Applications

UNDERSTANDING VISITOR BEHAVIOR

The Dream

Transform the visitors into customers and retain the existing ones

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.

Improving the relationship between the web site and the user

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

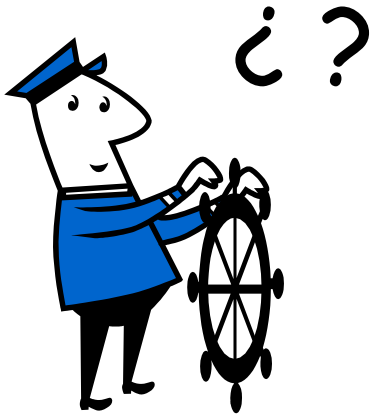
Intelligent Web site [Coenen00, Perkowitz98, Velasquez04b]

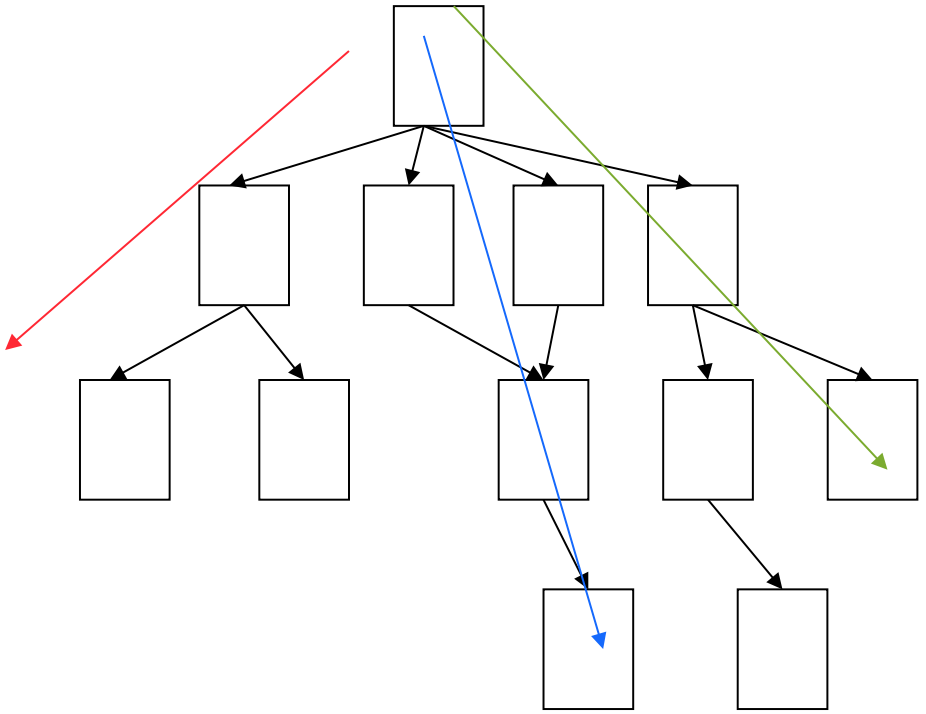
- It represents the main concepts behind the "new portal generation".
- They are systems that *"based on the user behavior, allow to implement changes in the current web site structure and content".*

What we should do? [Mombasher01]

- Data and information abut the visitor behavior ina Web site.
- 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





Visitor browsing behavior (2)

- Three variables are considered: the web page path, its content and the time spent when it is visited by a visitor.
- The visitor behavior vector is defined and a similarity measure between visitor session is introduced.

$$\vec{v} = [(p_1, t_1)...(p_n, t_n)]$$

- Where (p_i, t_i) is a component that represents the page content, its path and percentage of time spent in the i^{th} page visited.
- The vector maintains the page visit order.

Vector space model: A variation proposed

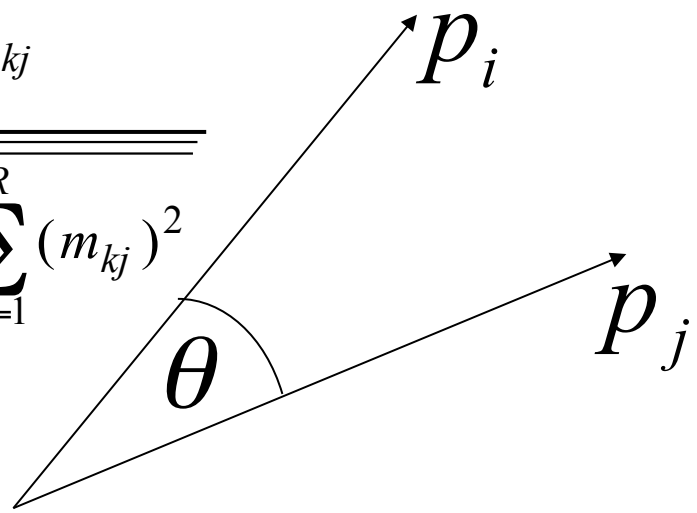
$$M = (m_{ij}) = f_{ij}(1 + \frac{sw_i}{TR}) * \log(\frac{Q}{n_i})$$

sw: special words array

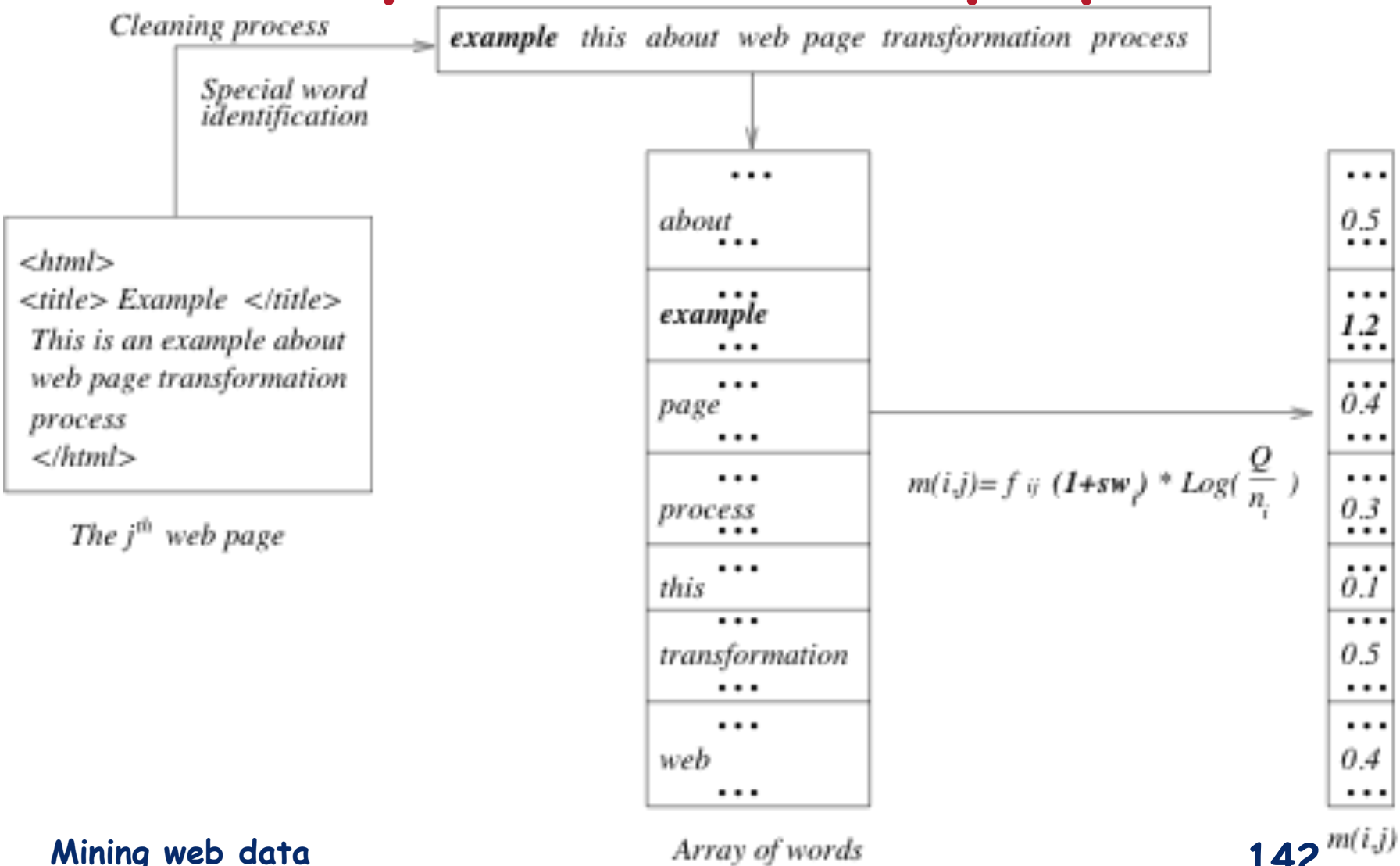
TR: Total special words

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

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

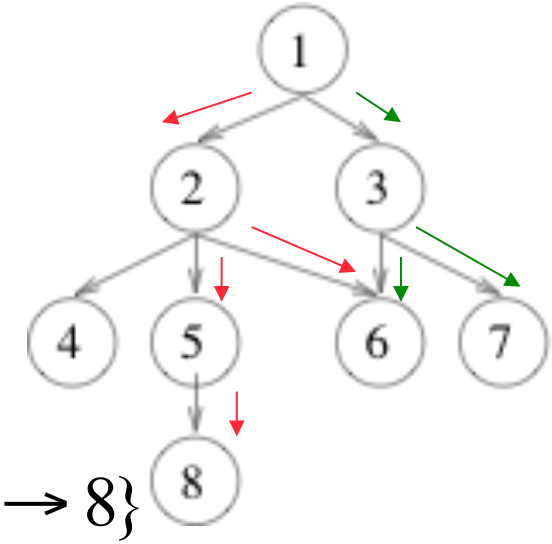


An example of variation proposed



Comparing the sequences [Runkler03, Velasquez04a]

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



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

$S_2 = (1,3,6,7)$

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

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

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

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

- 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"

$L(S_1, S_2) = 3$

$L(S_3, S_4) = 4$

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

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

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

Comparing browsing behavior

- Then the similarity measure is:

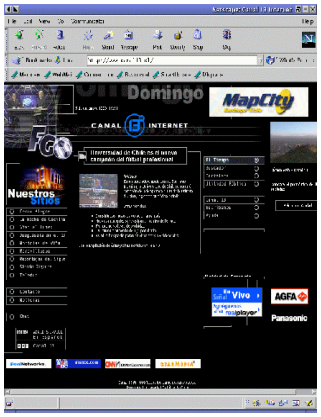
$$sm(\vec{\alpha},\vec{\beta}) = dG(p_\alpha^h, p_\beta^h)\frac{1}{L}\sum_{k=1}^{L}\min(\frac{t_k^\alpha}{t_k^\beta},\frac{t_k^\beta}{t_k^\alpha})dp(p_{\alpha,k}^c, p_{\beta,k}^c)$$

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

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

and *dG* is a "graph distance", i.e., how similar are the paths 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

Visitor preferences

- The main focus is to understand the text preferences.
- Web site keywords: the word or the set of words that make the web page more attractive for the visitor.
- From the visitor behavior vector, we want to select the most important pages, assuming that the degree of importance is correlated with the percentage of time spent on each page.

Important pages vector [Velasquez05b]

- We can suppose that the most important content for the visitors is included in the web pages that concentrate the greater amount of time spent in their visit.
- Then, in order to know the most important web page content, we can select a component set from v. Let be ι the quantity selected.

$$\upsilon_\iota(v) = [(\rho_1, \tau_1), \ldots, (\rho_\iota, \tau_\iota)]$$

- (ρ_k, τ_k) being the component in v that represents the k^{th} page most important and the time spent on it.
- This information is obtained after sorting the v components by time and select the first ι ones.

