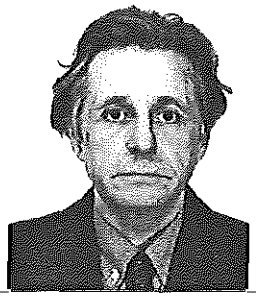
Luc Grivel*, Peter Mutschke, Xavier Polanco***
*Institut de l'Information Scientifique et Technique (INIST), Nancy
**Informationszentrum Sozialwissenschaften (IZ), Bonn

Thematic Mapping on Bibliographic Databases by Cluster Analysis: A Description of the SDOC Environment with SOLIS

Grivel, L., Mutschke, P., Polanco, X.: **Thematic mapping on bibliographic databases by cluster analysis: A description of the SDOC Environment with SOLIS.**
Knowl.Org. 22(1995)No.2, p.70-77, 13 refs.
The paper presents a coword-analysis-based system called SDOC which is able to support the intellectual work of an end-user who is searching for information in a bibliographic database. This is done by presenting its thematical structure as a map of keyword clusters (themes) on a graphical user interface. These mapping facilities are demonstrated on the basis of the research field Social History given by a set of documents from the social science literature database SOLIS. Besides the traditional way of analysing a coword map as a strategic diagram, the notion of cluster relationships analysis is introduced which provides an adequate interpretation of links between themes.
(Authors)

Luc Grivel (b.1962) is a computer scientist, graduated in 1985 from the University of Nancy. Since 1990 working in the informatics research program at INIST, CNRS. Heads now the SDOC project and another one on hypertext databases dedicated to scientific watch. Author of several articles.

Peter Mutschke (b.1961) is a computer scientist at the research department of the Informationszentrum Sozialwissenschaften (IZ). Subjects of interest: knowledge-based and object-oriented systems, esp. regarding large bibliographic databases. Author of several articles.



Xavier Polanco is head of the Informetrics Research Program at INIST, CNRS. More than ten years of experience in scientometrics and information science. Editor of a book on the rise and development of world-science (1990) and author of numerous articles in technology and science studies. Visiting appointments in Europe and America.

1. Introduction

Bibliographical information in public databases are, as Brookes (2,p.9) says, "abundantly generated and systematically stored but not yet efficiently used". The present paper addresses the problem of an end-user who is searching for information in a database. Usually, he needs to get an idea of the state of the art in his special domain of interest. In order to support the intellectual work of analysing retrieved documents in this respect, a coword-analysis method has been developed which discovers the thematical structure of a database and presents it as a map of themes on a graphical user interface. The SDOC-system from INIST (Institut de l'Information Scientifique et Technique) is an implementation of this method, and aims at mapping scientific research fields in large databases. Our goal is to demonstrate the thematic mapping facilities of SDOC with a German bibliographical database, in this case the SOLIS database of the Informationszentrum Sozialwissenschaften. SOLIS provides information mainly about German-language scientific literature, journal articles, contributions in compilations, monographs, and "grey literature".

Document-based retrieval systems normally use an indexing vocabulary to describe the content of its documents, and an online system to access these documents. The output of such a system in response to the user's query is a set of individual references. In this study, we imagine a French user who is searching for information in SOLIS concerning the field of social history in Germany. He selects all the literature processed over a three-year period (1989-90-91) in the SOLIS database having 'social history' as primary or secondary classification code and indexed by the keyword 'Germany'. This yields 285 bibliographical references. Traditionally, the user could only browse sequentially these documents with the difficulty of determining the importance of the topics and the links between them. By examining the indexing vocabulary, he can define certain topics manually and search for related documents. But even if the sample is not big, this iterative process is long and fastidious. The problem faced by all users of information systems is the need to reduce the amount of information to a manageable number of items to be examined.

