

## **Clustering and Mapping Web Sites. For Displaying Implicit Associations and Visualizing Networks.**

Xavier Polanco, Moses A. Boudourides, Dominique Besagni, Ivana Roche

Modification : IRZ (10/12/2001)

---

**Abstract:** This paper describes the implementation of an association analysis method to process Web site data. The association analysis makes use of Web site links in order to produce a representation of sites association structure. A description of the association network is the goal to reach. The data refers to 791 university sites from the 15 countries belonging to the European Union. The purpose is to perform the analysis of implicit associations discovered from this sample. In order to realize it, we propose to translate the *co-word analysis* and applying it as *co-site analysis*. Given a set of Web sites, two sites are considered to be associated when they are both present as outgoing links in other sites. We expose the association coefficient, the clustering and mapping techniques that are used. Their results are also exposed. In this context, the problem of information society indicators will be discussed.

The subject of this paper was presented as slide presentation titled: "*Clustering and Mapping European University Web Sites Sample for Displaying Associations and Visualizing Networks*" and cosigned by Xavier Polanco, Moses A. Boudourides, Dominique Besagni and Ivana Roche. ETK-NTTS Conference, Crete, June 18-22, 2001. ETK = Exchange of Technology and Know-How; NTTS = New Techniques and Technologies for Statistics. International Conference supported by EUROSTAT and Joint Research Centre: Institute for Systems, Informatics and Safety of the European Commission.

See:

A first version was the source of the X. Polanco slide presentation at the Lisboa Workshop, 25-27 June 2001, organized by RICYT, Red de Indicadores de Ciencia y Tecnología Iberoamericana e Interamericana, and OCT, Observatório das Ciências e das Tecnologias of the Portugal.

The work reported here has taken place in the EICSTES research project (IST-1999-20350). EICSTES, *European Indicators, Cyberspace and the Science-Technology-Economy System*, is supported by the Fifth Framework Program of R&D of the European Commission.

See:

---

## PLAN

### 1. INTRODUCTION

### 2. DATA

### 3. METHOD

3.1 Data transformation

3.2 Association coefficient

### 4. CLUSTERING

4.1 Clustering Process

4.2 Cluster Structure

**Figure 1:** The structure of a cluster

4.3 Analysis of Clusters

**Table 1:** The clusters and their numeric characteristics

4.4 Analyzing Cluster Relationships

**Table 2:** The categories of clusters

### 5. MAPPING

5.1 Constructing Maps

**Figure 2:** The map and the  $q_1$

5.2 Map Significance Analysis

**Table 3:** The types of clusters

### 6. INDICATORS

Centrality index

Density index

Ratio  $c/d$

Transformation index

### 7. CONCLUSION

### 8. REFERENCES

## 1. INTRODUCTION

The central subject of this paper is the association analysis method that we shall present and illustrate by means of a particular application on the Web, namely a set of academic sites. We propose to call it co-site analysis in analogy to co-word analysis. Previous experience on the analysis of scientific and technological information includes descriptive statistics, cluster analysis and cartography or mapping algorithms that represent the generated clusters in the form of maps. A hypertext interface generator provides the user with a friendly interface displaying the global map, the clusters and the documents set and then it gives access to useful information organized by clusters. This summarizes the approach that we are currently applying in the field of bibliographic databases and that we shall try now to extend to the Web domain.

The approach that we shall expose can be related to *Web mining*. The goal of the work often called data mining has been defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from given data" (Frawley et al, 1991). Justly what is called in the title "implicit associations" represents the "implicit, previously unknown, and potentially useful information." The Web mining task is extracting it from a given set of Web sites. The extracting method that will be exposed is founded on clustering and mapping procedures.

The goal is here to extract information from hidden patterns in a large set of Web data. This goal is called *Web mining*. These "implicit associations" can be considered as the hidden pattern from which we try to extract information. Web mining has been decomposed into four subtasks, namely: resource finding, information selection, generalization, and analysis. The last two subtasks characterize our work. The task is called "generalization" when it automatically discovers general patterns at individual Web sites as well as across multiple sites. "Analysis" is the validation and/or interpretation of the mined pattern. On the other hand, Web mining has been categorized into three areas: Web content mining, Web structure mining, and Web usage mining (see Kosala and Blockeel, 2000). The application exposed in this paper corresponds to "Web structure mining."

