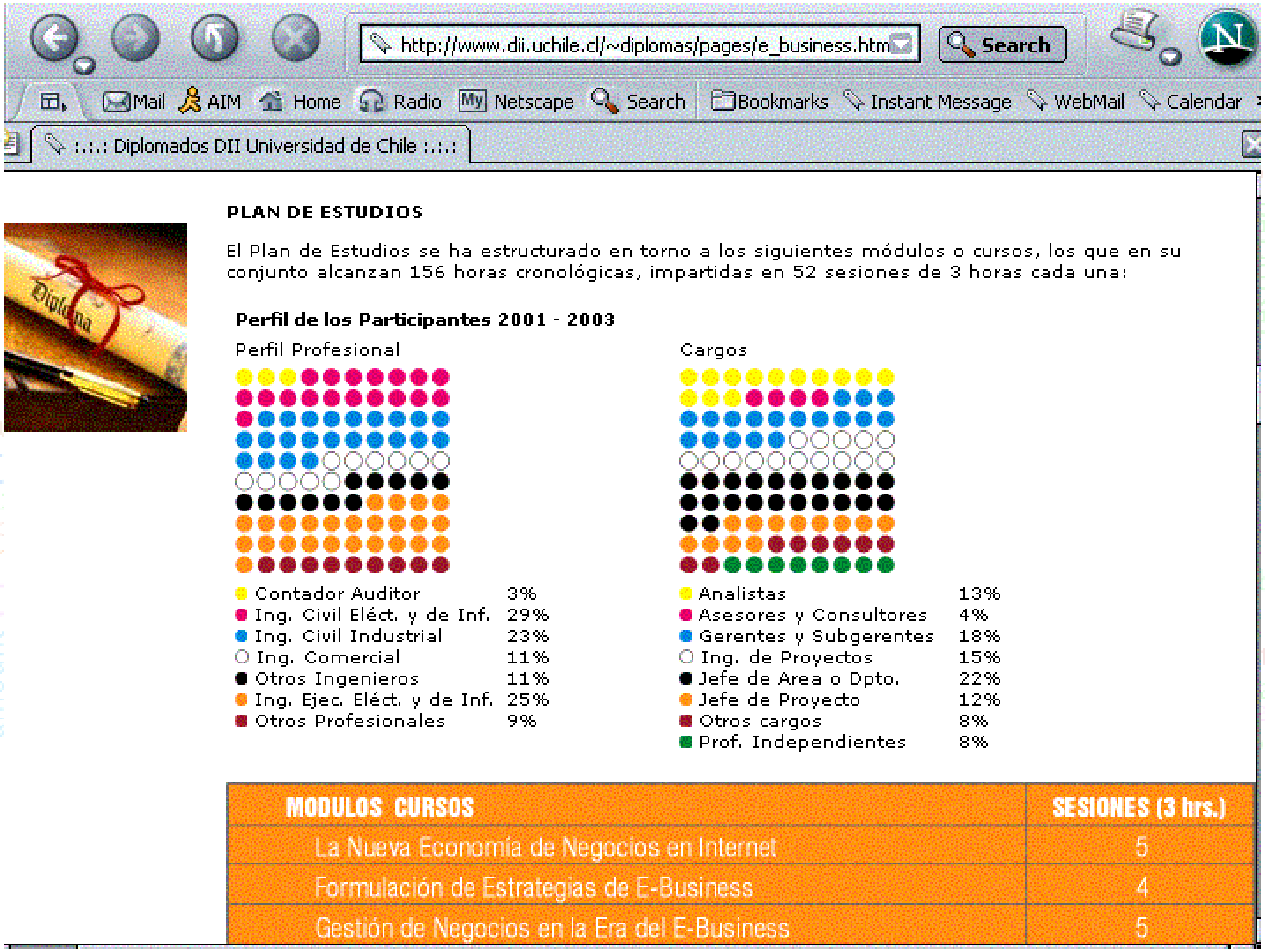
Cluster	Pages Visited	Time spent in seconds
1	(2,15,60,42,70,62)	(3,5,113,67,87,43)
2	(5,43,65,75,112,1)	(4,53,40,63,107,10)
3	(6,47,67,7,48,112)	(4,61,35,5,65,97)
4	(10,51,118,87,105,1)	(5,80,121,108,30,5)
5	(11,55,37,87,114,12)	(3,75,31,43,76,8)
6	(13,57,41,98,120,107)	(4,105,84,63,107,30)

Educational web site: offline recommendations

- To make the structure of the web pages more uniform. The visitor prefers to see the same type of information in course pages.
- The visitors are looking principally for information about student profile, schedule, contents and teacher's curricula in this order of priority.
- This information was contained in several links, making visitors feel “lost in the hyperspace".