Similarity measure using text

$$st(\upsilon_\iota(\alpha),\upsilon_\iota(\beta)) = \frac{1}{\iota}\sum_{k=1}^{\iota} \min\{\frac{\tau_k^\alpha}{\tau_k^\beta}, \frac{\tau_k^\beta}{\tau_k^\alpha}\} * dp(\rho_k^\alpha, \rho_k^\beta)$$

- This equation combines the content of the most important visited pages with the time spent on each one of them by a multiplication.
- In this way we can distinguish between two pages with similar content, but different spent times.

Identifying web site keywords

- A clustering algorithm can be applied to find groups of similar visitor sessions.
- Based on the clusters centroids, the most important words for each cluster will be identified.
- Using the vector space model and a method to determine the most important keywords and their importance in each cluster is proposed.
- A measure (geometric mean) to calculate the importance of each word relative to each cluster, is defined as:

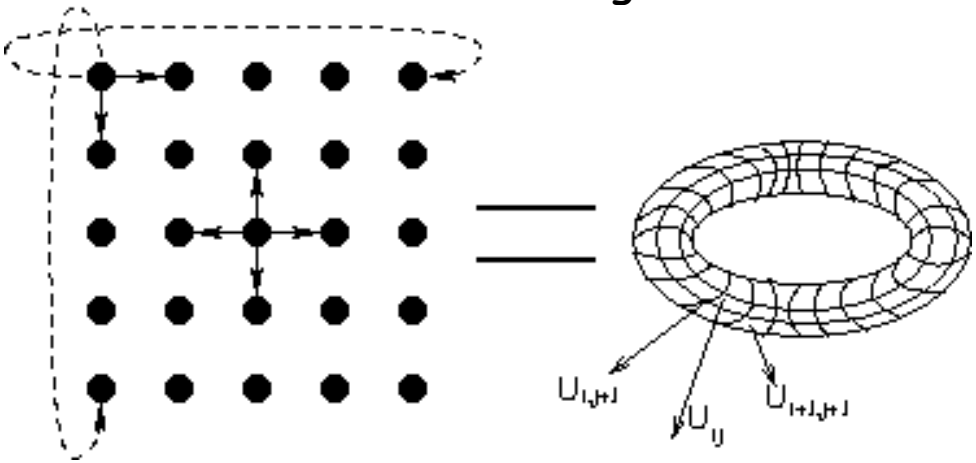
$$kw[i] = \sqrt[\iota]{\prod_{p \in \zeta} m_{ip}}$$

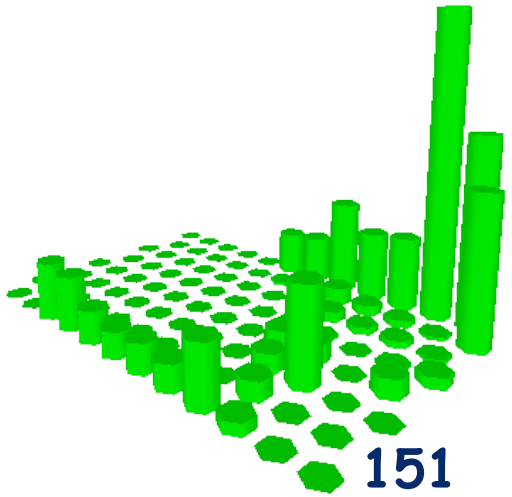
ζ : Cluster with ι vectors

$$i : 1, \ldots, R$$

Applying clustering algorithms [Hinnserbury03, Theodoridis99, Velasquez03

- For example, Self Organizing Feature Maps.
- Schematically, a SOFM is presented as a two-dimensional array in whose positions the neurons are located.
- Each neuron is constituted by an n-dimensional vector, whose components are the synaptic weights.
- The notion of neighborhood among the neurons provides diverse topologies.
- In this case a toroidal topology was used, which means that the neurons closest to the ones of the superior edge, are located in the inferior and lateral edges.





Working with real web sites

Web site selection

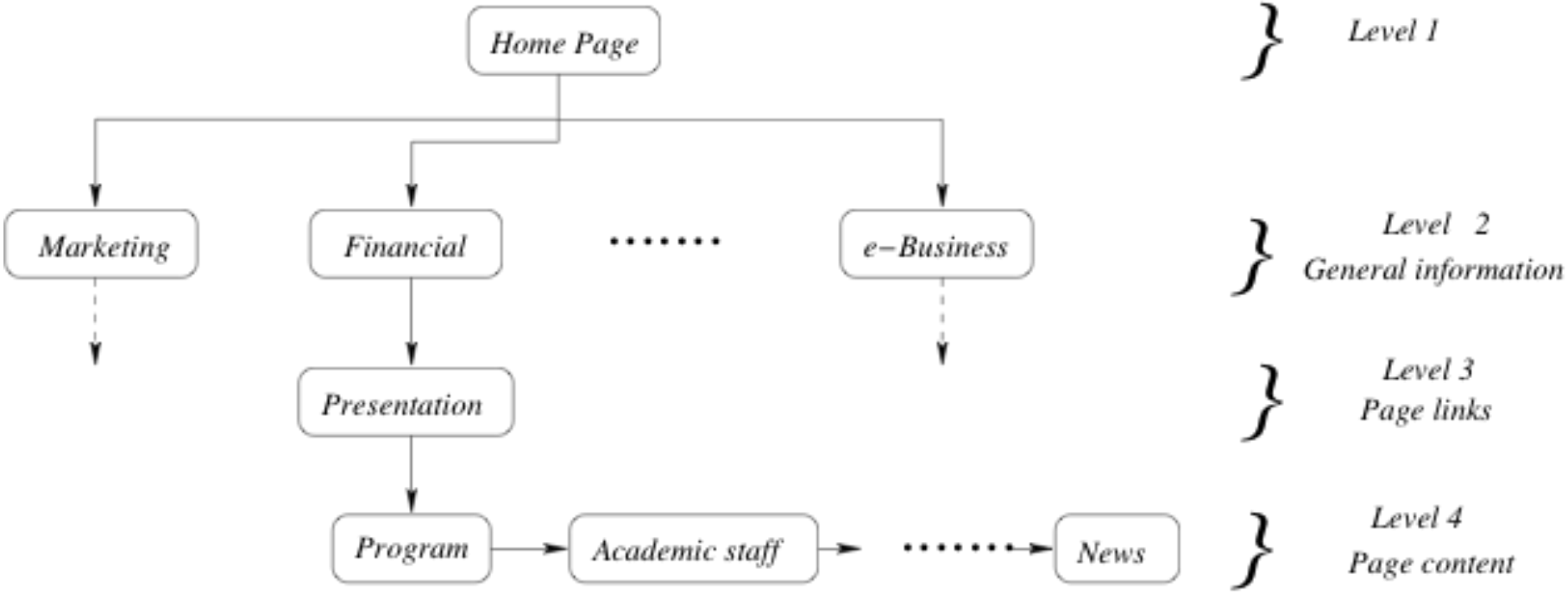
- Education program in the Department of Industrial Engineering at the University of Chile.
- A Chilean virtual bank, during the period of analysis (January to March, 2003) approximately eight million of raw registers were captured.

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

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

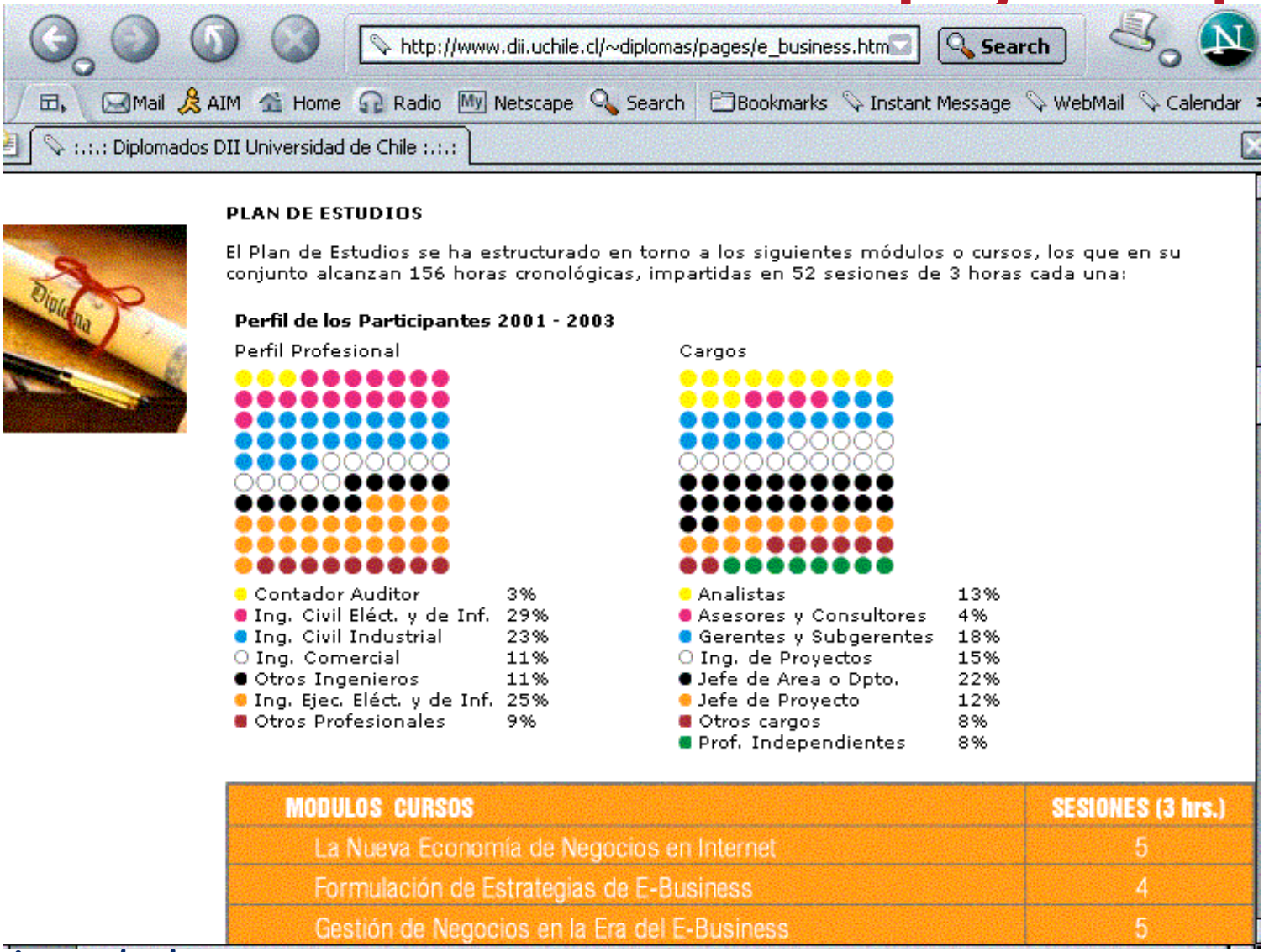
Visitor behavior clusters for the educational web site

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)

