

Diapositive 1/30

## Co-Site Analysis & Indicators for Information Society

Xavier Polanco  
URI-INIST-CNRS

Diapositive 2/30

## EICSTES

- European Indicators, Cyberspace and the  
Science-Technology-Economy System

– IST – 1999-20350 :

- » <http://www.EICSTES.ORG/>
- » Participants: National Research Council (CSIC) Spain; Austrian Research Center Seibersdorf (ARCS) Austria; National Center for Scientific Research (CNRS) France; Computer Technology Institute of Patras (CTI) Greece; University of Amsterdam (UvA) Netherlands; University of Surrey (UNIS) United Kingdom; Statistic Institute of Catalonia (IDESCAT) Spain

## Diapositive 3/30

The data on which this application is based were collected in January 2001 by the team of Moses A. Boudourides Computer Technology Institute University of Patras, as part of the EICSTES project.

An autonomous intelligent agent operating on the Alta Vista search engine and for every academic site was used.

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      3

## Diapositive 4/30

### Introduction

- Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services
- Three Web mining categories:
  - Web structure mining
  - Web content mining
  - Web usage mining

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      4

## Diapositive 5/30

**$D \Rightarrow A \Rightarrow C \Rightarrow M$**

From data (D) (hyperlink web site matrix) to Associations (A):  
D = HLWS  
D  $\Rightarrow$  A; A  $\neq$  HL

From associations (A) to clusters (C):  
A  $\Rightarrow$  C

From clusters (C) to map (M): C  $\Rightarrow$  M

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      5

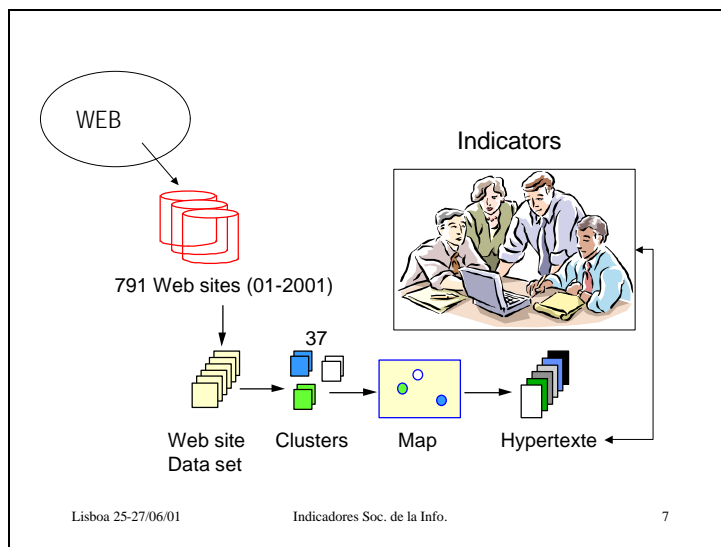
Diapositive 6/30

## Method

1. Transformation of the raw data matrix into the association matrix
2. Application of an association coefficient
3. Cutting the network of associations into clusters
4. Placing the clusters on a 2D map

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      6

Diapositive 7/30



Diapositive 8/30

## Data Matrix

Data refers to  $N$  academic sites from the European Union countries ( $N=791$ )  
N-square matrix  $D$  defined:

$$D=d(i,j);i=1,N;j=1,N$$
$$d(i,j)\geq 0$$

where:  
 $d(i,j)$  = number of links between site  $i$  and site  $j$ ;  
 $d(i,i)$  = number of internal links of the site  $i$

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      8

Diapositive 9/30

**STEP 1: Data Transformation**

each site  $s$  of  $D$  matrix is calculated the site associations (C as symbol -)

$$C(i, j) = \begin{cases} 1 & \text{If sites } i \text{ and } j \text{ together in the site } s \\ & \text{and } d(s, i) = 0 \text{ AND } d(s, j) = 0 \\ & \text{and } d(s, i) \neq 0 \text{ AND } d(s, j) \neq 0 \end{cases}$$