Educational web site: new page template


http://www.dii.uchile.cl/~diplomas/pages/e_business.htm
Search
Mail
AIM
Home
Radio
Netscape
Search
Bookmarks
Instant Message
WebMail
Calendar
:.:.: Diplomados DII Universidad de Chile :.:.:
Diploma

PLAN DE ESTUDIOS

El Plan de Estudios se ha estructurado en torno a los siguientes módulos o cursos, los que en su conjunto alcanzan 156 horas cronológicas, impartidas en 52 sesiones de 3 horas cada una:

Perfil de los Participantes 2001 - 2003

Perfil Profesional

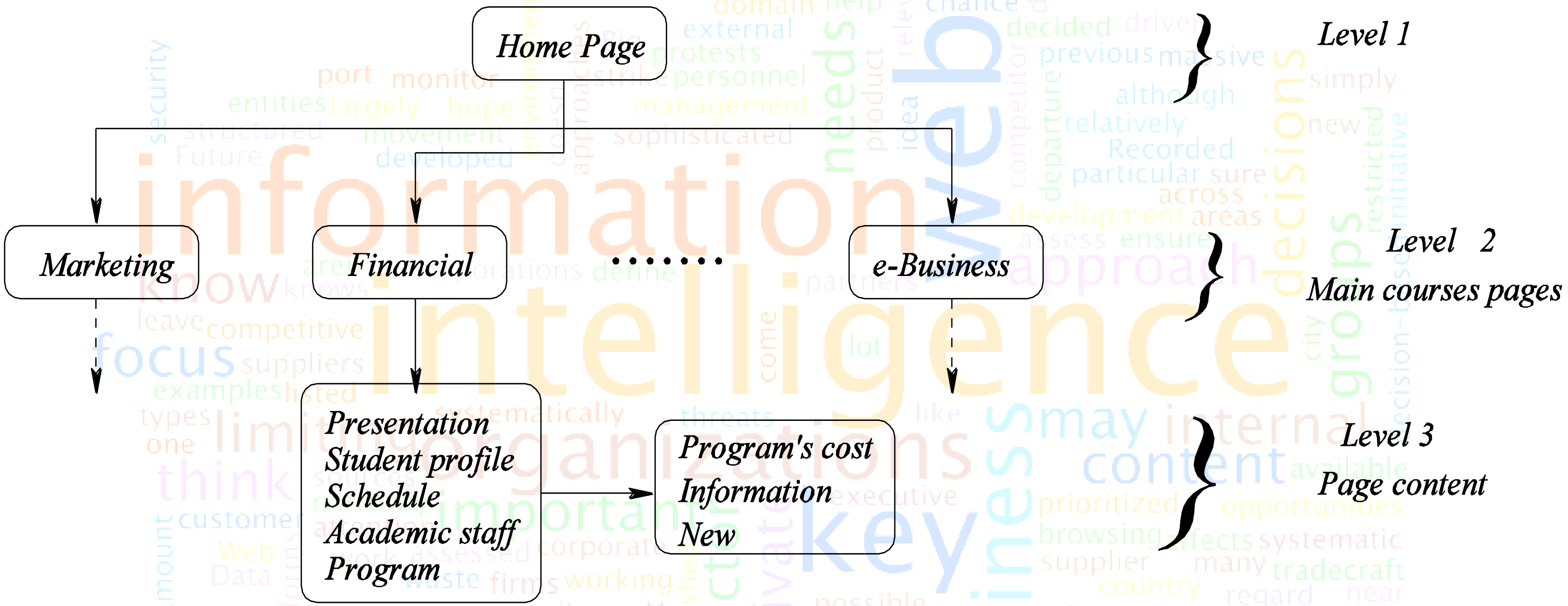
🟡 Contador Auditor	3%
🔴 Ing. Civil Eléct. y de Inf.	29%
🔵 Ing. Civil Industrial	23%
○ Ing. Comercial	11%
● Otros Ingenieros	11%
🟠 Ing. Ejec. Eléct. y de Inf.	25%
🟥 Otros Profesionales	9%