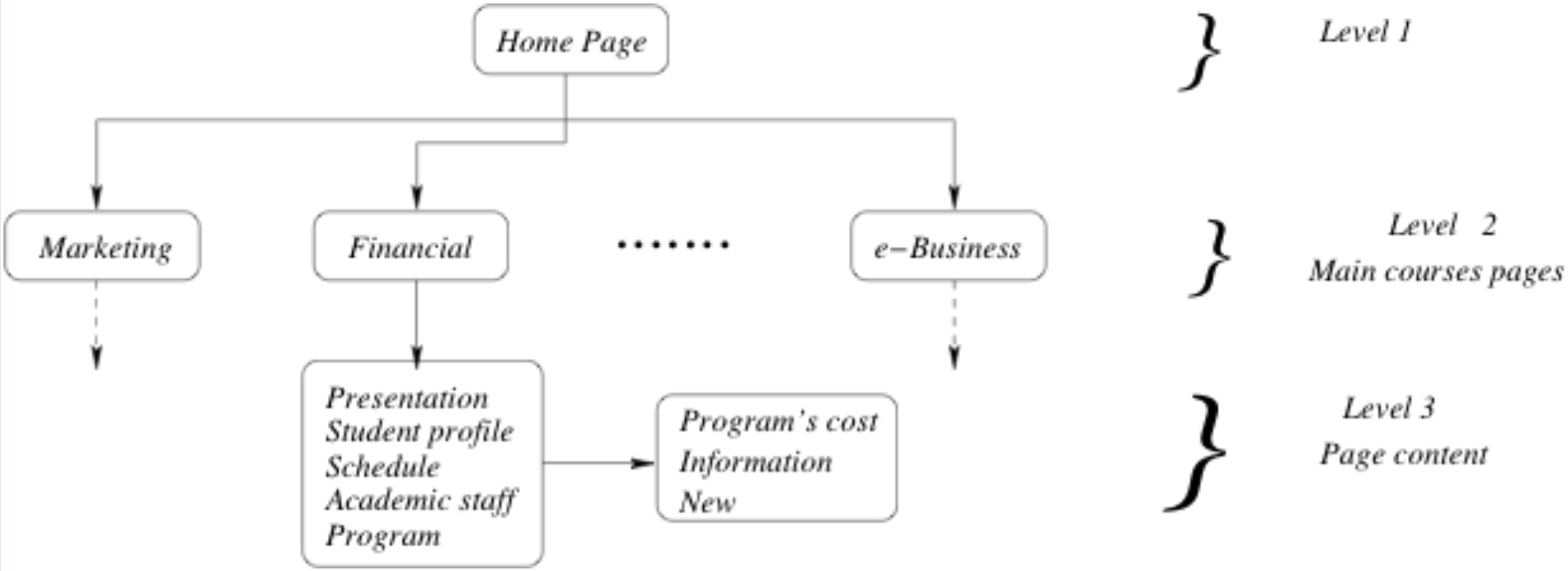
Educational web site: offline recommedations

- 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''.

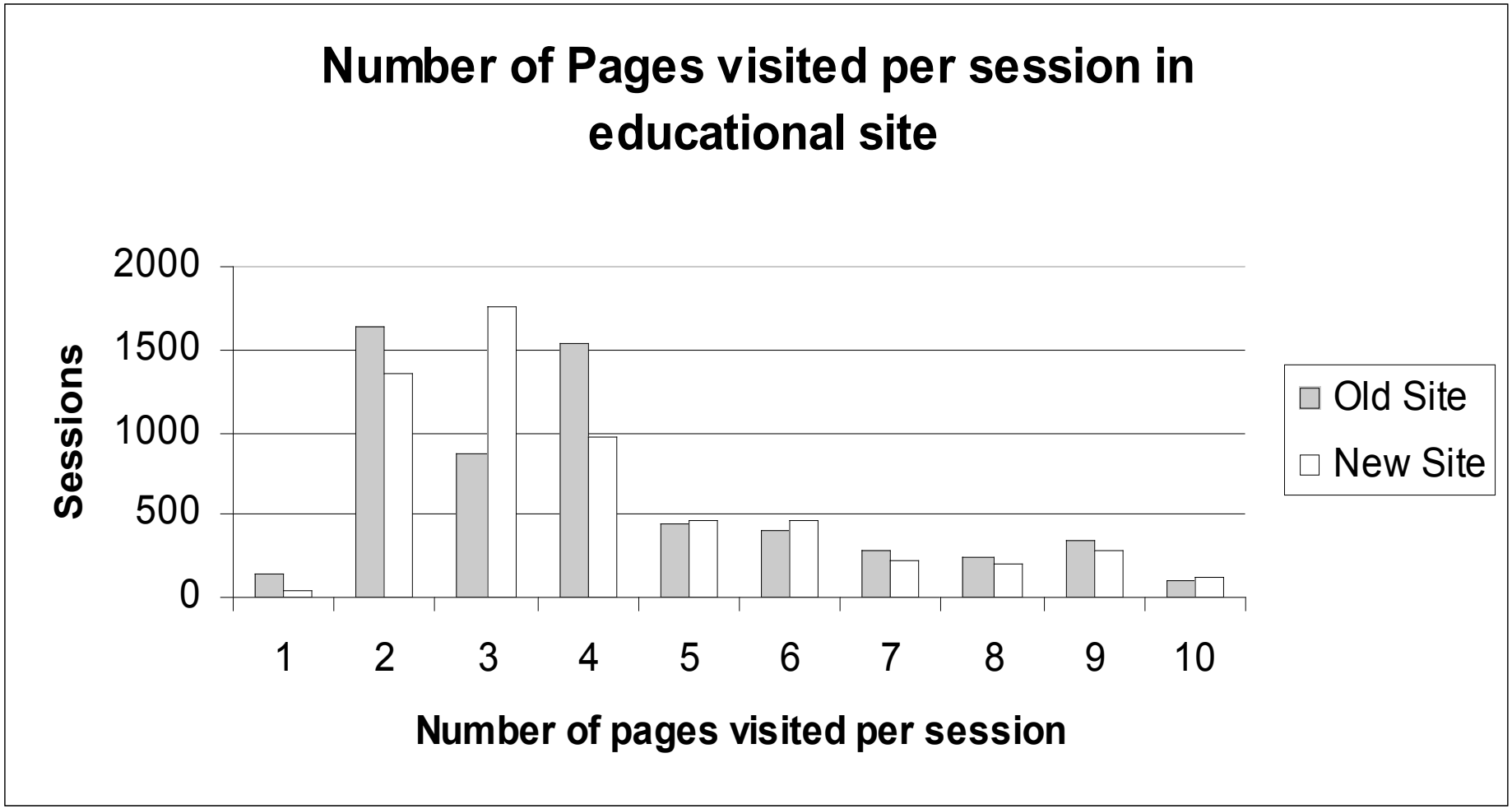
Educational web site: new page template



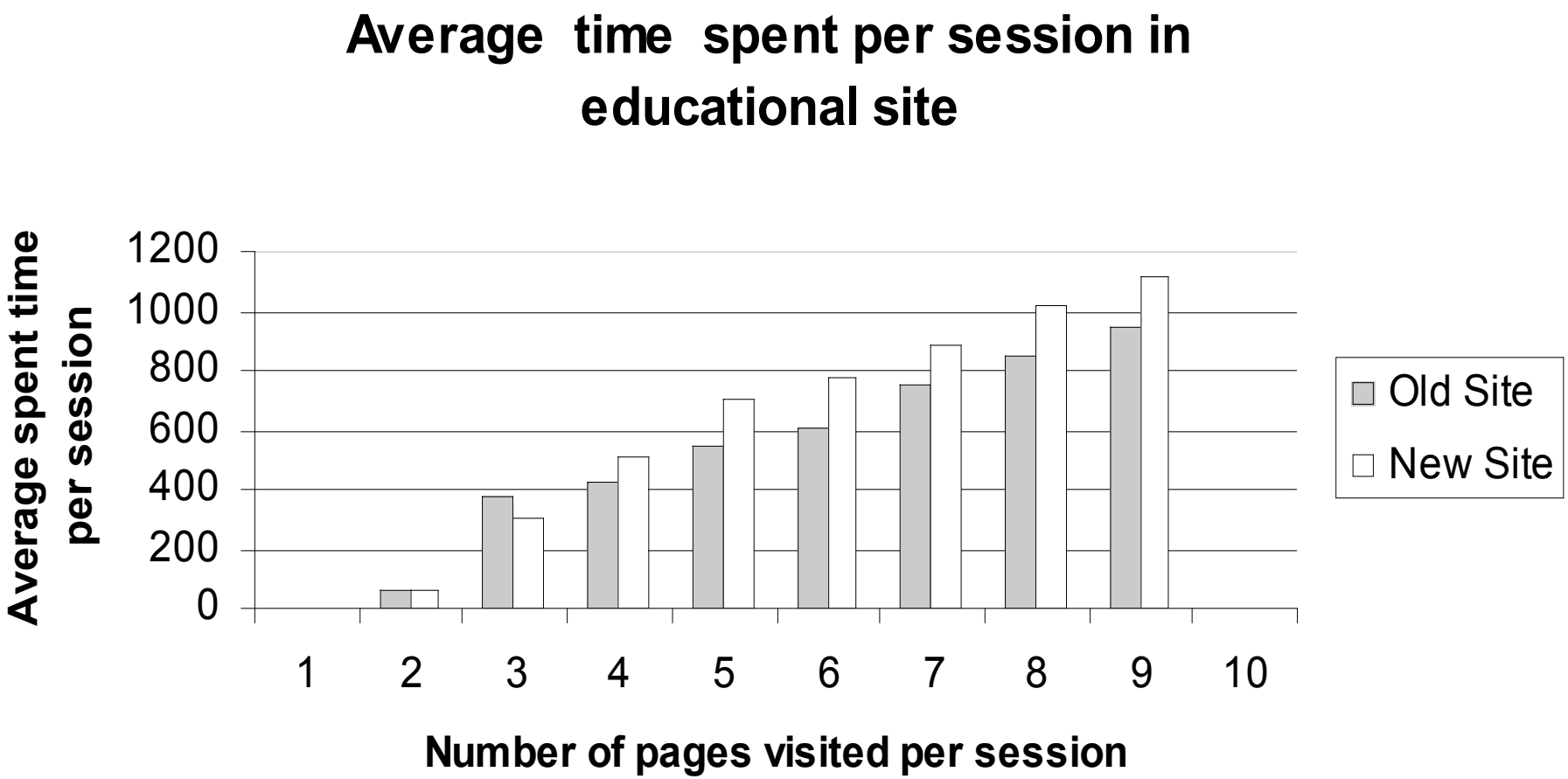
Educational web site: new site layout



Educational web site: impact of modifications



Educational web site: impact (2)