The paper organization is the following. At first the data are briefly described (section 2). Then the method is closely exposed at its four steps in sections 3, 4 and 5: the data

transformation into an association matrix is exposed, and the use of an association coefficient is introduced (section 3). This is followed by the hierarchical agglomerative clustering method description (section 4). Finally we examine the mapping of clusters by density and centrality (section 5). The indicators are briefly discussed in section 6.

## 2. DATA

The data on which this paper is based were collected in January 2001 by the Computer Technology Institute of Patras, Greece, as part of the project EICSTES. The original data set refers to 1064 academic Web sites from 22 European countries including the all of the European Union. For every academic site, an autonomous intelligent agent operating on the AltaVista search engine was used to measure the number of links from every academic site to another of the data set together with the internal links in these sites.

From these data we obtain a  $N$ -square matrix noted  $\mathbf{D}$  where  $N$  is equal to the number of sites considered in the data set. Then we have  $\mathbf{D}(i, j)$ , the number of links from the site  $i$  to the site  $j$ , and  $\mathbf{D}(i, i)$ , the number of internal links in the site  $i$ .

We have reduced the initial data set to the 15 countries of the European Union. This subset represents 791 university Web sites ( $N = 791$ ).

## 3. METHOD

The problem faced by all analysts of the Web is the need to reduce the amount of information to a manageable form to be examined. For this purpose, we propose a clustering and mapping method based on the co-site association analysis.

We will focus on the analysis of the four steps of the method. We want to demonstrate how this approach can be used to support the analysis of the relevant structure of the Web sites data set.

The *first step* consists on the transformation of the raw data matrix into an association matrix.

The *second step* is the computation of an association coefficient.

The *third step* consists of decomposing the network of associations into clusters.

The *fourth and last step* consists of placing the obtained clusters to a two-dimensional map.

The software that we are using is the co-word analysis program (called SDOC) that has been adapted now to support co-site analysis. This proves that both co-word and co-site analyses are specific cases of a more general approach: the association analysis method. Moreover, the analysis of citations and co-citations that are also applied in the Web domain both belong to this general association analysis.

In the remaining of this section, we will examine the first two steps of the method.

### 3.1 Data transformation

In order to draw a network of associations, we introduce the notion of co-occurring sites: if there is at least one site  $s$  ( $s \neq i$  and  $s \neq j$ ) with links to site  $i$  and to site  $j$ , then sites  $i$  and  $j$  are called co-occurring. Two sites that occur together in a large number of sites are considered to be closely related. So, we need to determine for each couple of sites how many times their links co-occur in the  $N-2$  other sites of the considered data set. For this purpose for each site  $s$  (row of  $D$ ) we calculate the co-occurrences  $C_s(i, j)$  equal to 0, if sites  $i$  and  $j$  do not occur in site  $s$  and equal to 1, if sites  $i$  and  $j$  occur in site  $s$ .

$$C_s(i, j) = \begin{cases} 0, & \text{if } [D(s,i)=0 \text{ OR } D(s,j)=0] \\ 1, & \text{if } [D(s,i) \neq 0 \text{ AND } D(s,j) \neq 0] \end{cases}$$

$s=1, N$   
 $i=1, N-1; i \neq s$   
 $j=i+1, N; j \neq s$

To obtain the total number of co-occurrences of couples of sites, we calculate:

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

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 is the first step producing as result a very large association matrix. The co-site analysis is based on these associations. The links between sites were given in the data. The associations were not empirically given but they are hidden in the data. The task is double: extracting these implicit associations between Web sites from data, and measuring the strength of these associations. Up to now, we have presented the extracting task.

### 3.2 Association coefficient

The next step is to compute an association coefficient. This type of computation is used to establish similarity between cases described by binary values: 1 refers to the presence of a variable, and 0 to its absence. In our case, the variable is the association between two sites,  $i$  and  $j$ . Old good classical references for those interested in association coefficients are Sneath and Sokal (1973), Clifford and Stephenson (1975), and Everitt (1980). At this step the goal is to obtain normalized associations.

The value of  $C(i, j)$  signifies the number of times when two sites are associated in the outgoing links of other sites. Alone, this kind of co-occurrence values is not enough to measure the strength of associations. Because it favors high-frequency couples compared to those with low frequency. Using an association coefficient, the measure of associations between two sites can be normalized. Thus we obtain values between 0 and 1. For this purpose, we apply the association coefficient called Equivalent coefficient (Michelet, 1988).