SDOC belongs to a family of methods which use term associations and clustering techniques to solve this problem. Callon, Courtial, Turner and Bauin (3) call it "coword analysis" and Salton (12) "term clustering". This technique was early used in the SMART automatic document retrieval system (11). The use of term associations in automatic information retrieval has been studied since a long time, whereas coword analysis[1] has been implemented in the eighties into the LEXIMAPPE program to highlight the dynamics of scientific and technical development. In the latter context, cowords are used for identifying and visualizing the centres of interest in scientific literature by means of coword maps (3).

Like LEXIMAPPE, SDOC[2] produces a classification of themes, i.e. clusters of closely tied keywords, characterizing the domain studied, which can be the complete database or a subset of it referring to a special query. Such clusters are structured internally by means of relation-

ships between the keywords of a cluster, and externally by interrelations between different clusters. The topics are visualized in a two-dimensional space or *Thematic Map* according to the semantic strength of their internal (*Density*: Y-axis) and external associations (*Centrality*: X-axis). Figure 1-1 shows an example of such a map of themes obtained from the 285 retrieved documents, saying, for instance, that *German Question*[3] was a central and intensively discussed theme of Social History research 1989-92. In this way, the user obtains an aggregation of thematic information. Furthermore, SDOC generates a hypertext system. Thus, the user can navigate through the generated knowledge space (map of themes). SDOC is described more detailed in Section 2.

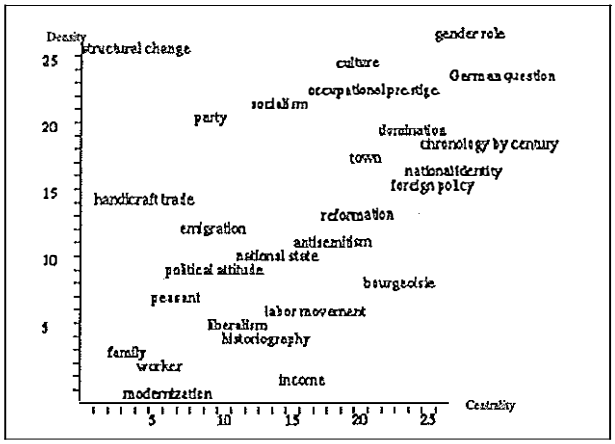


Figure 1-1: A General Map of Social History Themes

On the basis of such thematical maps two types of *information analysis* can be considered: One is the analysis of the thematic structure of the database itself ("What is in the database?"), the other is the observation of the research field ('Who does what, where and when?'). A researcher or teacher in social history at least needs to know the thematic structure of the database he is consulting to satisfy his information request. The role of information analysis here is to provide the user with a state of the art of a certain domain of interest, in order, for instance, to get its most relevant scientists or journals, or to compare the scientific discussion in different countries (10).

In this paper, we will focus on the analysis of the thematic structure of the database. By applying SDOC to the SOLIS data file (see Section 3), we want to demonstrate how this tool can be used to support this kind of analysis on the basis of bibliographical data.

2. Thematic Mapping

2.1. Coword Analysis

Coword analysis used in SDOC is an analytical method for identifying and visualizing the centres of interest in scientific literature (3). The method is founded on the use of keywords as indicators of information content. The essential concept is the co-occurrence of content-describing keywords belonging to the same document. It is based on the idea that two keywords i and j which are used together in the description of a single document are related. It is clear that the co-occurrence value Cij (number of co-occurrences of words i and j in a given set of documents) is not the best measure of the strength of a keyword association because very frequently used keywords have an advantage over those used less often. In order to normalize the proximity value of keyword pairs the *Equivalence index* Eij = Cij2 / (Ci * Cj) (square of Ochiai index also called Salton index) is used, where Ci is the frequency of i and Cj the frequency of j in the data set. The keyword *German question*, for instance, co-occurres three times with the keyword *reunification*; thus, their association has an Equivalence index of 0.3, since *German question* has a frequency of ten, whereas *reunification* appears only three times in the datafile.