Bank web site

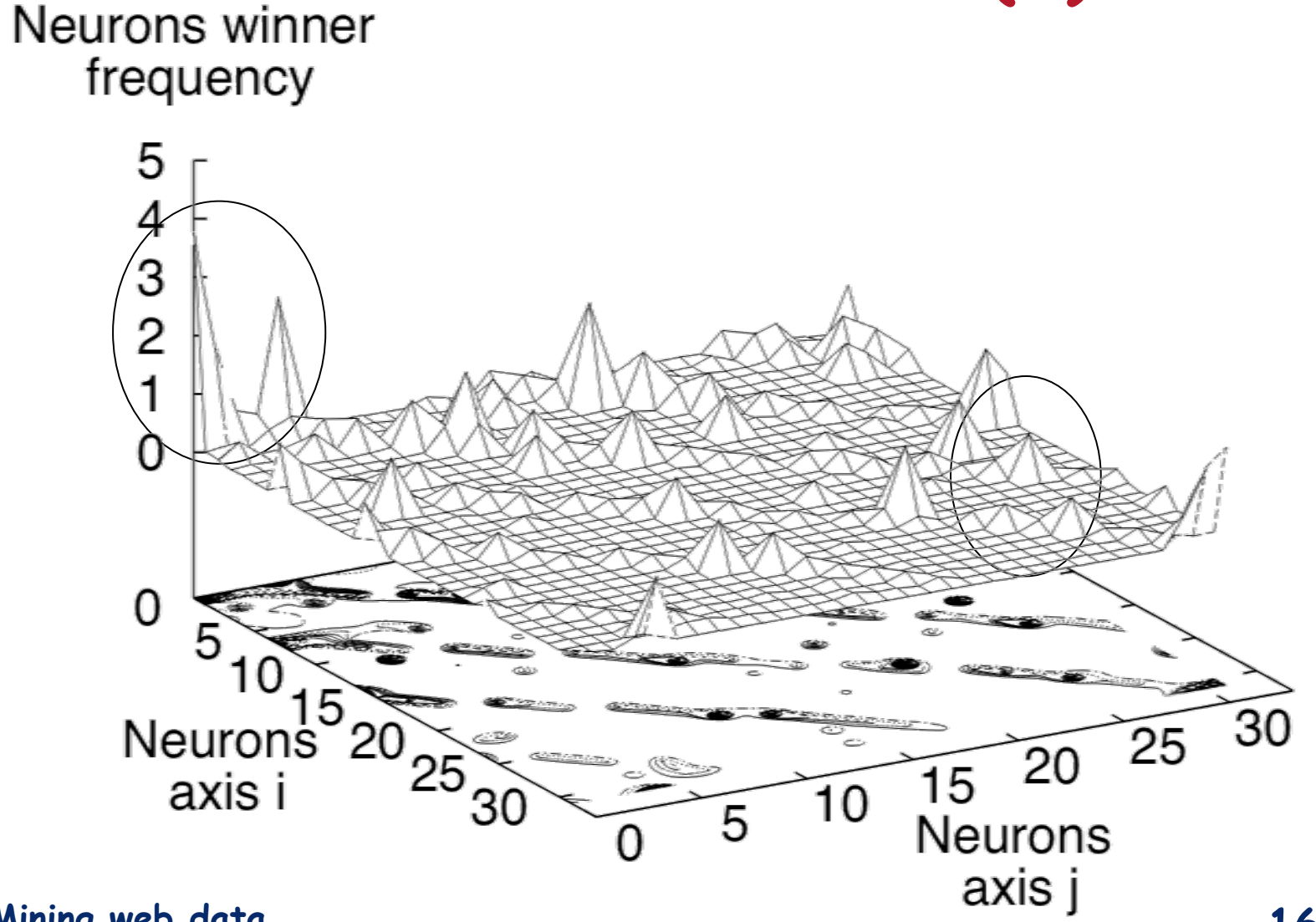
- It belongs to a Chilean virtual bank, i.e., a bank that doesn't have physical branches and where all the transactions are made using electronic means, like e-mails, portals, etc.
- Written in Spanish.
- 217 static web pages.
- Approximately eight million raw web logs registers corresponding to the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

Bank web site home page

Visitor Behavior Vectors

- Only 16% of the visitors visit 10 or more pages and 18% less than 4.
- The average of visited pages is 6.
- With these data, we fixed 6 as the maximum number of components in a visitor behavior vector
- Finally, applying the above described filters, approximately 200,000 vectors were identified.

Visitor Behavior Vectors (2)



Cluster analysis

- The cluster identification and validation is generally a subjective tasks.
- It is necessary the support of a business expert.
- In this case bank business expert.
- Eight clusters were identified but only four of the were accepted by the business expert.
- The criterion is "What cluster make sense for the business expert"

Results

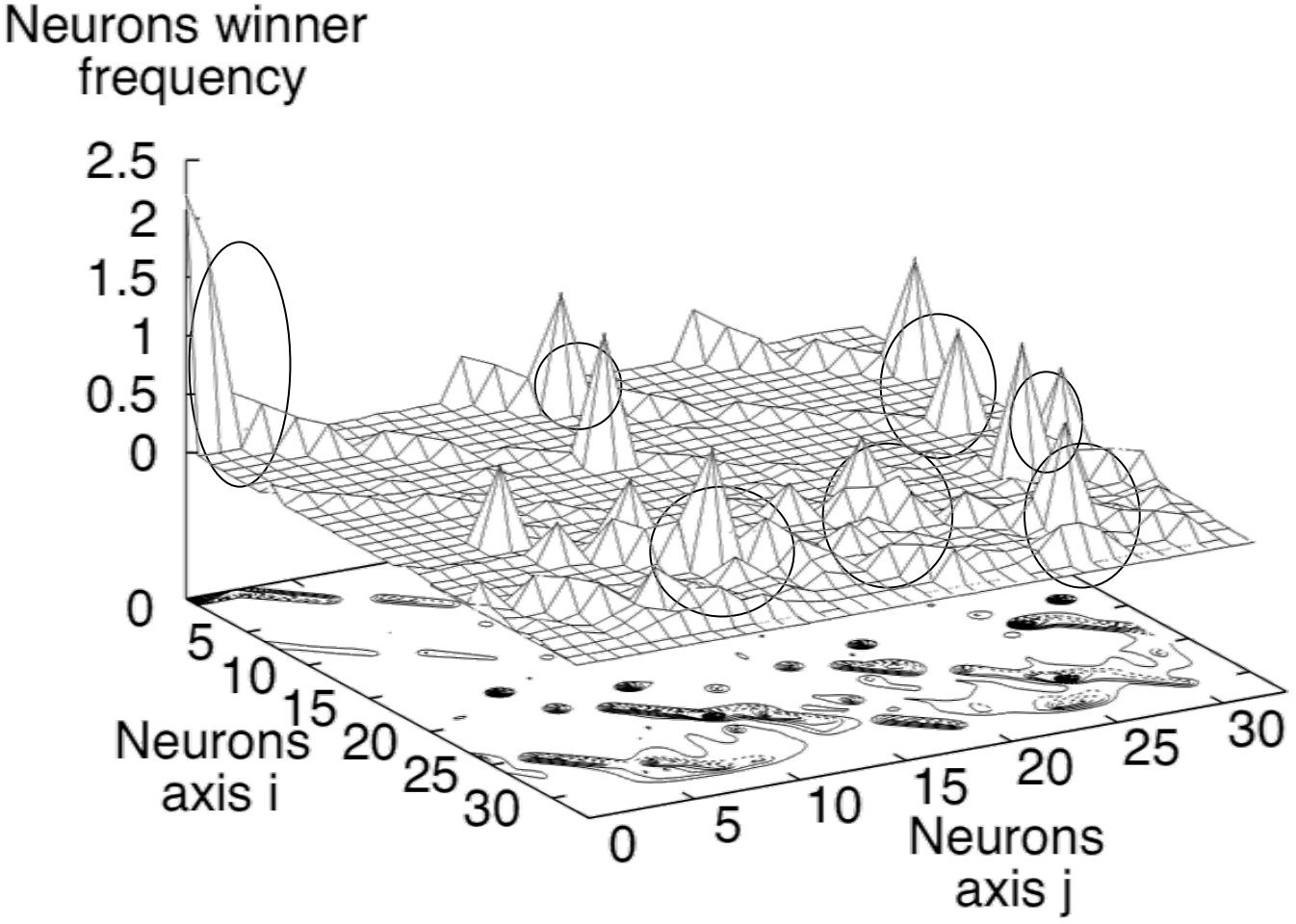
Bank web site pages and their content

Pages	Content
1	Home page
2, . . . , 65	Products and Services
66, . . . , 98	Agreements with other institutions
99, . . . , 115	Remote services
116, . . . , 130	Credit cards
131, . . . , 155	Promotions
156, . . . , 184	Investments
185, . . . , 217	Different kinds of credits

Visitor behavior clusters

Cluster	Pages Visited	Time spent in seconds
1	(1,3,8,9,147,190)	(40,67,175,113,184,43)
2	(100,101,126,128,30,58)	(20,69,40,63,107,10)
3	(70,86,150,186,137,97)	(4,61,35,5,65,97)
4	(157,169,180,101,105,1)	(5,80,121,108,30,5)

Important page clustering



Cluster analysis

- The cluster must content only pages related by main topic.
- The business expert define the main topic per page.
- From twelve identified clusters, only eight were accepted.

Results

Cluster	Pages Visited
1	(6,8,190)
2	(100,128.30)
3	(86,150,97)
4	(101,105,1)
5	(3,9,147)
6	(100,126,58)
7	(70,186,137)
8	(157,169,180)

Cluster analysis

- Review which words in each page cluster are the same.
- Using the vector page definition, it is possible to get the key words and their relative importance in each cluster.

$$M = (m_{ij}) = f_{ij}(1 + \frac{sw_i}{TR}) * \log(\frac{Q}{n_i})$$

$$kw[i] = \sqrt[\iota]{\prod_{p \in \zeta} m_{ip}}$$

- Example

$$\zeta = \{6,8,190\}$$

$$kw_i = \sqrt[3]{m_{i6} m_{i8} m_{i190}}$$

Cluster analysis

Cluster	Keywords	Sort *kw* sorted by weight
1	$(w_8, w_{1254}, w_{64}, w_{878}, w_{238}, w_{126}, w_{3338}, w_{2343})$	(2.51,2.12,1.41,1.22,0.98,0.95,0.9,0.84)
2	$(w_{200}, w_{2321}, w_{206}, w_{205}, w_{2757}, w_{3948}, w_{1746}, w_{1949})$	(2.33,2.22,1.12,1.01,0.93,0.91,0.90,0.89)
3	$(w_{501}, w_{733}, w_{385}, w_{684}, w_{885}, w_{2326}, w_{3434}, w_{1564})$	(2.84,2.32,2.14,1.85,1.58,1.01,0.92,0.84)
4	$(w_{3005}, w_{2048}, w_{505}, w_{3675}, w_{3545}, w_{2556}, w_{2543}, w_{2654})$	(2.72,2.12,1.85,1.52,1.31,0.95,0.84,0.74)
5	$(w_{4003}, w_{449}, w_{895}, w_{867}, w_{2567}, w_{2456}, w_{767}, w_{458})$	(2.54,2.14,1.98,1.58,1.38,1.03,0.91,0.83)
6	$(w_{105}, w_{959}, w_{212}, w_{2345}, w_{3456}, w_{3267}, w_{1876}, w_{384})$	(2.64,2.23,1.84,1.34,1.11,0.97,0.89,0.81)
7	$(w_{345}, w_{156}, w_{387}, w_{387}, w_{458}, w_{789}, w_{1003}, w_{376})$	(2.12,1.87,1.42,1.13,0.95,0.87,0.84,0.78)
8	$(w_{2323}, w_{1233}, w_{287}, w_{4087}, w_{594}, w_{587}, w_{2575}, w_{257})$	(2.35,1.93,1.56,1.32,1.03,0.92,0.83,0.76)

Some identified keywords

#	Keywords	
1	Crédito	Credit
2	Hipotecario	House credit
3	Tarjeta	Card
4	Promoción	Promotion
5	Concurso	Concourse
6	Puntos	Points
7	Descuento	Discount
8	Cuenta	Account

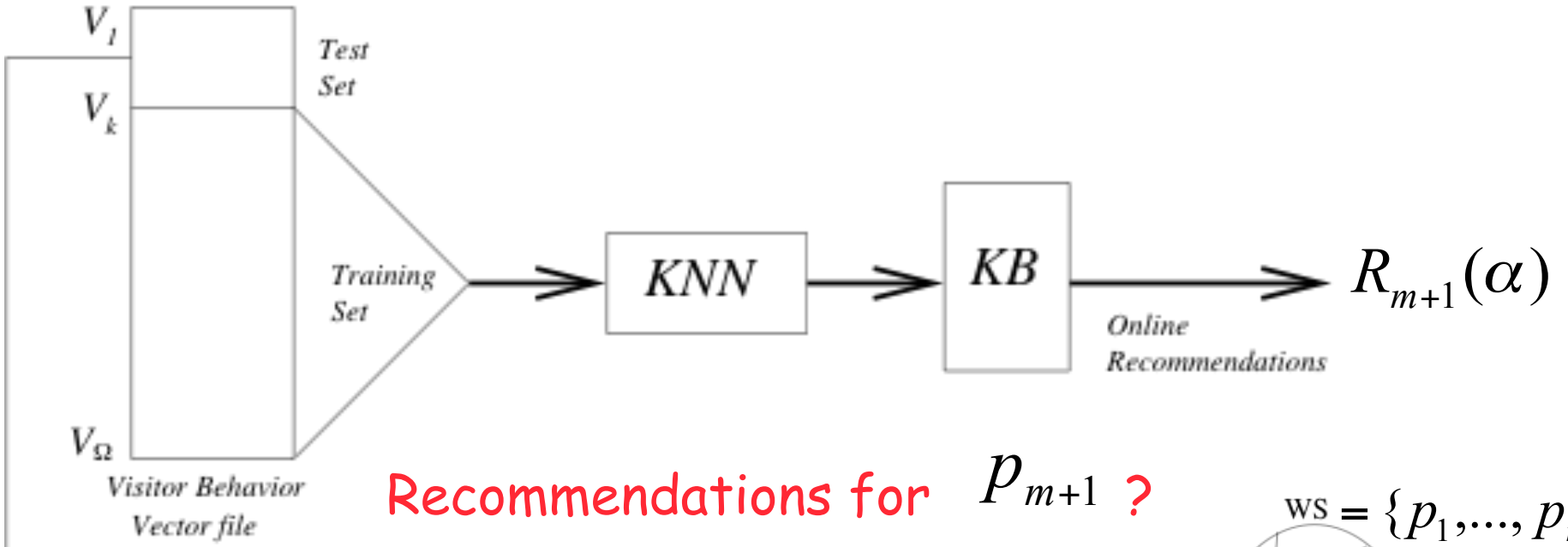
Offline recommendations

- 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.