The association coefficient applied here is denoted:

$$E_{ij} = \frac{C(i, j)^2}{C(i) \times C(j)} = \frac{C(i, j)}{C(i)} \times \frac{C(i, j)}{C(j)}$$

where:

$C(i,j)$  is the total number of associations of sites  $i$  and  $j$ , calculated above in section 3.1.

$C(i)$  is the total number of times that site  $i$  appears in the links of other sites

with:

$$C(i) = \sum_{s=1, N} D(s, i); \forall i = 1, N; s \neq i$$

We obtain an N-upper triangular matrix, denoted  $A$ , composed of the association coefficients values. The association matrix  $A$  gives a normalized measure of the strength of associations between sites:

$$A(i, j) = E_{ij}$$

with:

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

and

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

Let us now review the Equivalence index properties. This is a simple, local and homogenous association coefficient.

The Equivalence index is said to be a simple association coefficient because it is a function only of the local information of the type type  $A_{ij} = F(N, C_i, C_j, C_{ij})$ . All the other elementary statistical data result from these calculations. From these statistics, it is possible to calculate any coefficient or index of association between two objects  $i$  and  $j$ .



A simple association index is a function  $F$  that for each quadruplet  $(N, C_i, C_j, C_{ij})$  depending on  $(i, j)$  associates a real-value  $A_{ij}$ . From these simple coefficients, it is also possible to calculate vector coefficients starting from the scalar product of two vectors, as for instance, the Dice, cosine and Jaccard coefficients. In a vector space, the similarity between vectors  $x$  and  $y$  can be measured by the vector product  $x \times y = |x| |y| \cos \alpha$ , where  $|x|$  is the length of  $x$  and  $\alpha$  is the angle between the two vectors.

The Equivalence index is said also to be local and homogeneous. A local index is an index that does not use the number  $N$  of the size of the data. An association index is said to be more homogeneous if it remains unchanged when one multiplies all the variables by a constant factor. Local and homogeneous association indexes are more adapted to build a network of associations, if one does not know in advance the general structure of the studied field.

#### 4. CLUSTERING

The *third step* of the *co-site analysis* consists on decomposing the large network of associations into clusters. This section deals with the clustering algorithm that is used (section 4.1), the structure of obtained clusters (section 4.2), the analysis of the clusters (section 4.3), and the analysis of site cluster relationships (section 4.4.) While the cluster algorithm is a standard hierarchical agglomerative method following the single linkage procedure, the result is not a traditional single linkage hierarchy but a network of clusters linked together by inter-cluster associations. As we shall see this is the effect of the parameters of the used algorithm. We shall also examine more closely the structure of obtained clusters. In fact, an analysis of truly clustered sites will be based on this structure, and also on their characteristics.

Various uses of cluster analysis can be subsumed under four goals: development of a typology or classification, investigation of useful conceptual schemes for grouping entities, hypothesis generation through data explorations and hypothesis testing, or attempt to determine if types defined through other procedures are in fact present in a data set. Among these goals, the creation of classification probably accounts for the most frequent use of clustering methods

(see Aldenderfer and Blashfield, 1990, p. 9; Theodoridis and Koutroumbas, 1999, p. 354-355.)

Considering more common applications in which clustering is used, we find that our orientation consists here in data reduction and hypothesis generation. When the size of available data  $N$  is very large, cluster analysis can be used to group the data into a number of clusters  $m$  ( $\ll N$ ), and to process each cluster as a single entity. This is called data reduction. Cluster analysis is applied to a data set in order to infer some hypotheses concerning the nature of the data. These hypotheses must then be verified using other data sets.

Classification and clustering are distinguished as two different tasks in data mining. Classification is said to be a learning function that classifies a data item into one of several predefined classes, and clustering is considered as a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data (see Fayyad et al, 1996, p. 12-16). On the other hand, the term classification is often used in two different senses. One sense is to assign a new object/case to one of an existing set of possible classes. It is called supervised classification. Classification is also used in the sense of finding the classes themselves from a given set of unclassified objects/cases. It is called unsupervised classification (as it is used by Chesseman and Stutz, 1996.) In our case we can say that we apply an unsupervised classification method.

#### **4.1 Clustering Process**

The algorithm used is an adaptation of the single-link clustering in accordance with readability criteria on the size of the cluster (minimum and maximum number of items belonging to it), and on the maximum number of item associations constructing the cluster.

The algorithm is as follows:

1. Initially, each item is considered as a cluster.
2. The list of item pairs, sorted by decreasing value of the Equivalence coefficient, is examined sequentially to build the clusters.