2.2 SDOC's clustering process

These weighted coword-relations are the basis to construct a thematic representation (keyword clusters) of scientific areas and the relationships between research themes. The clustering-method aims at aggregating the keywords into groups of closely linked keywords. The algorithm implemented in SDOC is an adaptation of the single-link clustering in accordance with readability criteria: size of the cluster (minimum and maximum number of keywords belonging to it), and the maximum number of keyword associations constructing the cluster. The algorithm used is the following: Initially, each keyword is considered as a cluster. The list of keyword pairs, sorted by decreasing value of the Equivalence index, is examined sequentially to build the clusters. If both elements of a given pair belong to the same cluster, the link between these keywords is considered as an internal association of that cluster. If they belong to two different clusters, the algorithm tries to aggregate the clusters into one by merging them. This is authorized if the size of the resulting cluster complies with the readability criteria. 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 following example (see Figure 2-1) illustrates the building of the clusters *German Question* and *Foreign Policy* including their relationships (the links are valued by the Equivalence index of the respective keywords association). At a given time, *German Question* is composed of the links *Berlin <-> cold war, Berlin <-> reunification, cold war <-> german question, Berlin <-> german question, reunification <-> german question, german question <-> policy of detente, policy of detente <-> security policy, policy of detente <-> international relations, reunification <-> SED and GDR <-> SED*; the cluster *Foreign Policy* is only defined by *german policy <-> foreign policy*, and there is no link between these clusters. When the algorithm examines the associations *security policy <-> foreign policy* and *security policy <-*

> *german policy*, the two clusters can not be merged because of the size criteria. Therefore, these links are stored as external associations. Each further association between keywords of *German Question* and *Foreign Policy*, such as *german question* <-> *german policy*, is represented as external link.

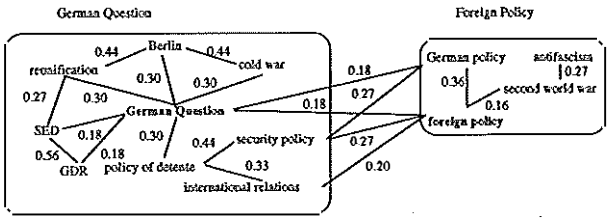


Figure 2-1: The building of clusters German Question and Foreign Policy

The user can modify the parameters used to compute the associations and construct the clusters. The goal here is to find a compromise between good readability of the results (not too many clusters) and what we accept to lose in terms of information. The parameters for this particular study are put in parenthesis.

Indexing vocabulary:
Minimum frequency of keywords (2)
Suppression of keywords that are used too often (Germany)
Associations :
Selection of a statistical index (Equivalence index)
Minimum keyword cooccurrence (2)
Clustering :
Saturation strategy, i.e to saturate the largest cluster
Minimum size and maximal size of clusters (4 and 10 keywords)
Maximum number of internal and external associations (20)
Maximum number of external associations (10)

2.3 The Structure of a Cluster

A cluster represents a special theme or centre of interest found in a set of documents. The keywords appearing in its internal associations are called *internal keywords*. The number of internal keywords defines the size of the cluster. Those keywords rejected during the clustering because they do not meet the "maximum cluster size" criteria are recorded as *external keywords*[4]. Each keyword has a weight indicating its centrality in the cluster. For a given cluster C, N being the number of internal and external associations and Fi the number of occurrences of term i in the associations, the weight $W(i)$ of term i of cluster C is defined by $W(i) = Fi/N$. The internal keyword with the highest value is chosen to name the cluster automatically[5]. In the following the keywords defining the cluster *German Question* are shown:

Weight	Frequency	Keyword
0.47	10	German question
0.18	5	Socialist Unity Party of Germany (SED)
0.18	3	security policy
0.18	3	policy of detente
0.18	3	reunification
0.18	3	Berlin
0.12	9	German Democratic Republic (GDR)
0.12	4	international relations
0.12	3	cold war
0.18	5	foreign policy*
0.12	5	Germany policy*