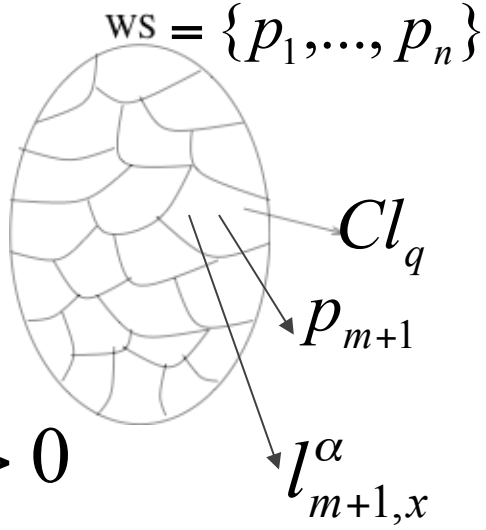
A methodology to test online navigation recommendations



Recommendations for p_{m+1} ?

$$\alpha = [(p_1, t_1), \ldots, (p_m, t_m), \ldots, (p_H, t_H)]$$

$$R_{m+1}(\alpha) = \{l^{\alpha}_{m+1,0}, \ldots, l^{\alpha}_{m+1,k}, \ldots, l^{\alpha}_{m+1,j}\}$$

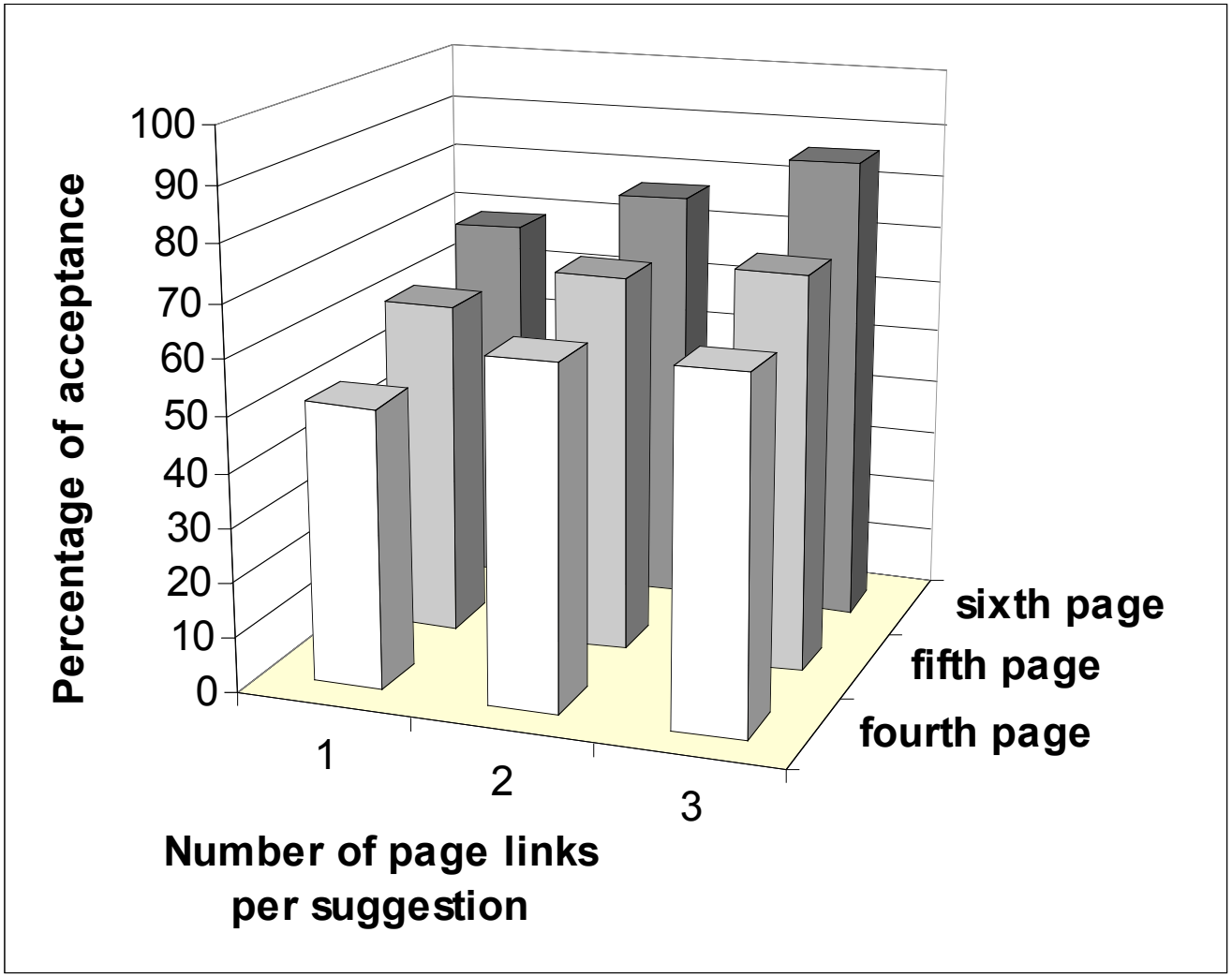


$$p_{m+1} \in Cl_q \quad \exists l^{\alpha}_{m+1,j} \in R^{\alpha}_{m+1} / l^{\alpha}_{m+1,j} \in Cl_q, j > 0$$

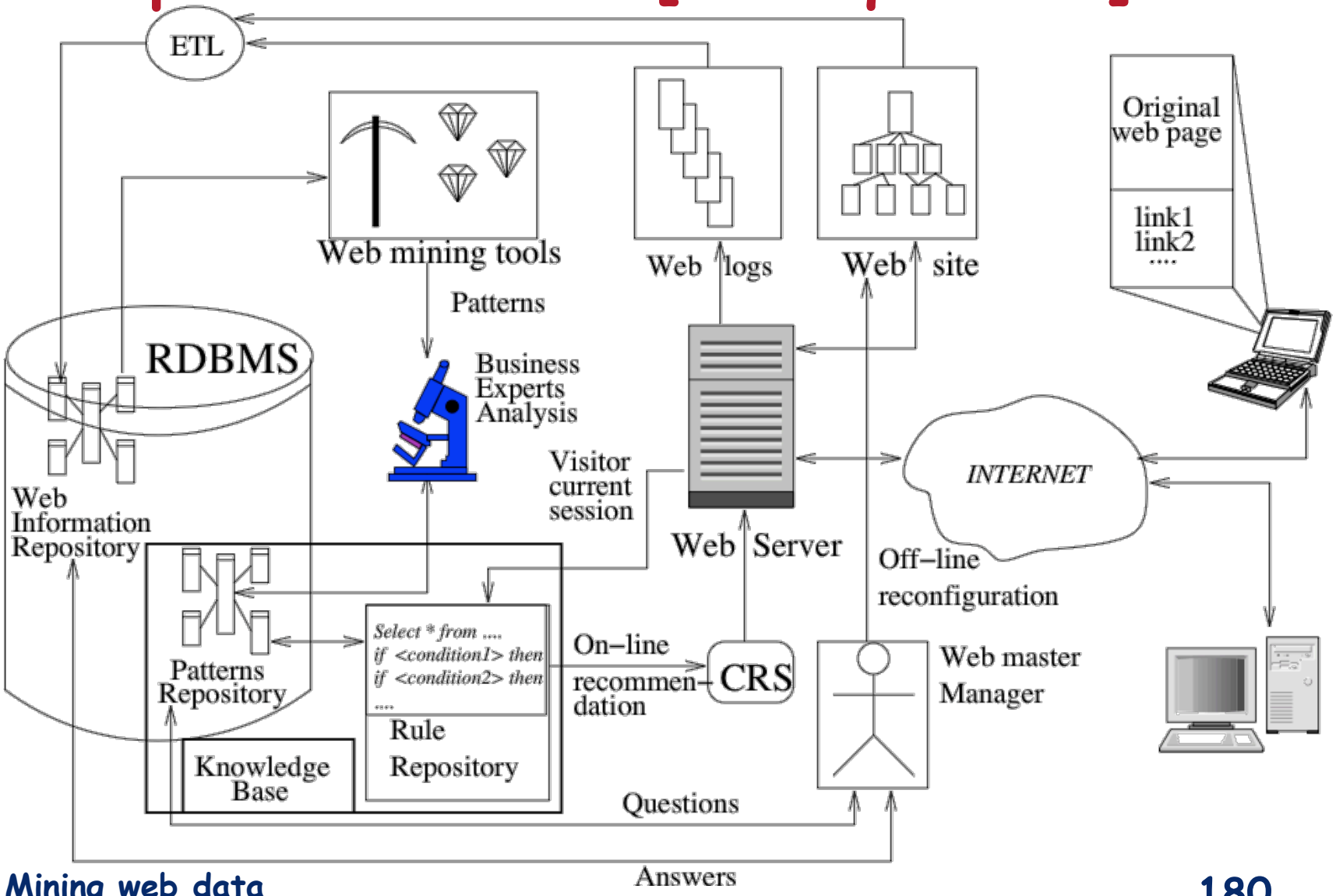
Online recommendations (2)

select navigation, statistics into S
from pr_fact, time, browsing_behavior, wmt where "star join"
and "fix technique" and "fix time" and
sm(pattern,current_visitor) $> \epsilon$;
...
$C_A \rightarrow$ [(78,7),(81,61),(150,43),(193,18),(138,82),(81,93)]
$\alpha \rightarrow [(60, 12), (92, 70), (142, 22)]$ % current visitor
ws $\rightarrow \{p_1, \ldots, p_{217}\}$ % current web site pages
S.navigation $\rightarrow \{p_{200}, p_{121}, p_{187}, p_{128}, p_{212}, p_{112}\}$
S.statistic $\rightarrow \{0.2, 2.2, 1.7, 0.7, 1.2, 0.9\}$
Case C_A and SuggestionPage=4 :
  Prepare_suggestion(p_0,0,L); % default "no suggestion"
% L: link page suggestion, 0: statistic associated
while S not null loop
 if (S.navigation not in ws) then
  S.navigation = compare_page(ws,S.navigation);
 elsif ((S.navigation <> $\alpha_{p_{1\ldots3}}$) and (S.statistic $> \gamma$)) then
  Prepare_suggestion(S.navigation,S.statistic,L);
  Pop(S); % Next element in S
 end if;
end loop;
send(Extracted_Three_Links(L)); % $L \rightarrow \{p_{121}, p_{187}, p_{212}\}$