Cargos

🟡 Analistas	13%
🔴 Asesores y Consultores	4%
🔵 Gerentes y Subgerentes	18%
○ Ing. de Proyectos	15%
● Jefe de Area o Dpto.	22%
🟠 Jefe de Proyecto	12%
🟥 Otros cargos	8%
🟩 Prof. Independientes	8%

MODULOS CURSOS	SESIONES (3 hrs.)
La Nueva Economía de Negocios en Internet	5
Formulación de Estrategias de E-Business	4
Gestión de Negocios en la Era del E-Business	5

Educational web site: new site layout



Home Page

Marketing

Financial

. . . .

e-Business

Presentation
Student profile
Schedule
Academic staff
Program

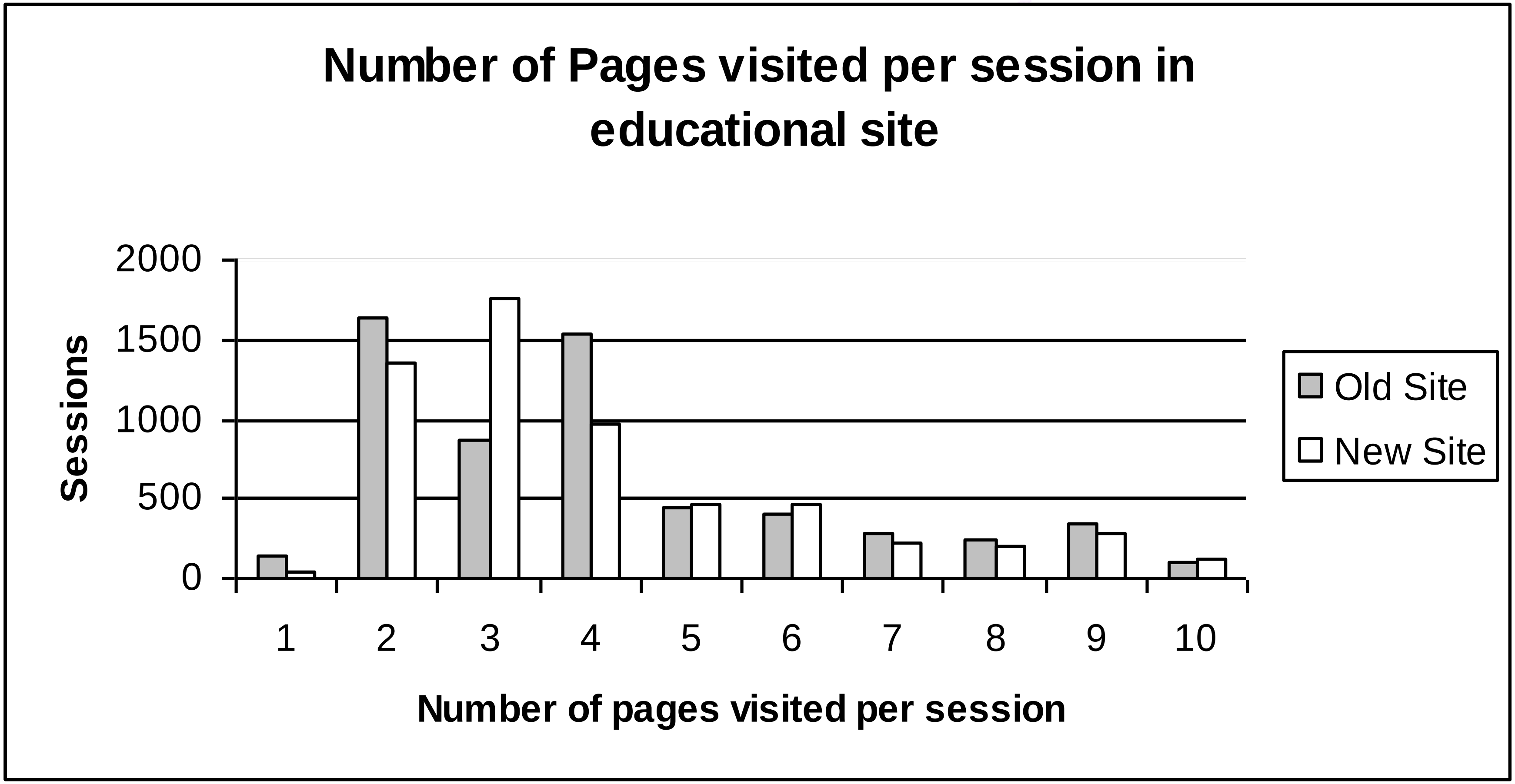
Program's cost
Information
New

Level 1

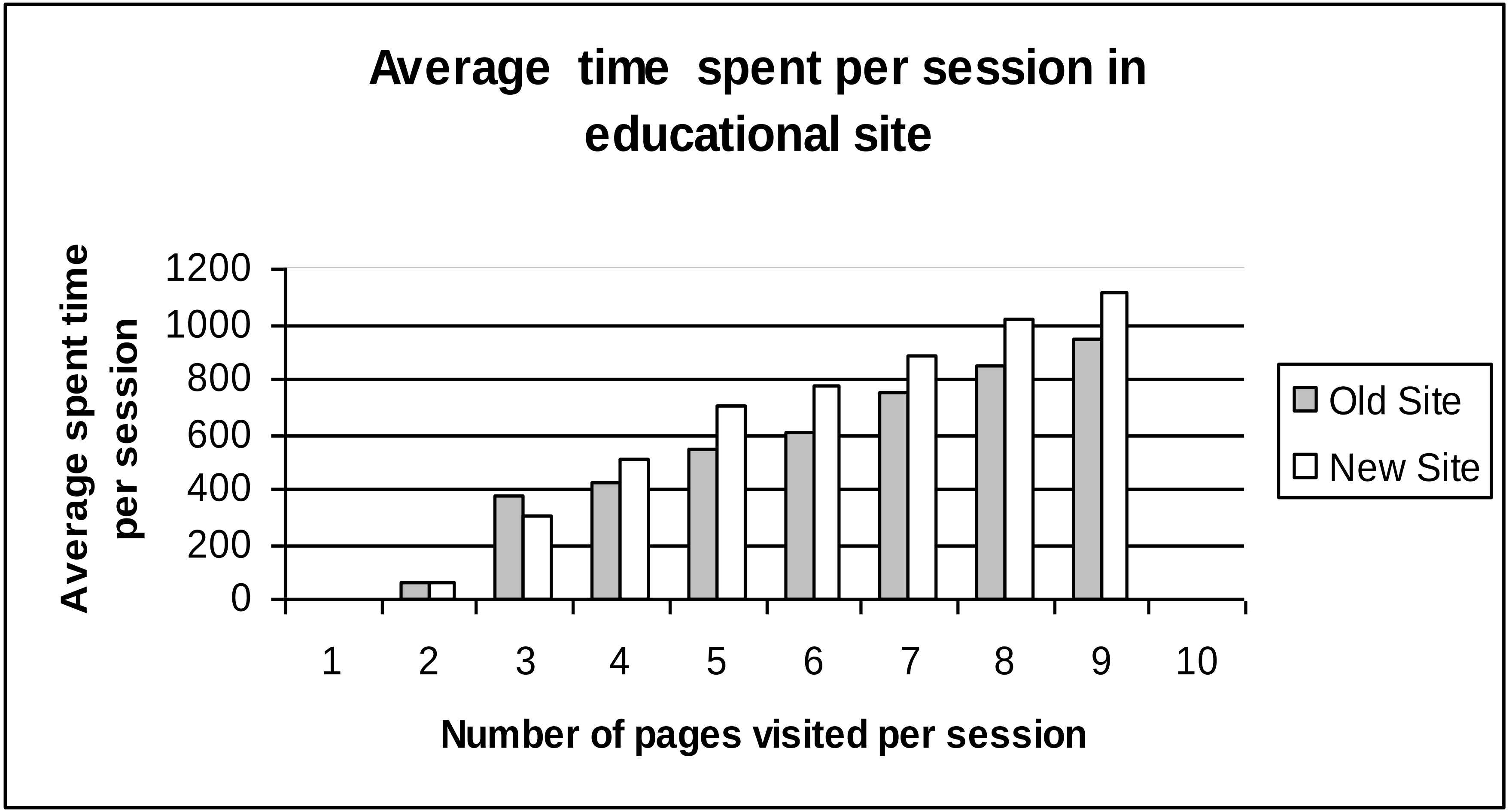
Level 2
Main courses pages

Level 3
Page content

Educational web site: impact of modifications



Educational web site: impact (2)