3. If both elements of a given pair belong to the same cluster, the association between the items is considered as an internal association of that cluster.
4. If they belong to two different clusters, the algorithm tries to aggregate the clusters into one by merging them.
5. This is authorized if the size of the resulting cluster complies with the readability criteria.
6. Otherwise, the association is taken to be an external association.

Three saturation options are available when an aggregation fails because of the readability criteria: [1] forbid any new aggregation for these two clusters, [2] forbid any new aggregation of the larger of these two clusters, [3] do nothing

The user can modify the parameters to compute associations and to construct clusters. The goal here is to find a compromise between good readability of the results and what we accept to lose in terms of information. The clustering parameters for this particular study are the following:

1. Saturation strategy, i.e. to saturate the largest cluster ...
2. Minimal size of clusters ...
3. Maximal size of clusters ...
4. Maximal number of (internal and external) associations ...
5. Maximal number of external associations ...

Clusters are formed by associated sites and correspond to their *internal associations*. On the other hand, clusters can have relations between each other: these are said to be *external associations*. An *external association* occurs when there is an association between two sites belonging to different clusters and if the size of the new cluster, obtained by the union of the two initial clusters considered, is greater than the cluster maximal size authorised by the readability criteria.

If a site has at least a couple of outgoing links corresponding to an internal or external association of a cluster, then this site is related to this cluster.



In the case of our Web data, the number of obtained clusters is 37. A cluster represents a class, a category, of associated sites. One can investigate the nature of these associations if



one analyzes the *information content* that the associated sites represent. This information content requires another level of analysis which as we have already said is known as Web content mining.


## 4.2 Cluster Structure

The sites appearing in internal associations are called *internal sites*. The number of internal sites defines the size of a cluster. Those sites rejected during clustering because they do not meet the “maximum cluster size” criterion are recorded as *external sites*. Each site has a weight indicating its centrality in the cluster. For a given cluster  $C1$ ,  $n$  being the number of internal and external associations and  $F_i$  the number of occurrences of site  $i$  in the associations, the weight  $W(i)$  of site  $i$  of cluster  $C1$  is defined by  $W(i) = F_i/n$ . The internal site with the maximal value is chosen to name the cluster automatically. Note this is only a label suggested by the algorithm. The user may change the name.

### Figure 1

In figure 1  the structure of a cluster  is shown.

The Equivalence indices (the values) of the internal associations describe the strength of the sites associations defining the internal structure of a cluster. A cluster can be graphically represented as a graph. In order to have an indicator of its degree of cohesiveness that is called Density, the mean value of the internal associations is used. The density of  $C1(i)$  is for instance . The external associations are the associations existing between the sites of the cluster (internal sites) and sites belonging to other clusters (external sites). The mean value of the external associations of a cluster that is called Centrality is an indicator of its degree of dependence with regard to other clusters. The centrality of  $C1(i)$  is . (Note that these notions of *Density* and *Centrality* are different from the once in graph theory).

The saturation threshold of a cluster is the *Equivalence index* of the last internal association added before the cluster becomes saturated. The saturation threshold of  $C1(i)$  is . This value characterizes the relationship between density and centrality of a cluster. The centrality index

of  $Cl(i)$ , for instance, is below its saturation threshold, showing that this cluster can be extended to  $Cl(j)$ . In next section, we will show that the saturation threshold provides a important information for interpreting interrelations between clusters.

The number of external associations authorised for a given cluster may be limited. This is one parameter of the application. Thus, the external associations are not necessarily bi-directional. We introduce the idea of *reference* to indicate how many times the sites of a cluster appear in the external associations of other clusters. When a cluster refers to another one by its external associations, the latter is said to be referenced by the former as a related item of information. Here,  $Cl(i)$  is referenced ..... times by other clusters indicating that its influence goes beyond the field described by the sites of its cluster.

A cluster gathers together not only its forming internal sites, but also the source sites contributing to the cluster. A relevance weight should be computed for each source site related to the considered cluster. This is the sum of the site weights ( $W(i)$ ) of the sites belonging to both, the set of outgoing links of the source site, and the internal sites cluster set, divided by the total number of outgoing links present in the source site. It is possible to observe some source sites with multiple values of relevance weight. Indeed, as the same source site contributes often to two and plus clusters, it can have a different calculated value of relevance weight in each cluster.