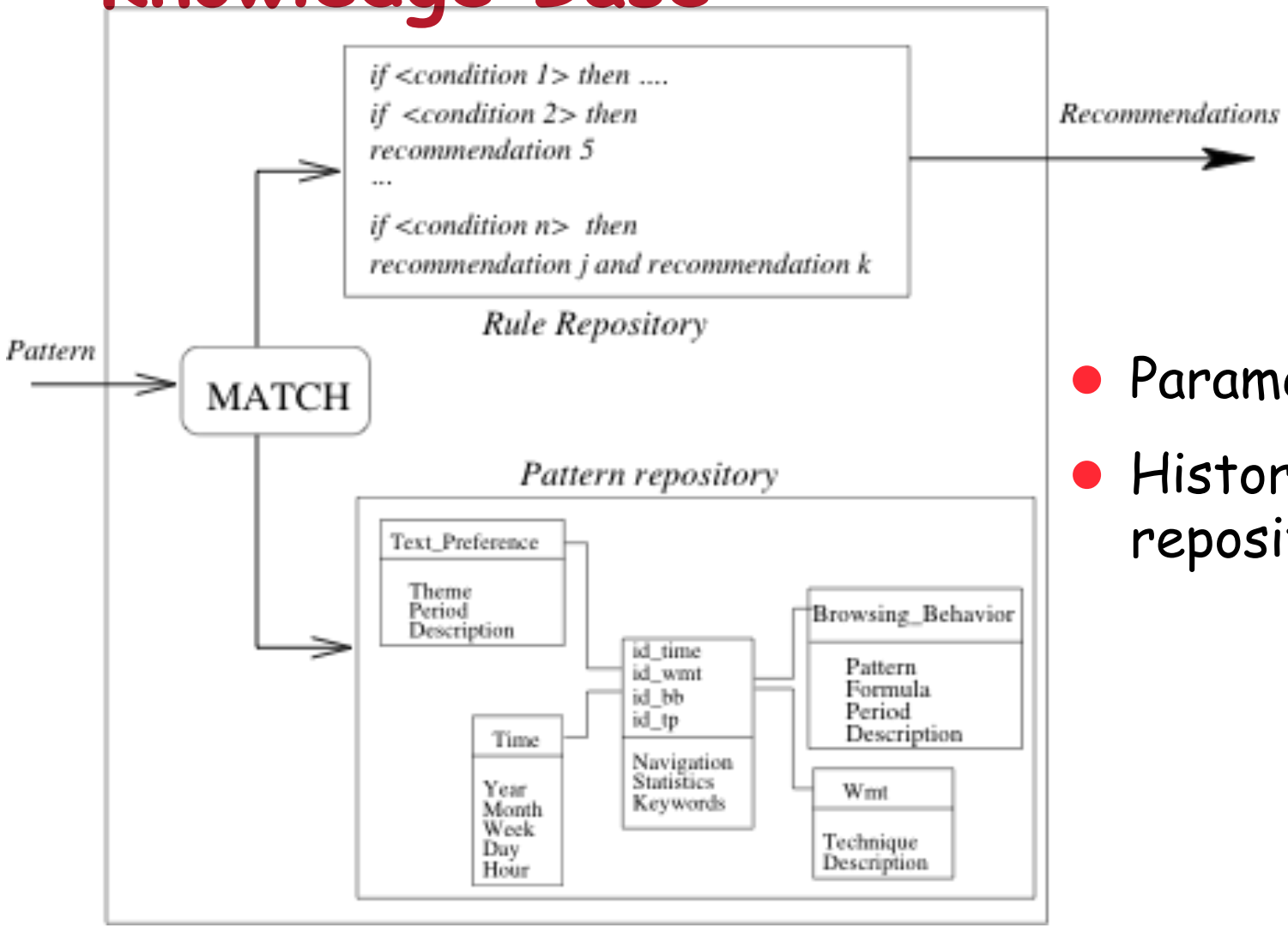
Testing the online navigation effectiveness



Adaptive web site [Velasquez04b]



Knowledge Base



- Parametric rules
- Historic pattern repository

Applications

WEB USAGE AND CONTENT MINING

Definitions

- Web Object: *Any structures, group of words or multimedia resources within a web page that has metadata to describe its content*
- Website Keyobjects: *The Web objects or groups of web objects that attracts web users' attention and that characterizes the content of a given webpage or website.*

[Dujovne L.E., Velásquez J.D., 2009]

Representation of Web Objects

- Every object is described by a bag of concepts each of which is associated to a certain "Category"


```
<?xml version="1.0 encoding="ISO-8859-1">
<Metadatos>
<Page>12</Page>
<objeto mdidobj="obj1" tipo="flash">
<concepto cat="geomarketing">
    censo hogares persona
</concepto>
<concepto cat="cartografía digital">
    mapas digitales urbanos
</concepto>
       …
<concepto cat="información empresa">
    alianzas estratégicas empresa
</concepto>
</objeto>
</Metadatos>
```

[Dujovne L.E., et al., 2009]

- An object can be described by a string that considers every category an object belongs to
  - Each category is listed from A-Z
  - Obj = E|D|L|L|M maybe a valid string that describes the categories which in turn describes the concepts that define each object

Objects Similarity Measure and Sessionization

- Given two objects, every concept is matched by comparing concepts

$$do(O_1, O_2) = 1 - \frac{L(O_1, O_2)}{\max(|O_1|, |O_2|)}$$

- By using the web long and object description a vector is created
  - O is the object, and t is the time spent by each user in that object

$$v = [(o_1, t_1), \ldots, (o_n, t_n)]$$

  - ϑ is the session, α, β are users, min{·,·} is the ratio of the time spent by the users in each object, which is defined above

$$st(\vartheta_i(\alpha), \vartheta_i(\beta)) = \frac{1}{n} \sum_{k=1}^{n} \min\left\{\frac{\tau_k^\alpha}{\tau_k^\beta}, \frac{\tau_k^\beta}{\tau_k^\alpha}\right\} do(o_k^\alpha, o_k^\beta)$$

Extracting Significant patterns

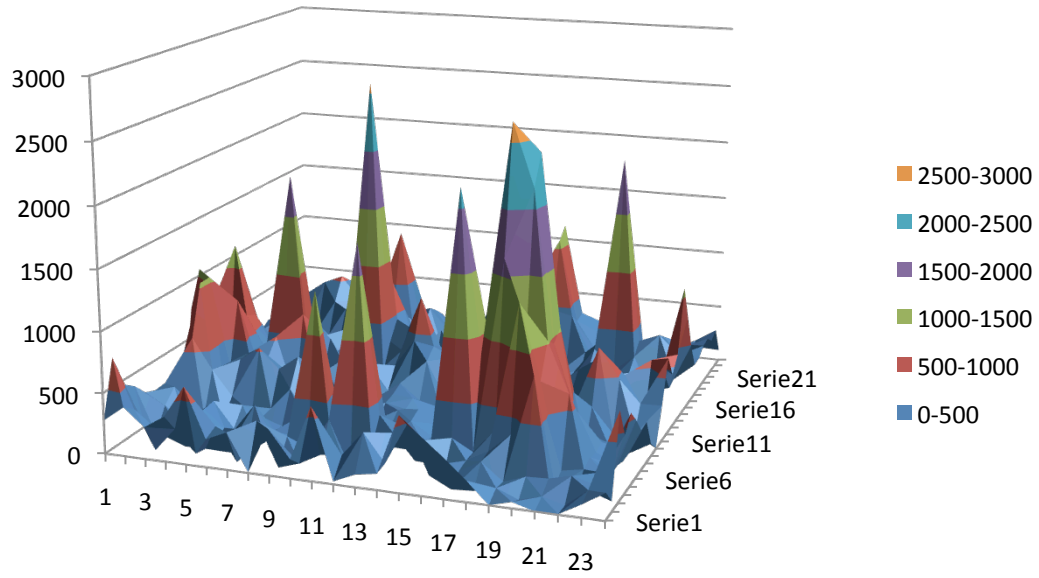
- We applied three different mining algorithms for extracting significant patterns and comparing results.
- These were:
  - SOFM with Thoroidal Topology
  - K-Means (Cross-checking)
  - Association Rules (Cross-checking)
- In the first two algorithms, it is very important the object representation and similarity measure used.
- Association Rule was used as a comparison result method.

The Test Site

- Corporate website of the GIS Service Provider dMapas
- The site contains information of the company, their products and services
- Weblog considers the period of June 2007
- 29 static pages
- 39 Objects
- 27.899 page requests
- 5.866 unique sessions

Results

- SOM:
  - Visualization of Clusters:



  - Concepts Generated By Clusters:

CARTOGRAPHY	CARTOGRAPHY & BUSINESS	DEMOS Y GIS
GEO-BUSINESS	CARTOGRAPHY & DEMOS	BUSINESS Y GIS
CARTOGRAPHY & GIS	CARTOGRAPHY DEMOS & GIS	DEMOS
GEO-BUSINESS & GIS		

Results (3 algorithms)

- At least 5 clusters per algorithm (SOFM found the most clusters (10))

SOM	KMEANS	ASSOCIATION RULES
CARTOGRAPHY	CARTOGRAPHY	CARTOGRAPHY
GEO-BUSINESS	GEO-BUSINESS	GEO-BUSINESS
CARTOGRAPHY & GIS	CARTOGRAPHY & GIS	CARTOGRAPHY & GIS
GEO-BUSINESS & GIS	GEO-BUSINESS & GIS	GEO-BUSINESS & GIS
CARTOGRAPHY & BUSINESS		CARTOGRAPHY & BUSINESS
CARTOGRAPHY & DEMOS	CARTOGRAPHY & DEMOS	
CARTOGRAPHY DEMOS Y GIS		CARTOGRAPHY & GEO-BUSINESS
DEMOS Y GIS		DEMOS Y GEO-BUSINESS
BUSINESS Y GIS		
DEMOS		

The website keyobjects are considered to be the top 10 objects that appears the most

Object	Type	Concept	Count
Index	Flash	General Information about dMapas Products	28
Cartography 1	Text	Technical Information about Cartography	23
GIS Products 1	Text	General Information about GIS Products	17
GIS Products 2	Text	General Information about GIS Products	14
Cartography 2	Text	Technical Information about Cartography	14
Cartography Products	Text	Information about Cartography Provided by dMapas	13
About GIS	Text	General Information about GIS Systems	10
Demo 2	Flash	Demonstration of a GIS Application	10
Geobusiness	Image	Information about Geobusiness	9
Cartography 3	Text	Technical Information about Cartography	9

Web content mining application

Extracting "opinion vectors"

- An electronic news or forum, is a system where the users discuss about a topic.
- Sometimes the topic is very general and the opinions could be a huge amount.
- By using Web Content Mining over the free text in the forum, clusters about the most important opinions could be extracted.
- From the identified clusters, the centroid shows significant information about the a "consensus" opinion

Analyzing the education quality in Chile

- The Chilean Education Ministry have a forum where the users discuss about the quality of the education.
- This forum contain the opinion of five regions (II,V, VIII, X and Metropolitana)
- The total amount of opinion are 6000 approximately.
- We want to extract the "opinion vectors", i.e, the opinion that represent a group of persons with similar thinking about the same topic.

Data source

- Opinions by location and persons' role

Roles	Pto Montt	Concepción	Valparaíso	Antofagasta	Santiago	Chile Total
Father's association	89	102	146	104	269	710
Fathers	194	102	141	29	93	559
Experts	61	37	5	29	26	158
Teachers	176	167	188	154	294	979
Students	83	100	98	19	151	451
Principal	50	48	36	45	124	303
Mineduc	40	0	0	0	0	40
Teacher's association	2	14	0	0	43	59
Business investment	28	15	0	13	4	60
Supervisor	0	7	0	0	0	7
Observer	0	56	13	28	50	147
Advisor	0	0	0	3	0	3
Total	723	648	627	424	1054	3476
Words	66774	65750	69554	65339	147063	414480

Data source (example)

Región II (Antofagasta), Opinion 1 from a student parent : "I choose this topic because I'm interested in the quality of the people that work in a kinder garden, I'm mean I'm worry for the psychological test, because I believe that it is important before to contract somebody.