The Equivalence indices of the *internal associations* describe the strength of the keyword associations defining the internal structure of a cluster. In order to have an indicator of its degree of cohesiveness (*Density*), the mean value of the internal associations is used (density of *German Question*: 0.34). The *external associations* are the associations existing between the keywords of this cluster (internal keywords) and keywords belonging to other clusters (external keywords). The mean value of the external associations of a cluster (*Centrality*) is an indicator of its degree of dependance with regard to other clusters (centrality of *German Question*: 0.22). 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 *German Question* is 0.27). This value characterizes the relationship between density and centrality of a theme. The centrality index of *German Question*, for instance, is below its saturation threshold, showing that this theme can be extended to *Foreign Policy*. The saturation threshold is therefore an important information for interpreting interrelations between clusters (see Section 3.4 Analysing Cluster Relationships).

The number of external associations displayed for a given cluster may be limited. This is one parameter of the application. Thus, the external associations are not necessarily bidirectional. We introduce the idea of *thematic reference* to indicate the number of times that keywords of one 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, *German Question* is referenced 13 times by other clusters indicating that its influence goes beyond the topic described by the keywords of the cluster (Section 3.3 illustrates these relationships).

Considered as a classification unit, a cluster gathers together not only keywords, but also a set of documents. A document is assigned to a cluster if it is indexed by a couple of two internal keywords or a couple of one internal and one external keyword of the cluster. A document may therefore belong to several clusters. A relevance weight is computed for each document. This is the sum of the weights of keywords in the cluster indexing the document, divided by the number of keywords belonging

to it. In the following, the documents dealing with the German Question topic are shown:

Weight:	Title:
0.14	The social-democratic intra-party discussion on security, detente and German unity
0.11	Between the Cold War and detente: security and Germany policy within the system of the allied powers in the years 1953-1956
0.11	From "civil war" to the responsible community
0.10	The four-sector city of Berlin in the German press 1945-1949
0.10	Attitude of the SED and the GDR towards German unity 1949-1987
0.08	The German policy of the government of the U.S.A. in preparation and during the course of the Potsdam Conference
0.07	The Socialist Unity Party of Germany (SED) and the national issue
0.07	Neither a hammer nor an anvil?: observations on the present-day situation in Germany (1973)
0.06	Contributions on the history of the Berlin democracy: 1919-1933/1945-1985
0.05	The Socialist Unity Party of Germany (SED) in history and the present age
0.05	The political obstruction to modernization in France during the interwar period
0.03	On the appearance of the first volume of the "History of the SED"
0.02	The German-Japanese relations during the Third Reich
0.02	The Socialist Unity Party of Germany (SED) and German history

Additional information such as a list of authors, a list of sources (journals, books etc.) or institutional affiliations, can also be assigned to the clusters if this information is in the bibliographical reference. The weight assigned to each item is the sum of the weights of the documents where the item appears.

2.4 Constructing Thematic Maps

The measures of Density and Centrality allow the visualization of themes and their relationships in a two-dimensional space (map), where the x-axis corresponds to Centrality and the y-axis to Density[6]. In order to support a consultation of the clustering results, SDOC integrates this map in a graphical hypertext-based user interface (s. Fig. 2-2).

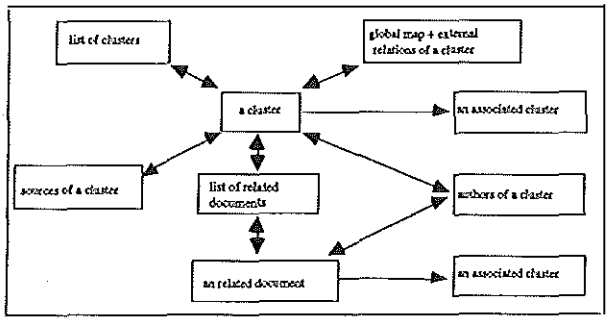


Figure 2-2: Browsing the organization of a topic, the key figures and the sources of information