Additional information can also be assigned to clusters if this information is in the data set. Such could be the number of outgoing links, the number of outgoing links in English, the number of Web pages inside the site, the number of Web pages in English, the number of images, the number of audio files and the number of video files inside the site. All this content information has been collected for our data and therefore it could be used but however it is not done here.

### 4.3 Analysis of Clusters

A cluster represents a class, a category, of associated sites. The characteristics of the 37 clusters obtained are the following:

- [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 with other clusters
- [8] Number of times a cluster is referenced by others
- [9] Number of sites related to a given cluster
- [10] Number of sites exclusively related to the clusters

CLUSTERS	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
ed.ac.uk	0.922	0.925	0.918	7	4	10	10	13	708	1
shef.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
ruu.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
jyu.fi	0.887	0.885	0.888	7	5	10	6	5	675	0
upc.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-ravensburg.de	0.878	0.895	0.878	10	1	17	1	14	506	0
ualm.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
bton.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
nhl.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
unav.es	0.856	0.861	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
unile.it	0.759	0.775	0.762	7	3	10	6	0	505	0
esc-clermont.fr	0.749	0.803	0.752	6	6	10	9	0	80	0
unilim.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
unirioja.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

**Table 1:** The 37 clusters and their numeric characteristics

Table 1 shows the 37 clusters in rows with their 10 numeric characteristics in columns. Column [1] permits to identify the order in which the clusters have been "frozen" during clustering. It is used in combination with column [3] for analysing cluster relationships (see below section 4.4). The values of columns [2] and [3] are used to plot clusters in a two-dimensional space representation. To get a more detailed idea of the structural diversity of clusters, a connection can be made between these mean values [2] and [3], and the number of internal and external associations [6] and [7] of each cluster.

The cluster size [4] is the number of distinct sites appearing in the internal associations [6] and its mean value [2] represents the density of the cluster. This characterises cluster cohesion. The sum of values of column [4] gives the number of sites kept in the clusters. Here 242 sites appear in the 37 clusters. This can be compared with the initial number of sites (791) to evaluate the "data reduction", corresponding to 30% in this case.

The number of external associations [7], the mean value of these associations [3], the number of external sites involved in these external associations [5], and the number of times a cluster is referenced by the others [8] give an idea regarding the role it plays within the network of clusters describing a certain Web domain (see below section 4.4).

Columns [9] and [10] indicate the number of sites related to each cluster. Since site classes can overlap, the total number of sites related to a given cluster [9] is not the same as the number of sites exclusively associated to that cluster [10]. In the obtained results we have 10 sites related to only one cluster. Therefore, the 760 remaining sites are associated to at least 2 clusters. The sum of values of [9] permits to calculate the average number of clusters with which a site is related to: in our case, this value is equal to 24.45. Overlaps like this are indicators of "content" relationships. More of 97% of the sites in the initial data set of 791 sites are covered by the 37 clusters. We may stress that we have obtained a manageable number of items (37 clusters) without losing too much site information.

### 5.3 Analyzing Cluster Relationships

Co-site analysis is not only a method to classify Web sites in clusters representing a category. It also provides the possibility of analyzing the associations between clusters. This analysis relies on the distinction between internal and external associations, the notion of cluster saturation threshold, and the size of the clusters. In this section, we propose the following procedure of analysis (Grivel et al., 1995).

[A] are those clusters whose external association mean value is higher than the saturation threshold; external associations are as strong as most internal associations.

[B] are those clusters whose external associations mean value falls below the saturation threshold; internal associations are much stronger than external associations. In this latter category, we distinguish between those clusters whose external associations are relatively strong [B1] from those ones whose external associations are very weak [B2].

The table 2 describes the two categories of clusters A and B and the subcategories B1 and B2.



Clusters of category A identify sites that are secondary insofar as they are of weak internal cohesiveness, whereas their associations with other clusters are relatively strong; they seem to be sub categories of these clusters. Clusters of category B1 could be qualified as mainstream if their internal associations are numerous and relatively strong. A typical example is ... Clusters of category B2 represent periphery because the external associations are very weak. In this category, ... is a good example of such a cluster.