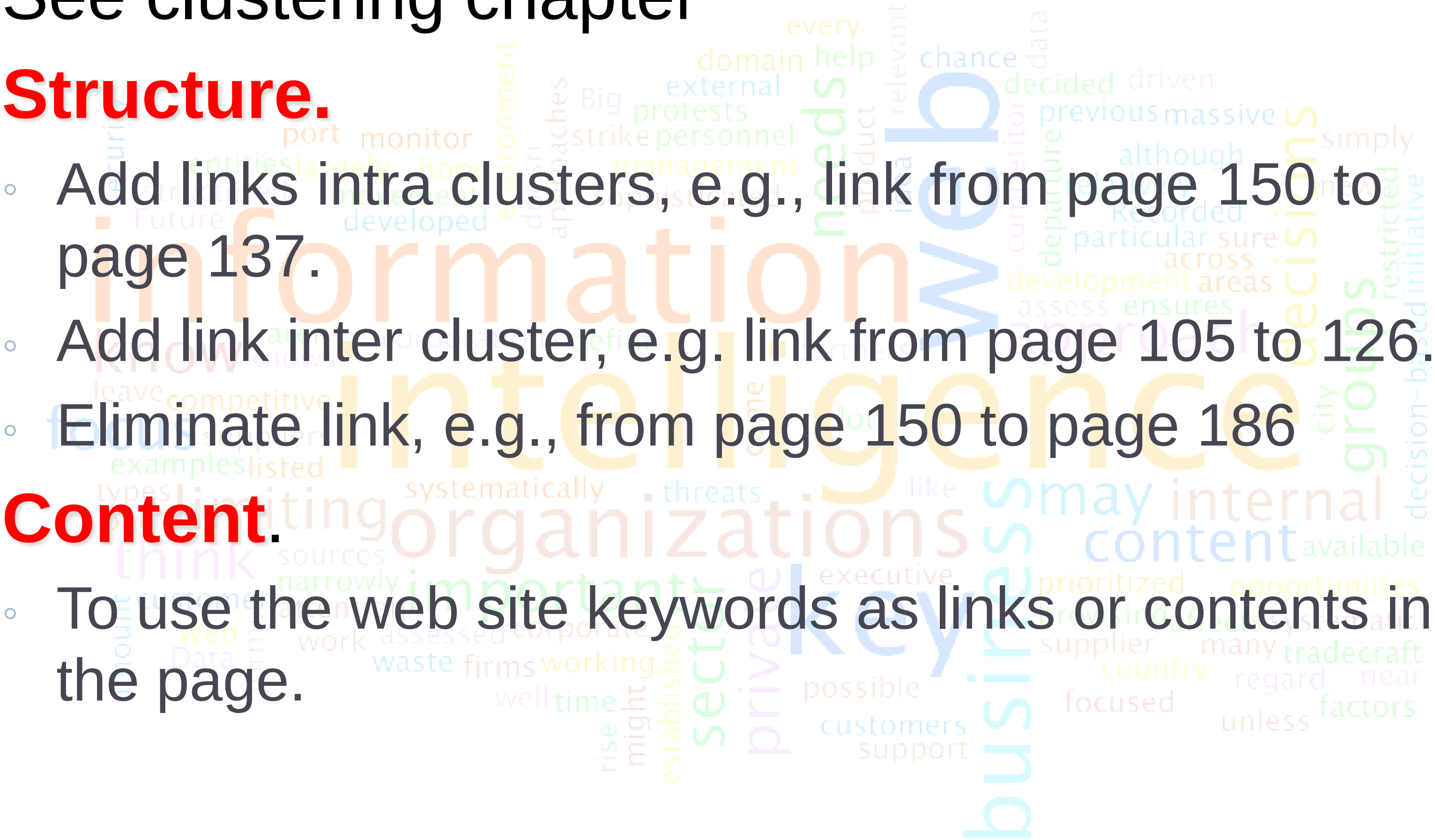


Bank web site

- It belongs to a Chilean virtual bank, i.e., a bank that doesn't have physical branches and where all the transactions are made using electronic means, like e-mails, portals, etc.
- Written in Spanish.
- 217 static web pages.
- Approximately eight million raw web logs registers corresponding to the period January to March, 2003.