The starting point for the navigation is the list of clusters sorted by the saturation threshold. This corresponds to the order in which they have been "frozen" during the clustering. The user selects the cluster name and points to its description. He can then examine: a) the characteristics of the cluster (number of documents, authors and sources, saturation threshold, density, centrality, number of citations by the other clusters); b) the characteristics of the keywords in the cluster (weight, frequency) and their associations (Equivalence index, co-occurrence); and c) the associated clusters including a description of the external associations involved.

3. Information Analysis of the SOLIS Datafile

3.1 The Indexing Vocabulary

Keywords are primarily used for information retrieval by Boolean queries. Here, they are used as content indicators to which the SDOC analysis is applied. The vocabulary indexing the 285 retrieved Social History documents consists of 892 controlled terms manually assigned on the basis of the Social Science thesaurus of the Informationszentrum Sozialwissenschaften. For this coword analysis, the English keywords of SOLIS are used, with the exception of the keyword "Germany", because, given the search query, this keyword yields no information. The 499 keywords of frequency 1, which represent 56 % of the indexing vocabulary, are excluded as input to the coword analysis. They complicate the keywords association network with potentially noisy information. So the effective number of keywords as input to the clustering is 392.

In order to analyse this datafile, we will first study the variables which characterize a cluster as an indicator of a research theme. Then we will focus on the use of the hypertext maps as a means to explore the thematic structure of the database by theme. Finally, we will analyse the cluster relationships.

3.2 Coword Clusters as Knowledge Indicators

Applying SDOC on the Social History document set provides 27 clusters in all (s. Fig. 1-1: A General Map of Social History Themes). Table 3-1 shows these clusters with the following characteristic data:

[1] Cluster saturation threshold
[2] Density
[3] Centrality
[4] Number of internal keywords
[5] Number of external keywords
[6] Number of internal associations
[7] Number of external associations with other clusters
[8] Number of thematic references of a subject by other topics
[9] number of bibliographical references related to the cluster
[10] number of bibliographical references exclusively related to the cluster

Name	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
Antisemitism	0.125	0.212	0.106	10	4	16	4	4	22	3
Bourgeoisie	0.133	0.185	0.129	9	6	12	7	24	89	1
Chronology by Century	0.200	0.296	0.160	9	7	13	7	38	188	21
Culture	0.173	0.376	0.122	10	5	14	6	5	19	0
Domination	0.118	0.296	0.131	8	5	10	8	10	25	1
Emigration	0.083	0.218	0.071	10	1	18	1	6	22	6
Family	0.111	0.148	0.033	4	9	3	10	1	12	1
Foreign Policy	0.160	0.262	0.143	4	7	3	10	6	11	1
Gender Role	0.213	0.527	0.196	8	2	18	2	5	10	2
German Question	0.267	0.337	0.219	9	2	12	5	13	14	0
Handicraft Trade	0.167	0.222	0.019	5	4	4	10	0	12	1
Historiography	0.082	0.163	0.086	8	8	7	9	2	18	2
Income	0.114	0.137	0.103	9	5	13	7	6	19	0
Labor Movement	0.091	0.169	0.096	9	8	10	10	7	46	1
Liberalism	0.062	0.166	0.079	7	6	6	9	2	18	1
Modernization	0.071	0.093	0.039	4	6	3	9	0	16	2
National Identity	0.188	0.289	0.147	9	2	14	2	8	19	6
National State	0.078	0.194	0.087	9	10	10	10	9	33	0
Occupational Prestige	0.190	0.315	0.115	9	6	12	6	8	18	0
Party	0.133	0.297	0.076	6	7	11	9	3	11	2
Peasant	0.089	0.184	0.060	7	7	9	10	0	14	0
Political Attitude	0.114	0.186	0.066	5	6	4	9	2	13	0
Reformation	0.111	0.221	0.121	8	3	14	6	4	13	0
Socialism	0.167	0.309	0.095	8	5	8	10	6	15	0
Structural Change	0.200	0.486	0.000	8	0	20	0	5	4	0
Town	0.113	0.289	0.124	10	6	12	7	16	60	1
Worker	0.067	0.142	0.057	6	8	6	8	1	15	0