A	B1	B2
uia.es	unilim.fr	unirioja.es
teilar.gr	tvu.ac.uk	iulm.it
sghms.ac.uk	ull.es	wales.ac.uk
esc-clermont.fr	nhl.nl	hua.gr
unile.fr	u-strasbg.fr	ense.fr
khs-linz.ac.at	ba-ravensburg.de	eap.fr
iniv-mlv.fr	brookes.ac.uk	
unav.es	hb.se	
lamp.ac.uk	rwth-aachen.de	
bton.ac.uk	casa-unimo.it	
ub.es	ed.ac.uk	
uni-potsdam.de		
ualm.es		
th-darmstadt.de		
upe.es		
jyu.fi		
uni-bonn.de		
kuleuven.ac.be		
ruu.nl		
shef.ac.uk		

**Table 2:** The categories of clusters

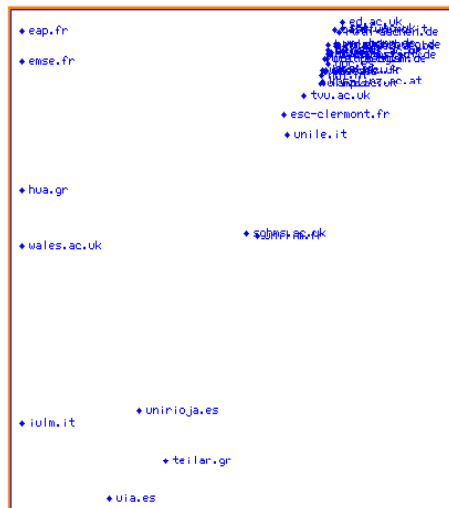
## 5. MAPPING

The *fourth and last step* of the *co-site analysis* consists on placing the obtained site clusters on a two-dimensional map. Map displays of clusters and their relationships can provide the analyst with an insight that is virtually impossible to get from looking at tables of clusters. In this section, we expose the map construction, the analysis of the map significance, and the analysis of cluster relationships on the map. This section deals with the method of construction of a map that is known also as *strategic-diagram*, and the significance of such a map as an analytical tool. We shall emphasize map displays not only as visualization tools but also as tools for analysis to confirm a hypothesis. In order to support a user friendly access to the results, a graphical hypertext interface displays the map of clusters. The user can thus explore and browse them through a Web navigation.

### 5.1 Constructing Maps

The measures of *density* and *centrality* allow the visualization of clusters and their relationships in a two-dimensional space: a map, where the X-axis corresponds to *centrality* and the Y-axis to *density*. To avoid recovering clusters having similar coordinates on the map, it is possible to plot the clusters by rank along these two axes. The scope of this map is to allow the user to understand the global and the local structure brought out by the clustering method from the associations matrix.

On the map (see figure below), the 37 clusters are arranged along the vertical Y-axis by order of increasing mean value of internal associations (density), and along the horizontal X-axis by order of increasing mean value of the external associations (centrality).



## 5.2 Map Significance Analysis

Each cluster has certain significance within the studied field expressed by its coordinates on the two axes. The fact that two clusters appear close to one another in the map does not mean that they are closely associated with one another. It only means that their values of density and centrality are similar.

The proximity of two clusters on such a map shows that they are structurally close, but it does not indicate any semantic closeness or “cognitive resemblance” (Peters et al., 1995). The question of “cognitive resemblance” is related to the content analysis of the sites of the clusters.

The maps are not only a mean of visualization, but they also represent an analytical method insofar as they can be used to evaluate the relative positions of the clusters in the geometric representation space. The higher a cluster is located on the Y-axis, the more coherent unit of information it is. The farther right it is on the X-axis, the greater are its links to other clusters.

The original use of the so-called "strategic diagram" proposes the following procedure of analysis (Callon et al, 1983).

Type 1	Type 2	Type 3	Type 4
ed.ac.uk		eap.fr	unirioja.es
shef.ac.uk		emse.fr	iulm.it
casa-unimo.it		hua.gr	teilar.gr
rwth-aachen.de		wales.ac.uk	uia.es
uni-bonn.de			
ba-ravensburg.de			
ruu.nl			
kuleuven.be			
ualm.es			
brookes.ac.uk			
jyu.fi			
hb.se			
th-darmstadt.de			
uni-potsdam.de			
u-strasbg.fr			
upc.es			
univ-mlv.fr			
bton.ac.uk			
unav.es			
ub.es			
nhl.nl			
khs-linz.ac.at			
ull.es			
lamp.ac.uk			
tvu.ac.uk			
esc-clermont.fr			
unile.it			

sghms.ac.uk			
unilim.fr			

Four types of clusters are distinguished: clusters with high density and centrality (type 1), with low density and high centrality (type 2), with high density and low centrality (type 3), and clusters with low values on both axes (type 4). This typology can be used to assess the strategic interest of clusters. In this kind of analysis, the mainstream clusters in the studied field should be represented by those clusters having the highest values on both axes. Clusters of type 2 are considered to correspond to a central position in the future. Clusters of type 3 are specialised clusters while clusters of type 4 are both peripheral and weakly developed and represent the margins of the network. The strategic diagrams are generally used to study the life cycle of the clusters. A case study can be found in (Callon et al., 1991) and farther analysis in (Callon et al., 1993, see pages 84-95).