Offline recommendations

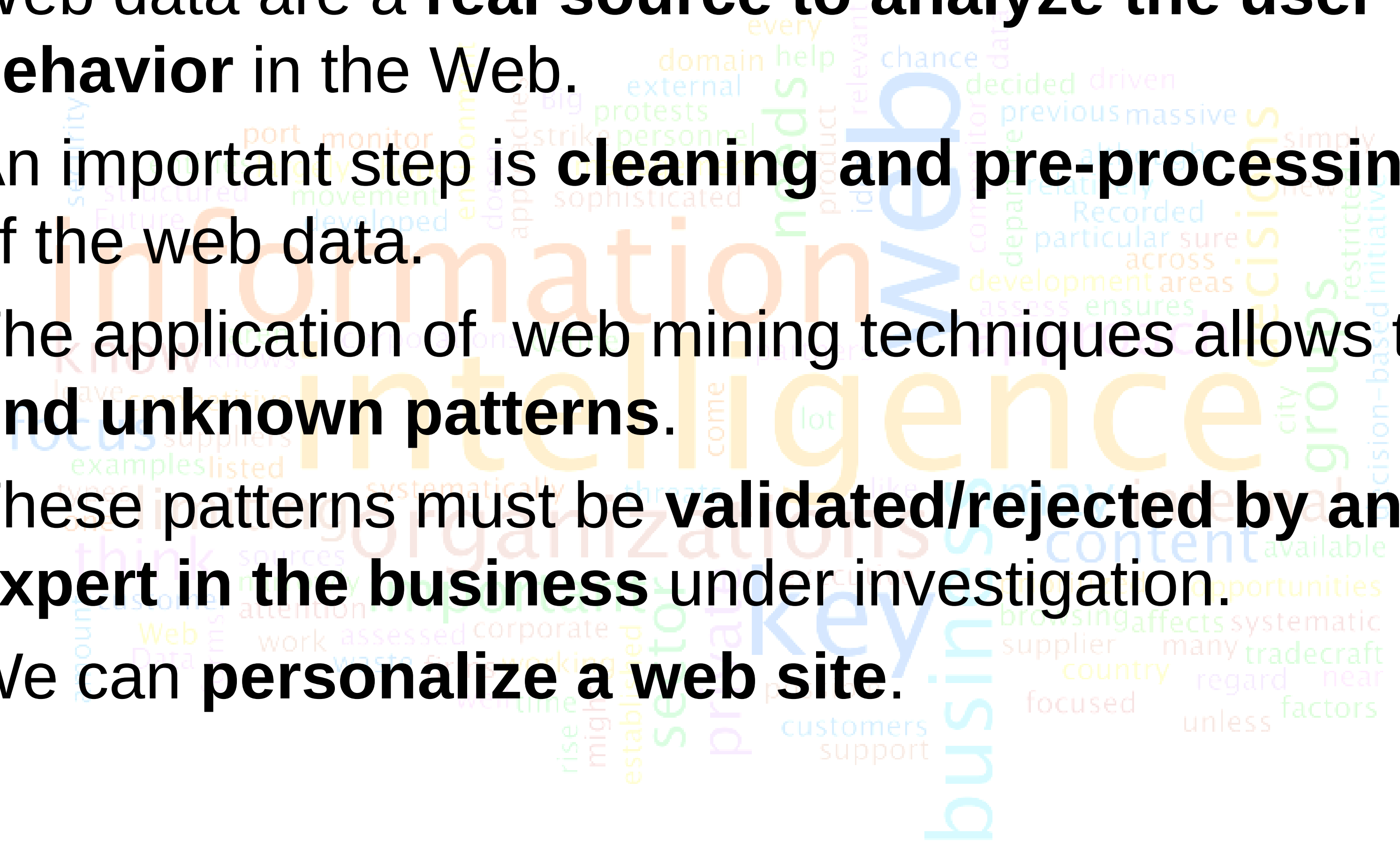
- See clustering chapter
- **Structure.**
  - Add links intra clusters, e.g., link from page 150 to page 137.
  - Add link inter cluster, e.g. link from page 105 to 126.
  - Eliminate link, e.g., from page 150 to page 186
- **Content.**
  - To use the web site keywords as links or contents in the page.



Online recommendations

- A current user session is classified into some clusters found.
- The online navigation recommendation is created as a set of links to pages belonging to the current web site.
- The user can select some links or not.
- **Default case:** *No recommendation.*
- It is too risky to apply the recommendations in the real web site. However it is possible to make a simulation.

Summary



- Web data are a **real source to analyze the user behavior** in the Web.
- An important step is **cleaning and pre-processing** of the web data.
- The application of web mining techniques allows to **find unknown patterns**.
- These patterns must be **validated/rejected by an expert in the business** under investigation.
- We can **personalize a web site**.