Table 3-1: Characteristics of the 27 clusters obtained (in alphabetical order)

Column [1] permits to identify the order in which the clusters have been "frozen" during the clustering. It is used in combination with column [3] for analysing cluster relationships (see Section 3.4). The values of columns [2] and [3] are used to plot the clusters in a two-dimensional space representation. To get a more detailed idea of the structural diversity of the 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 keywords appearing in the internal associations [6] whose mean value [2] represents the density of the cluster. This characterizes the cohesion of the cluster. The sum of the values of column [4] gives the number of keywords kept in the clusters. Here 208 keywords appear in the 27 clusters. This can be compared with the initial number of keywords (892) to evaluate the "data reduction".

The number of external associations [7], the mean value of these associations [3], the number of external keywords 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 themes describing a certain research context (see Section 3.4 Analysing Cluster Relationships).

Column [9] and [10] indicate the quantity of bibliographic information relative to each cluster. Since document classes can overlap, the total number of documents classified in a given cluster [9] is not the same as the number of documents exlusively associated to that cluster [10]. The sum of the values of [9] gives the number of documents belonging to the clusters. In this case, there are 756 document cluster associations, whereas the total number of distinct documents in the clusters is only 266. Of these 266, 52 are related to exclusively one cluster. Overlaps like this are indicators of theme relationships. More than 93% of the documents in the initial file of 285 documents are covered by the 27 clusters. We may stress that we have obtained a manageable number of items (27 clusters) without losing too much bibliographic information.

3.3 Mapping Knowledge: A Hypertext System

On our maps (s. Fig. 3-1 to 3-4), the 27 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). Each cluster has a certain thematic significance within the studied research field expressed by its position on the two axes. The fact that two clusters appear close to one another in the information space (or map) does not mean that they are closely associated with one another. It only means that their values of centrality and density are similar.

The higher a cluster is located on the Y-axis, the more it is a coherent unit of information. The farther right it is

on the X-axis, the greater are its links to other clusters. The authors of the coword analysis method traditionally distinguish four types of clusters: clusters with high density and centrality (type 1), with a low density and high centrality (type 2), with high density while peripheral from the point of view of centrality (type 3), and themes with low values on both axes (type 4). Callon, Courtial, Turner and Bauin (3) call this representation "strategic diagram" and use this typology to assess the strategic interest of the themes. In this kind of analysis, the mainstream themes in the research field studied should be represented by those clusters having the highest values on both axes (type 1 in table 3-2). Clusters of type 2 may correspond to central themes in the future. Clusters of type 3 are specialized themes while clusters of type 4 are both peripheral and weakly developed and represent the margins of the network. This categorization should be cautiously used in collaboration with an expert of the domain. The strategic diagrams are generally used to study the life cycle of the themes. A case study can be found in (6).

Type 1	Gender Role, Culture, German Question, Occupational Prestige, Domination, Sixteenth Century, Town, National Identity, Foreign Policy, Reformation
Type 2	Antisemitism, Bourgeoisie, Labor Movement, Income
Type 3	Structural Change, Socialism, Party, Handicraft Trade
Type 4	Emigration, National State, Political Attitude, Peasant, Liberalism, Historiography, Family, Worker, Modernization