- The text could contain orthographic and grammatical errors. It was checked for a human being, but the original message was maintained.

Characterizing the analysis

- Problems:
  - A lot of text and words.
  - It is not easy to find group of similar opinions in order to identify preoccupations .
- A possible solution:
  - Application of WCM in order to identify similar opinion group and by extraction of the cluster centroid, to know the cluster opinion, i.e. the forum user central opinion.

WCM: Methodology used on forum

1. Forums cleaning (stop words)
2. Stemming
3. Creating the "opinion/word" matrix.
4. Calculus of the word weight (TF*IDF)
5. Transforming the opinion in a feature vector.
6. Applying a clustering technique.,
7. Cluster identification and centroid extraction.
8. Opinion identification and invert cluster analysis.
9. Semantic interpretation of the original text.

Forums cleaning

- Cleaning stop words like articles, prepositions, conjunctions, etc.
- Applying a synonym table.
  - This table contain a root word and its synonyms.
  - With this procedure a vector dimension reduction can be get.

Stemming

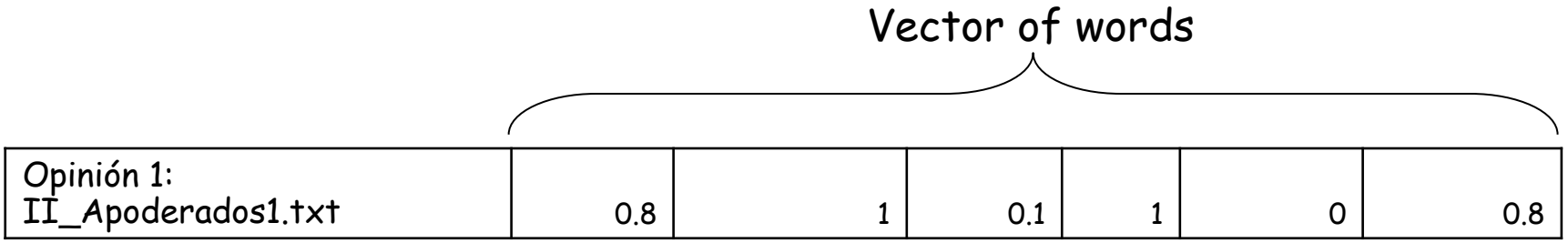
- Word reduction to the root term.
- Example: write, writing, wrote ➔ write.
- This process is very important because if the matrix's dimension is huge, the processing time will be enormous.
- Also it is important to considerate a low term's redundancy.

Matrix opinion/word

	Calidad	Formación	...	...	...	...
Opinion 1: II_Apoderados1.txt	0	1	1	1	0	0
Opinion 2: II_Apoderados2.txt	1	1	0	0	1	1
Opinion 3: II_Apoderados3.txt	1	0	0	1	1	0
... ...	...	...	...	...	...	...

Vector of words

Application of the TD*IDF



Steps 1. – 5.: “From text to vector space model”

Clustering

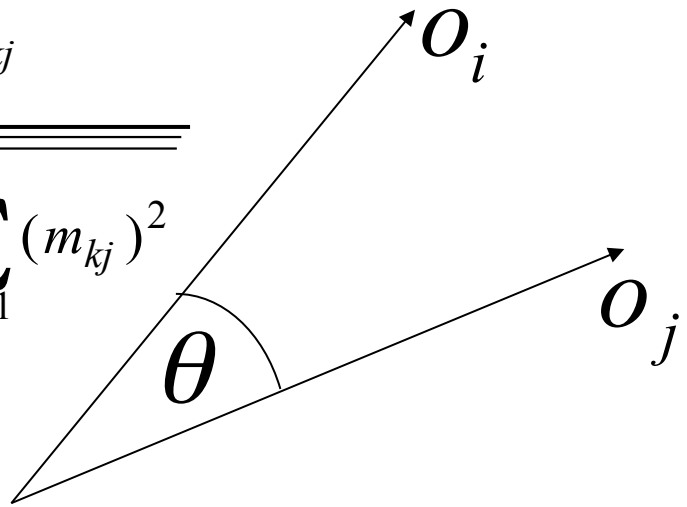
- Technique applied: SOFM
- Similar vectors will be group in the same cluster.
- Here is necessary the human expert collaboration in order to reject/accept the cluster.
- A visualization tool is a big help.
- The cluster creation need a similarity/distortion measure to compare the training set.

Comparing opinions

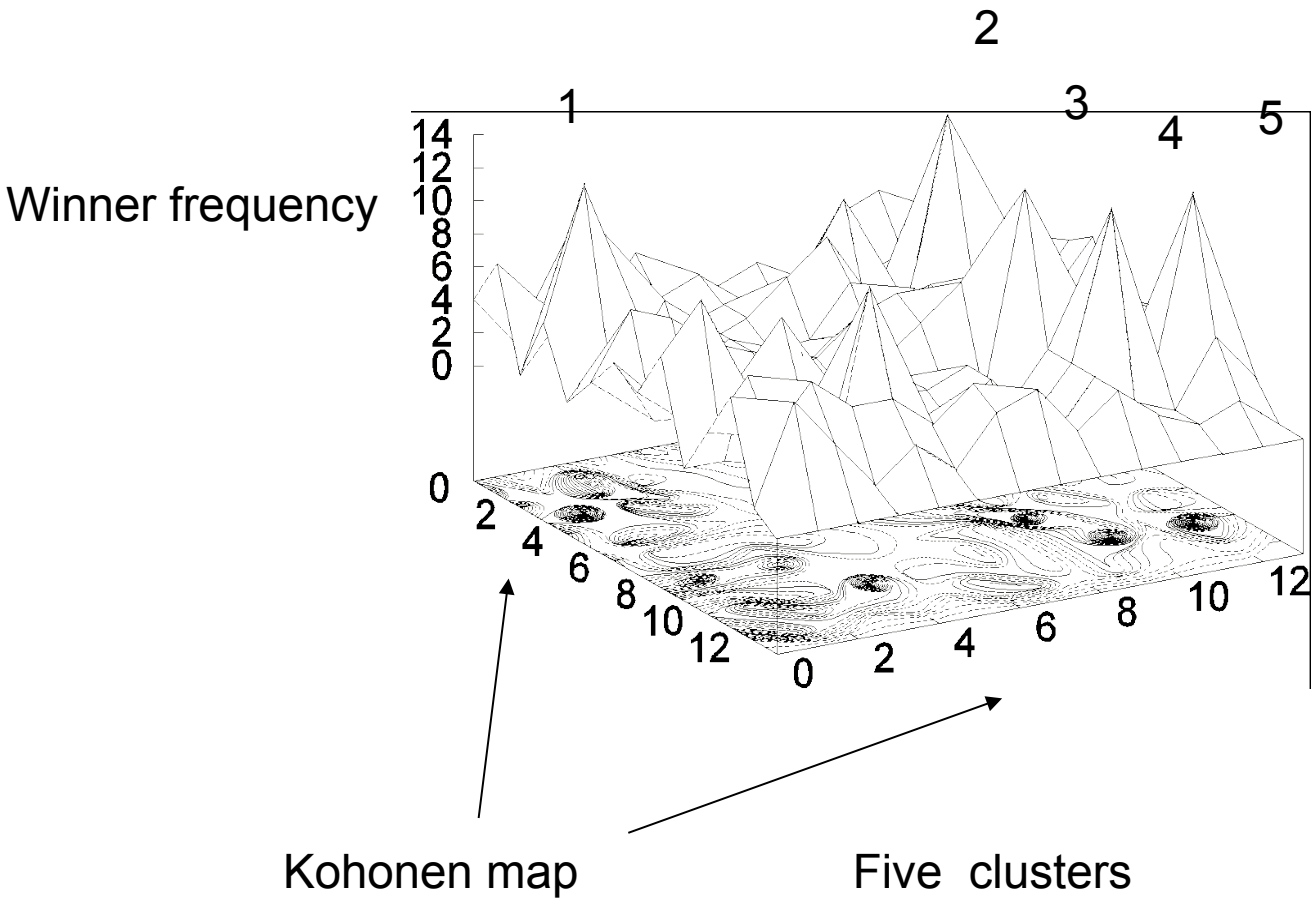
- Because each opinion is a vector, they can be compared by…

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

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



Cluster extraction



Identifying the opinion vector

- Each cluster centroid contain a number in its components.
- Using this representation it is not possible to know the words that represent the vector.
- It is necessary a reverse cluster analysis, i.e., to compare the centroid with the others opinion vector's representation.
- Next the most close vector is selected and the original text is extracted.

Identifying the opinion vector (2)

- The same procedure is applied with the vectors that belong to the cluster.
- The final result is, for example:

Cluster 1, Antofagasta contain the next eight opinions :

II_Director9.txt, II_Apoderados22.txt,
II_CentroApoderados12.txt, II_CentroApoderados56.txt,
II_CentroApoderados80.txt, II_CentroApoderados85.txt,
II_CentroApoderados92.txt, II_Consejero1.txt

Some results

- Three clusters found in Antofagasta:

Cluster 1: 8 opiniones:

5 del CentroApoderados, 1 Director, 1 Consejero, 1 Apoderado

Detalles: II_Director9.txt, II_Apoderados22.txt, II_CentroApoderados12.txt, II_CentroApoderados56.txt, II_CentroApoderados80.txt, II_CentroApoderados85.txt, II_CentroApoderados92.txt, II_Consejero1.txt

Cluster 2: 8 opiniones

6 del CentroApoderados, 2 Estudiantes

Más detalles: Ver informe

Cluster 3: 8 opiniones

4 del CentroApoderados, 1 Apoderado, 3 Profesores

Más detalles: Ver informe

Some results (2)

Cluster 3 in Antofagasta:

II_Apoderados24.txt:

"There is another thing about the child, the child's rights. Sometimes, we say anything to the student, maybe speaking in a strong way, and the student say "**you are harming my rights**", but where are the child duties?, it seem we need some child duties, for instance they must obey their parents"

Some results (3)

- Cluster 1: It is clear that we need an educative project, that take in consideration the differences among each school. The project must consider the entire actors in the student's educations, i.e., the parents must be involved at less in the education a project implementation. In the opinions there is consensus in the problem is not only of the actual government, else every education's actors are guilty, and they can contribute to an eventual solution.

- Cluster 2: It shows restlessness for involving with the best efficiency to the parents in the education process. For instance, to create initiative for the parents that are working can to support in the best way the student's education in the house. Also it is important to emphasize the passive role that sometime the parents have in relation with the school. If we want the parents help in the student's education, it is necessary to create instances where the parents can learn about how to teach, because a lot of them don't have idea about how to teach a particular topic to the child.

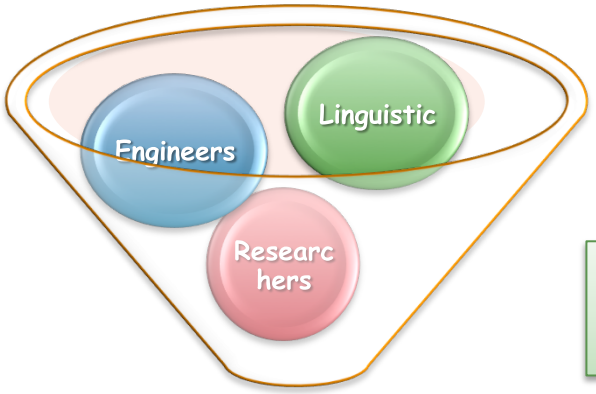
CURRENT PROJECT: DOcument COpy DEtector (DOCODE)

Problem & Opportunity

1. • **Usual Practice**: Writing based on other sources
2. • Unlimited access to information
   • Digital libraries, powerful web search engines, Web.
3. • **PROBLEM**: The document is not original, it is simply a copy of sources without mentioning original authors.
4. • **MOST AFFECTED INDUSTRY**: Education
   • The student acquires bad habits and the teaching process is compromised.
5. • **SOLUTION?**:The teacher has to revise the document's originality. This task consumes several recourses.