## 6. INDICATORS

Indicators discuss in this section are relational indicators. They are designed to measure relationships and interaction between fields of activity or items. In quantitative studies of science and technology, relational indicators were designed to make possible to measure and assess whether static or dynamic relationships between scientific or technological specialties and domains exist. This is our orientation (translating the co-word method into the co-site analysis) that we have followed. These indicators are specific to the applied method of clustering and mapping. Now, they can be used in the co-site analysis too. They represent good candidates to become indicators for an analysis of information society, which is a new context of their use.

We call "information society" the set of social practices that were emerging (sociological context) from the Internet and today popularized via the Web (technological context). These social practices can be in the fields of education and culture, social life, health, economy, policy and so on. Information tends to penetrate almost all human activities.

In our analysis, the indicators are *centrality* and *density indexes*, the *structural index*, and the index of *cluster transformation* over the time. These indicators arise from the co-site analysis that generates clusters and maps such as those that we have exposed. They have been formulated and largely used by French authors of co-word analysis.

The *centrality index* measures the number of associations between one cluster and other clusters in the field. The more central a given cluster is, the more it appears to be an obligatory passage point for others in the field.

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

A *structural indicator* is the ratio of centrality to density. A low value of this indicator often indicates that a cluster can be cohesive but located at the periphery of the network. A strong value of this indicator often indicates that a cluster can be central and thus strategic but with little cohesion.

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 that two subsequent clusters have in common by the total number of items of the two clusters:

$$T_i(t, t+1) = \frac{Cl_t(i) \cap Cl_{t+1}(i)}{Cl_t(i) + Cl_{t+1}(i)}$$

## 7. CONCLUSION

Our objective was to apply the association analysis implemented as clustering and mapping on the field of Web structure mining. A hypertext user interface provides the tool for exploring and visualizing maps and clusters. Our intention is to provide the analysts of the Web with a working environment to support their own discovering processes.

Two possibilities of using the clustering and mapping method implemented by SDOC were illustrated. The first one is to give an easy and manageable access to a large Web network. The second way is to use a map as a means of analyzing information. In addition to the traditional way of analyzing in terms of centrality and density indexes, we have introduced the analysis of clusters relationships taking into account some further important parameters of clustering: the saturation threshold, the size of clusters and the number of associations. Since this approach avoids some interpretation problems due to criteria of cluster size, it provides a more adequate interpretation of inter-cluster links.

The structure of associations among key sites within clusters and inter-clusters establishes a pattern for a given Web domain. This pattern may then be observed to change through time. Through the study of these changing patterns, co-site analysis provides a tool for analyzing the development of Web and for assessing the degree of interrelationship between sites in the Web.

## 8. REFERENCES

- M. S. Aldenderfer and R. K. Blashfield, *Cluster Analysis*. Newbury Park, SAGE Publications, 1990
- P. Chesseman and J. Stutz, Bayesian Classification (AutoClass): Theory and Results, in U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (editors), *Advances in Knowledge Discovery and Data Mining*. Menlo Park, AAAI Press – The MIT Press, 1996, p. 153-180.
- H. Clifford and W. Stephenson, *An Introduction to Numerical Taxonomy*, New York, Academic Press, 1975.
- B. Everitt, *Cluster Analysis*, New York, Halsted, 1980.
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth From Data Mining to Knowledge Discovery: Overview, in U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (editors), *Advances in Knowledge Discovery and Data Mining*. Menlo Park, AAAI Press – The MIT Press, 1996, p. 1-34

W. J. Frawley, G. Piatetsky-Shapiro, C. J. Matheus, Knowledge Discovery in Databases: An Overview, in G. Piatetsky-Shapiro and W. J. Frawley (editors) *Knowledge Discovery in Databases*, MIT Press, 1991, p. 1-27.