Table 3-2: Cluster categorization in a strategic diagram

Here, our use of the map is different. We use this representation to define an informational space or global context of research information where the local networks are highlighted, i.e. the associations between the clusters. The hypertext interface permits the user to follow the local networks of each theme (s. Fig. 3-1 to 3-4), and then to proceed to an analysis. If, for instance, he is interested in questions of nation and nationality in the framework of the *German Question,* he can see that this cluster (s. Fig. 3-1) is associated with one other cluster, *Foreign Policy.*

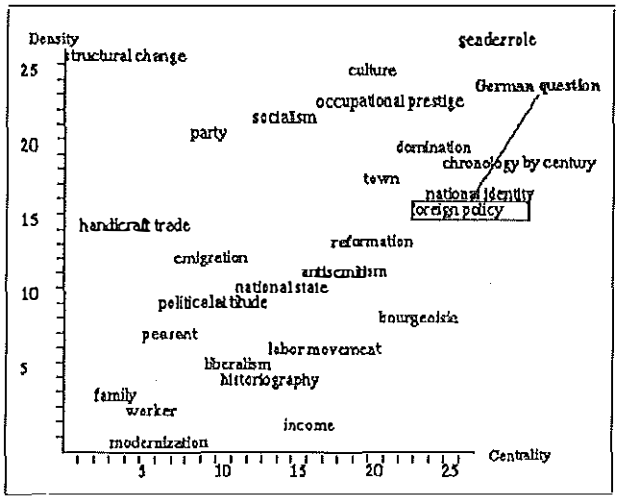


Figure 3-1: Cluster German Question

German Question and *Foreign Policy* are associated by way of five bidirectional associations (s. Fig. 2-1). The analysis of these associations shows that *Foreign Policy* is a subtheme of *German Question* because the saturation threshold of *German Question* is higher than the mean value of its external associations to *Foreign Policy*, and, vice versa, the strength of the external associations of *Foreign Policy* with *German Question* are higher than its saturation threshold. The relative position of Foreign Policy with respect to German Question (below, and more left) is an indicator but not a sufficient condition for the existence of such a relationship, because we need to know the saturation threshold and the strength of the external associations concerned. Figure 3-2 illustrates the local network of the theme *Foreign Policy.* Thus, the initial topic *German Question* is also associated with *National Identity, Labor Movement* and *Emigration.*

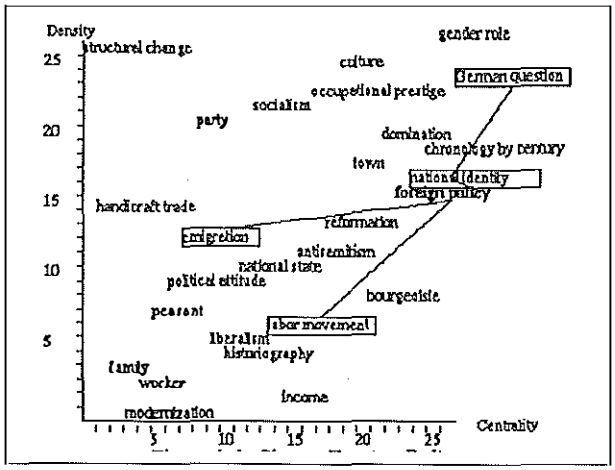


Figure 3-2: Cluster Foreign Policy

Suppose the user is now interested in the position of *National Identity.* Figure 3-3 shows that this topic is associated with the initial theme German Question, and refers to a new topic, *Socialism. National Identity contains the keywords: national identity, national consciousness, historical awareness, conception of history, German, Nazism, Hitler, Third Reich, nationalism.* It has external associations with German Question, by conception of history - Socialist Unity Party of Germany (SED), and with Socialism, by Nazism - socialist party.