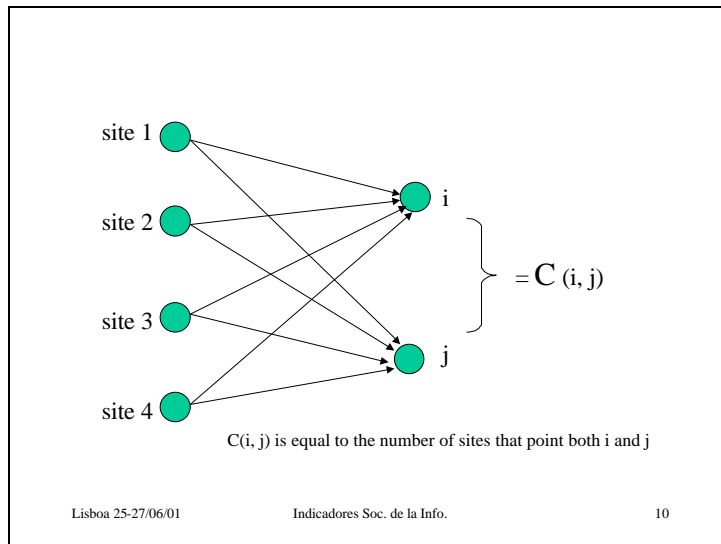
$d(s, i) = 0$  OR  $d(s, j) = 0$

$d(s, i) \neq 0$  AND  $d(s, j) \neq 0$

$C(i, j) = \begin{cases} 1, & i = j \\ -1, & i \neq j \end{cases}$

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      9

Diapositive 10/30



Diapositive 11/30

To obtain the total number of co-occurrences, we compute:

$$C(i, j) = \sum_{\substack{i=1, N-1 \\ j=i+1, N}} \sum_{\substack{s=1, N \\ s \neq i, s \neq j}} C_s(i, j)$$

with

$$C(i, j) \in [0, N - 2] \forall i, j = 1, N; i \neq j$$

$$C(i, i) = 0; \forall i = 1, N$$

This permits to highlight indirect hidden associations from hyperlinks data between Web sites

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      11

Diapositive 12/30

**STEP 2: The use of an association coefficient**

$$E(i, j) = \frac{C(i, j)^2}{C(i)C(j)}$$

With:  
C(i,j) = number of times sites i and j are associated  
C(i) = number of times site i appears  
C(j) = number of times the site j appears

Where:

$$C(i) = \sum_{s \neq i} d(s, i) \quad \forall i = 1, N \quad \text{and} \quad C(j) = \sum_{s \neq j} d(s, j) \quad \forall j = 1, N$$

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      12

Diapositive 13/30

Association matrix A gives a normalized measure of the strength of associations between the sites

$$A(i, j) \in [0, 1]; \forall i, j = 1, N; i \neq j$$

and

$$A(i, i) = 0; \forall i = 1, N$$

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      13

Diapositive 14/30

**STEP 3: CLUSTERING**

- 1 ⇒ Clustering process
- 2 ⇒ Cluster structure
- 3 ⇒ Analysis of clusters

Clustering is used to enhance search, browsing and visualization. In our case, clustering is used to enhance co-site analysis.

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      14

Diapositive 15/30

## Clustering process

(bottom up)

Single link clustering technique

Saturation strategy

2. Minimum size of clusters (4)
3. Maximal size of clusters (10)
4. Maximum number of associations (20)
5. Maximum number of external association (10)

Lisboa 25 27/06/01Indicadores Soc. de la Info.15

Diapositive 16/30

## Internal and External Associations

Clusters are formed by associated sites and corresponding to their internal associations or intra-clusters associations.