There is an opportunity to develop a tool to automate the detection of plagiarism in digital documents written in Spanish

Methodology



Join effort

Linguistic strategies

There will be a teacher inside the development team

IMPACT:
- Dissuasive tool
- Control tool
- Benchmarking

DOCODE
- Traditional methodology to algorithms' development
- Agile methodology to interface's development

PERFORMANCE AND IMPACT ANALYSIS
- PERFORMANCE: Manual checking

Diagnostics usage of DOCODE
- Training
- Manuals
- Surveys

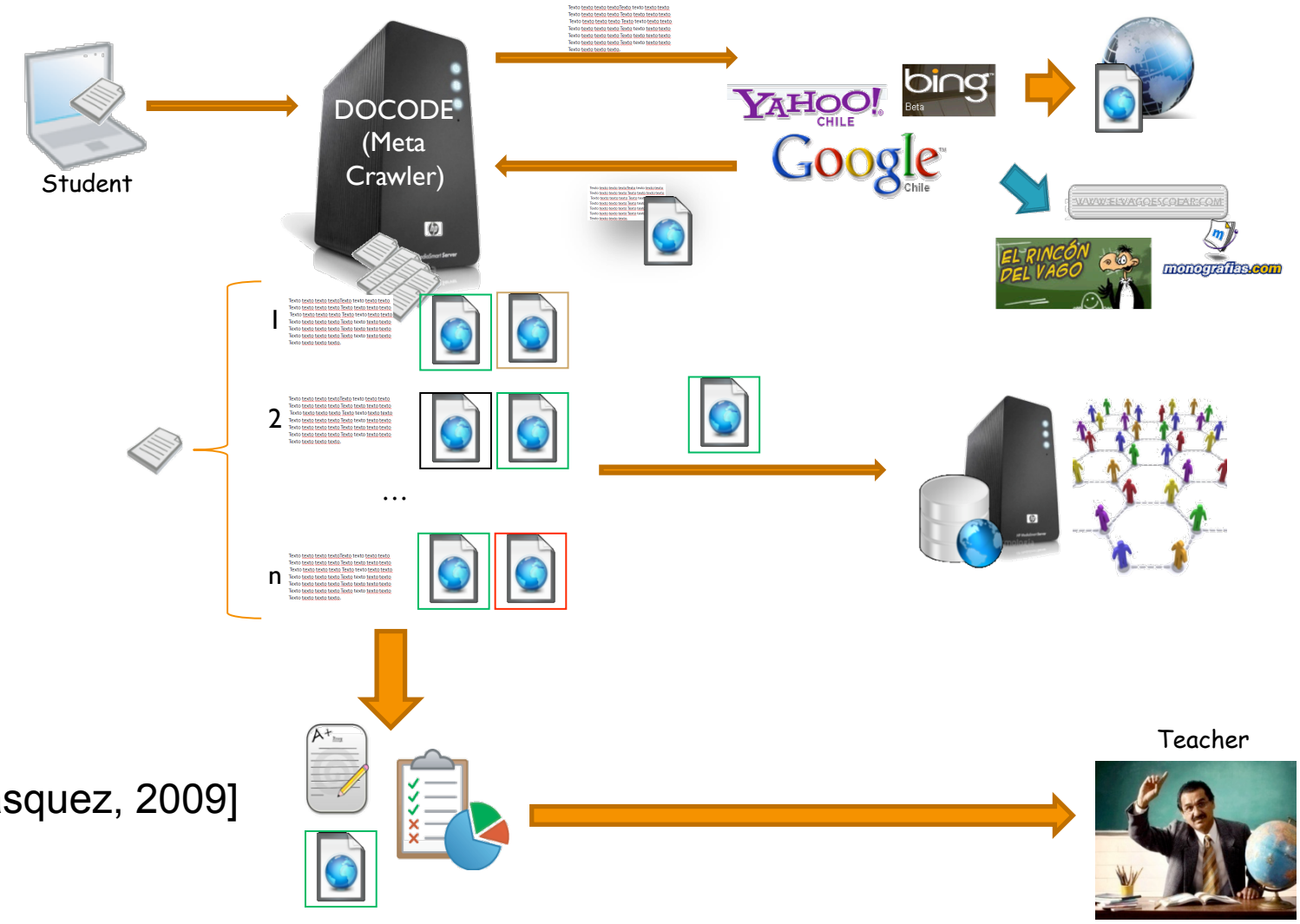
Generate Social Conscience
- Preliminary Survey
- Market Research
- Difusión of results

Spin-off company
- Business Plan
- Training
- Support

DOcument COpy DEtector

[J. D. Velásquez, 2009]

DOCODE Solution



[J. D. Velásquez, 2009]

SUMMARY

Summary

1. Web data is a real source to analyze the user behavior in the web.
2. An important step is the cleaning and pre-processing of the web data.
3. The application of web mining techniques permits one to find unknown patterns.
4. These patterns must be validated/rejected by an expert in the firm who is under investigation.
5. We can personalize a web site.

Muchas gracias por su atención

Thank you very much for your attention

ご清聴ありがとうございました

References

J.D. Velásquez, H. Yasuda, T. Aoki, R. Weber and E. Vera, 2003, Using self organizing feature maps to acquire knowledge about visitor behavior in a web site, Lecture Notes in Artificial Intelligence, 2.

J.D. Velásquez, H. Yasuda, T. Aoki and R. Weber, 2004, A new similarity measure to understand visitor behavior in a web site, IEICE Transactions on Information and Systems, E87-D(2): 389-396.

P.E. Roman and R. Dell and J.D. Velásquez, Web User Session Reconstruction Using Integer Programming. Procs. of The 2008 IEEE/WIC/ACM International Conference on Web Intelligence. Sydney, Australia.

P.E. Román and J.D. Velásquez, Cadenas de Markov para modelar la navegación del Usuario Web: Inferencia Estadística. Procs. of The XIV Latin Ibero-American Congress on Operations Research (CLAIO 2008). Cartagena, Colombia, 2008.

Dell, R., P. Roman, and J Velasquez. "Optimization Models for Construction and Analysis of Web User Sessions" 11th Informs Computing Society Conference, Charleston, South Carolina, USA, January 11-13 2009.

Luhn, H.P. A statistical approach to the mechanized encoding and searching of literary information. IBM Journal of Research and Development 1:4; 309-317; October 1957.

B. Huberman, P. Pirolli, J. Pitkow, R. Lukose. “Strong Regularities in World Wide Web Surfing”, **SCIENCE**, Vol 280, p. 95-97, 1998.

N. ZHONG and J. LIU. Web Intelligence. Springer, 2003.

J. D. VELASQUEZ and V. PALADE. Adaptive Web site: A Knowledge Extraction from Web Data Approach. IOS Press, 2008.

M. SPILIOPOULOU, B. MOBASHER, B. BERENDT, and M. NAKAGAWA. A framework for the evaluation of session reconstruction heuristics in web-usage analysis. INFORMS Journal on Computing, 15(2):171190, 2003.

Mobasher B., Berent B., S.M.W.J.: Measuring the accuracy of sessionizers for web usage analysis. In: Proceedings of the Web Mining Workshop at the First SIAM ICDM. (2001)

Román P., Dell R., Velásquez J.: "Optimization models for construction of web user sessions". Submitted to Informs Journal on Computing.

C.D. Manning, H. Schutze, Fundation of Statistical Natural Language Processing (The MIT Press, 1999).

J. D. VELASQUEZ, H. YASUDA, T. AOKI, and R. WEBER. A new similarity measure to understand visitor behavior in a web site. IEICE Transactions on Information and Systems, Special Issues in Information Processing Technology for web utilization, E87-D(2):389-396, February 2004.

Mobasher B.: Data Mining for Personalization. Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.). Lecture Notes in Computer Science, Vol. 4321, PP. 90-135, Springer, Berlin-Heidelberg, 2007.

J. D. Velásquez: "DOCODE, Document Copy Detector" National Project Foundation 2009, Conycit Chile.

Anderson, C., Domingos, P., Weld, D.: Adaptive web navigation for wireless devices. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence, Seattle, Washington (August 2001) 879–884.

Borges, J., Levene, M.: Data mining of user navigation patterns. In Masand, B., Spiliopoulou, M., eds.: Web Usage Analysis and User Profiling: Proceedings of the WE-BKDD'99 Workshop. LNAI 1836. Springer-Verlag (1999) 92–111.

Cadez, I., Heckerman, D., Meek, C., Smyth, P., White, S.: Model-based clustering and visualization of navigation patterns on a web site. Journal of Data Mining and Knowledge Discovery 7(4) (2003) 399–424.

Deshpande, M., Karypis, G.: Selective markov models for predicting web-page accesses. ACM Transactions on Internet Technology 4(2) (2004) 163–184.

Pitkow, J., Pirolli, P.: Mining longest repeating subsequences to predict www surfing. In: Proceedings of the 2nd USENIX Symposium on Internet Technologies and Systems, Boulder, Colorado (October 1999).

Dujovne L.E., Velásquez J.D.: "Design and implementation of a methodology for identifying web site key object" to appears in proceedings KES 2009.

Velásquez J.D., Gonzalez P.: "Expanding the possibilities of deliberation" to appears in The Information Society Journal November 2009.

Perkowitz, M., Etzioni, O.: Adaptive web sites: Automatically synthesizing web pages. In: Proceedings of the 15th National Conference on Artificial Intelligence, Madison, WI (July1998) 727–732.

Perkowitz, M., Etzioni, O.: Adaptive web sites. Communications of ACM 43(8) (2000) 152–158.

J. Sobecki: Web-Based System User Interface Hybrid Recommendation Using Ant Colony Metaphor, in proceedings of KES 2007, 1033-1040.

Kaski, S.: Dimensionality reduction by random mapping: Fast similarity computation for clustering. In Proceedings of IJCNN'98, International Joint Conference on Neural Networks, volume 1, pages 413-418. IEEE Service Center, Piscataway, NJ.

Z. MARKOV and D. T. LAROSE. Data Mining the Web: Uncovering Patterns in Web Content, Structure and Usage. John Wiley & Sons, 2007.

Spiliopoulou, M., Faulstich, L.: Wum: A tool for web utilization analysis. In: Proceedings of EDBT Workshop at WebDB'98. LNCS 1590, Springer Verlag (1999) 184–203.

Jin, X., Zhou, Y., Mobasher, B.: Web usage mining based on probabilistic latent semantic analysis. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD04), Seattle, WA (August 2004) 197–205.

Jin, X., Zhou, Y., Mobasher, B.: A unified approach to personalization based on probabilistic latent semantic models of web usage and content. In: Proceedings of the AAAI 2004 Workshop on Semantic Web Personalization (SWP'04), San Jose, CA (2004).

P. K. CHAN. Constructing web user profiles: A non-invasive learning approach. In WEBKDD, pages 39-55, 1999.

Y. Chen and C. Shahabi, *Improving User Profiles for E-Commerce by Genetic Algorithms *, In E-Commerce and Intelligent Methods Studies in Fuzziness and Soft Computing, Kulwer Academic Publishers, 2002, ISBN 3-7908-1499-7.

Abraham A., Ramos V., Web Usage Mining Using Artificial Ant Colony Clustering and Genetic Programming, Proceedings of the 2003 Congress on Evolutionary Computation CEC2003, 1384-1391.