L Grivel, P. Mutschke, X. Polanco, Thematic Mapping on Bibliographic Databases by Cluster Analysis: a Description of the SDOC Environment with SOLIS, *Knowledge Organization*, vol. 22, num. 2, 1995, p.70-77.

J. Hartigan, *Clustering Algorithms*, New York, John Wiley, 1975.

R. Kosala and H. Blockeel (2000) Web Mining Research: A Survey, *SIGKDD Explorations*, vol. 2, issue 1, p. 1-15.

H. P. F. Peters, R. R. Braam, A. F. J. van Raan, Cognitive Resemblance and Citation Relation in Chemical Engineering Publications, *Journal of the American Society for Information Science*, vol. 46, num. 1, 1995, p. 9-21.

P. H. Sneath and R. R. Sokal, *Numerical Taxonomy*, San Francisco, W. H. Freeman, 1973.

S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, San Diego, Academic Press, 1999.

### **Bibliographical note about the co-word analysis**

M. Callon, J-P. Courtial, W. A. Turner, S. Bauin, From Translation to Problematic Networks: An Introduction to Co-word Analysis, *Social Science Information*, vol. 22, 1983, p. 191-235. It is the first publication on the co-word analysis.

M. Callon, J. Law, A. Rip (editors), *Mapping the Dynamics of Science and Technology*. London, Macmillan Press, 1986.

B. Michelet, *L'analyse des associations*. Thèse de doctorat de l'Université de Paris VII, 1988. This work that represents the theoretical foundation of the Equivalence coefficient, becomes the source of many developments of the co-word analysis in France.

In the Handbook of Quantitative Studies of Science and Technology, edited by A.F.J. Van Raan (Amsterdam, North-Holland, 1988), the co-word analysis is referenced in two chapters: A. Rip, Mapping of Science: Possibilities and Limitations (p. 253-273), and W.A. Turner, G. Chartron, F. Laville, and B. Michelet, "Packaging Information for Peer Review: New Co-Word Analysis Techniques" (p. 291-323).

M. Callon, J-P. Courtial, F. Laville, Co-word Analysis as a Tool for Describing the Network of Interactions between Basic and Technological Research: The Case of Polymer Chemistry, *Scientometrics*, vol. 22, num. 1, 1991, p. 155-206. This article exposes one of the firsts more important studies using the co-word analysis.



J-P. Courtial, Introduction à la scientométrie. Paris, Anthropos – Economica, 1990. This book provides a pedagogical presentation of the co-word analysis in the context of the quantitative studies of science and technology

M. Callon, J-P. Courtial, H. Penan, *La Scientométrie*. Presses Universitaires de France. Collection : Que sais-je ? 1993. This is a scholar presentation of this discipline

M. Callon and J. P. Courtial, Using scientometrics for evaluation, in M. Callon, Ph. Larédo, Ph. Munstar (editors) *The strategic management of research and technology*. Paris, Editions Economica International, 1997, p. 165-219. This book takes stock of the methods and tools that are being developed and used in Europe today to ensure a strategic management of research and technology.

### **References about the citation analysis and co-citation analysis use in the Web analysis:**

S. Lawrence, C. L. Giles, K. Bollacker, Digital Libraries and Autonomous Citation Indexing, *IEEE Computer*, vol. 32, num. 6, 1999, p. 67-71.

S. Chakrabarti, B. E. Dom, D. Gibson, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, J. Kleinberg, Mining the Web's Link Structure, *IEEE Computer*, August 1999, p. 60-67.

S. Chakrabarti, B. E. Dom, D. Gibson, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, Experiments in Topic Distillation, authors describes recent experiments using their CLEVER system, in [URI/EICSTES/Clever/algorithm.html](http://URI/EICSTES/Clever/algorithm.html)

J. M. Kleinberg, Authoritative Sources in a Hyperlinked Environment, *Proceedings of the 9<sup>th</sup> Annual ACM-SIAM Symposium on Discrete Algorithms*, January 25-27, 1998, San Francisco, USA, p. 668-677. See p. 672, and references 9, 10, 17, 18, 19, 20, 26. In the full version of this paper that is available at in PDF, see p. 18-19, and references 27, 37 (in which the co-citation is used as a measure of similarity of Web pages), 45, 47, 54, 56. References 47 and 56 applied algorithm used in reference 54.

Authors have been used the term "sitation" to designate the directed links between sites on the Web (Rousseau 1997 quotes McKiernan 1996 and Aguillo 1996).