On the other hand, clusters have external associations or inter-cluster associations

Lisboa 25-27/06/01Indicadores Soc. de la Info.16

Diapositive 17/30

Lisboa 25-27/06/01Indicadores Soc. de la Info.17

Diapositive 18/30

Cluster Structure

The sites appearing in the internal ass. are called internal sites  
The number of internal sites defines the size of the cluster

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      18

Diapositive 19/30

### Analysis of clusters

1. Cluster saturation threshold
2. Density
3. Centrality
4. Number of internal sites
5. Number of external sites
6. Number of internal associations
7. Number of external associations
8. Number of times a cluster is referenced by th others
9. Number of sites classified in a given cluster
10. Number of sites exclusively related to the cluster

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      19

Diapositive 20/30

ed.ac.uk	0.922	0.925	0.918	7	4	10	10	13	708	1
shof.ac.uk	0.915	0.918	0.918	7	5	10	10	10	707	0
rwth-aachen.de	0.911	0.913	0.909	8	8	10	8	25	706	0
casa.unimo.it	0.908	0.915	0.897	10	1	16	1	29	533	0
tau.nl	0.902	0.895	0.906	5	6	10	9	8	697	1
kuleuven.ac.be	0.899	0.894	0.902	6	5	10	8	8	696	1
uni-bonn.de	0.897	0.897	0.902	7	5	10	10	6	691	1
juyu.fi	0.887	0.885	0.888	7	5	10	6	5	675	0
upcs.es	0.886	0.870	0.877	5	3	10	5	11	654	0
brookes.ac.uk	0.883	0.887	0.882	10	1	12	1	14	634	1
th-darmstadt.de	0.882	0.882	0.886	7	4	10	6	0	665	0
hb.se	0.881	0.884	0.880	6	3	10	10	2	524	0
ba-avensburg.de	0.878	0.895	0.878	10	1	17	1	14	506	0
uafm.es	0.878	0.889	0.880	9	5	11	9	12	571	0
eap.fr	0.876	0.914	0.000	8	0	20	0	9	60	0
u-strasbg.fr	0.873	0.877	0.869	8	4	10	5	10	684	0
uni-potsdam.de	0.871	0.877	0.875	9	4	10	6	8	638	1
ub.es	0.868	0.861	0.872	5	4	10	9	5	641	0
lton.ac.uk	0.867	0.862	0.872	4	5	6	8	2	607	0
lamp.ac.uk	0.865	0.844	0.867	4	8	6	9	2	575	0
emse.fr	0.864	0.874	0.000	9	0	19	0	0	658	1
rhl.nl	0.862	0.856	0.860	7	6	10	8	2	523	0
univ-mlv.fr	0.860	0.863	0.864	6	4	10	10	4	634	1
ull.es	0.858	0.844	0.855	5	9	10	9	2	578	1
univ.es	0.856	0.851	0.868	4	8	6	10	3	590	0
khs-linz.ac.at	0.849	0.848	0.864	4	9	6	10	2	506	0
tvu.ac.uk	0.834	0.827	0.810	4	6	6	7	3	481	0
wales.ac.uk	0.759	0.775	0.762	7	3	10	6	0	505	0
hva.gr	0.749	0.803	0.752	6	6	10	9	0	80	0
unilm.fr	0.678	0.640	0.676	4	2	6	2	0	521	0
sghms.ac.uk	0.639	0.644	0.644	7	2	10	4	0	443	0
wales.ac.uk	0.596	0.628	0.000	10	0	15	0	0	272	0
hua.gr	0.588	0.703	0.000	7	0	20	0	10	37	1
iulm.it	0.451	0.392	0.000	5	0	10	0	0	148	0
unioja.es	0.402	0.409	0.337	6	3	10	3	10	274	0
teilar.gr	0.345	0.342	0.414	5	5	10	10	0	76	0
uia.es	0.199	0.291	0.251	4	4	6	10	0	92	0

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      20

Diapositive 21/30

**STEP 4: MAPPING**

1 ⇒ Constructing Map  
2 ⇒ Significance Map Analysis  
3 ⇒ Analyzing Clusters Relationships

The map allows to understand the global and local structure brought out by clustering from the association matrix

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      21

Diapositive 22/30

**The Map Construction**

The measures of *density* and *centrality* allow the visualization of clusters and theirs relationships in a 2D space

Where:

the X-axis corresponds to *centrality* and the Y-axis to *density*

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      22

Diapositive 23/30

$$Density = \frac{\sum Ass.int}{N} = Y-axis$$

Average value of the internal associations

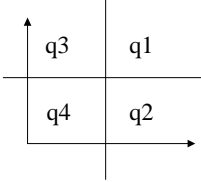
$$Centrality = \frac{\sum Ass.ext}{N} = X-axis$$

Average value of the external associations

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      23

Diapositive 24/30

**Significance Map Analysis**

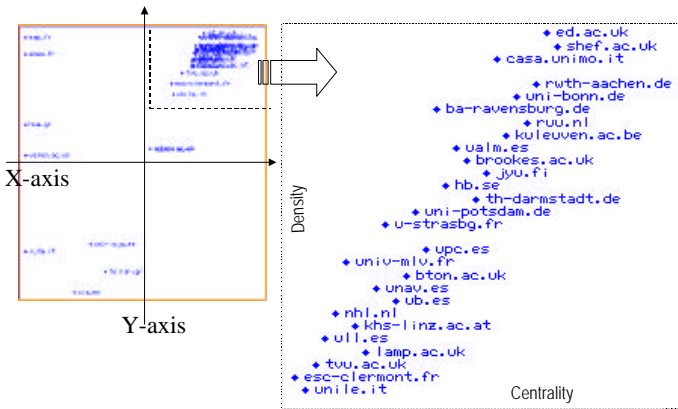


Four types of clusters can be distinguished:

Type 1: high density and centrality	= quadrant 1
Type 2: low density and high centrality	= quadrant 2
Type 3: high density and low centrality	= quadrant 3
Type 4: low density and centrality	= quadrant 4

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      24

Diapositive 25/30



Lisboa 25-27/06/01      Indicadores Soc. de la Info.      25

Diapositive 26/30

**INDICATORS**

Density index  
Centrality index  
Structural ratio  
Transformation index

These indicators are said relational indicators  
They are designed for measuring the relationships  
and interactions between clusters

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      26

## Diapositive 27/30

The *centrality index* measures the strenght of associations between one cluster and the other clusters in the field

The *density index* measures the degree of cohesion of clusters. The denser cluster, the more it will consist of tightly connected Web sites

The *structural index* is the report or ratio of the centrality to the density. A low value of this indicator can be indicated that a cluster can be cohesive but located at the periphery of the network. A strong value, a cluster can be central and thus strategic but little cohesive

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      27

## Diapositive 28/30

The *transformation index* measures change in the components of a cluster over time  
Continuity and change can be expressed by this index

The transformation index is obtained by dividing the number of items which two subsequent clusters have in common by the total number of items of the two clusters:

$$T = \frac{Cl(t) \cap Cl(t+1)}{Cl(t) + Cl(t+1)}$$

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      28

## Diapositive 29/30

### CONCLUSION

- \* The pattern of associations among key sites within clusters and inter-clusters establishes a structure for a given Web domain.
- \* This structure may then be observed to change over time.
- \* Through the study of these structure changes, co-site analysis provides a method for
  - ⇒ assessing the degree of interrelationships,
  - ⇒ and analyzing the development of Web fields

Lisboa 25-27/06/01      Indicadores Soc. de la Info.      29

## Diapositive 30/30

Finally, we can also consider to do a content analysis of Web sites applying truly co-word analysis to Web pages.

Each Web site will then be analyzed as a collection of texts indexed by keywords and using the same set of indicators

Thus, we have the same method for analyzing

⇒ the Web structure and

⇒ the Web content.

This is very advantageous