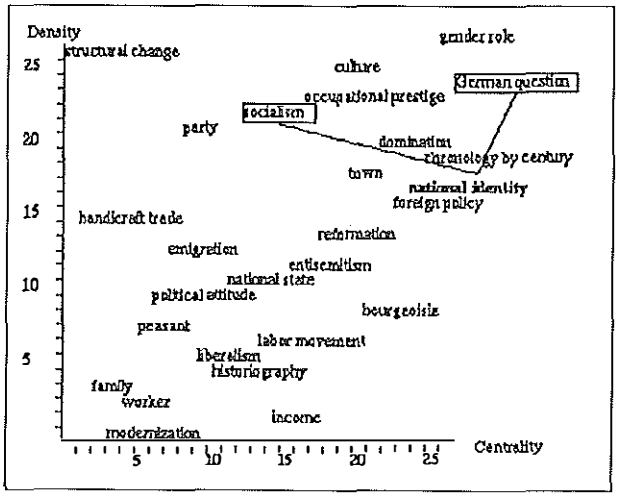


Figure 3-3: Cluster National Identity

The *Socialism* cluster refers back to *Labor Movement* and *National Identity,* and opens the network towards two other themes, *Party and Chronology by century* . Moving from one topic to another, the user explores the content of his data by examining a structured knowledge space. He can decide either to follow another informational network or to stop the navigation process and browse the literature aggregated under a topic.

3.4 Analysing Cluster Relationships

Coword analysis is not only a method for classifying bibliographical references in clusters representing a research theme. It also provides the possibility of analysing the associations between themes. This analysis relies on the distinction between internal and external associations, the notion of cluster saturation threshold, and the size of the clusters.

Table 3-3 describes two categories of clusters:

[A] those whose external associations mean value is higher than the saturation threshold, i.e. the external links are as strong as the most internal associations;

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

A	Domination, Town, Reformation, National State, Labor Movement, Liberalism, Historiography
B1	Gender Role, Culture, German Questio , Occupational Prestige, Socialism, Party, Sixteenth Century, Natio al Identity, Foreig Policy, Emigratio , Antisemitism, Political Attitude, Bourgeoisie, Peasant, Income
B2	Structural Cha ge, Ha dicraft Trade, Family, Worker, Modernization

Table 3-3: Categories of clusters

Clusters of category [A] identify themes which are secondary (in the datafile) insofar as they are of weak internal cohesiveness, whereas their associations with other clusters are relatively strong, i.e. they seem to be subthemes of these clusters. For instance, Liberalism seems to be secondary with respect to the theme *Bourgeoisie.* Furthermore, in this category of clusters, we can discover crossroad clusters (*Domination* and *Town*) which connect very heterogenous topics via one generic keyword (s. Fig. 3-4). Thus, crossroad clusters usually represent very generic research topics, which are crossing points of themes.

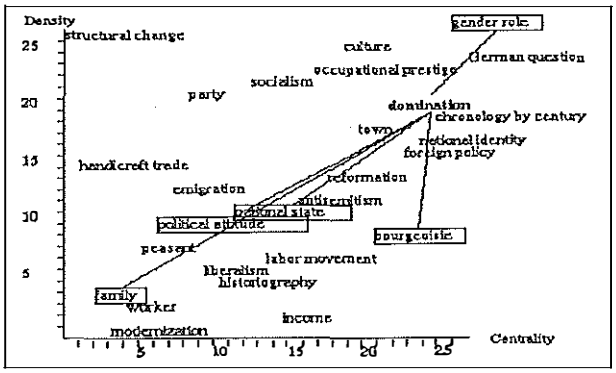


Figure 3-4: Cluster Domination: An example of crossroad cluster

Clusters of category [B1] could be qualified as mainstream themes if their internal associations are numerous and relatively strong. A typical example is *German Question* (s. Fig. 3-1 and 3-2) whose local network has been already studied. An analysis process should start with them because they are the main thematic nodes of the network.

Clusters of category [B2] represent peripheral themes because the links tying them to the network are very weak. In this category, *Handicraft Trade* is a good example of such a cluster. The only external associations it has are with *Chronology by century. Family, Worker* and *Modernization* have numerous but weak associations to other clusters. Since their internal structure is, moreover, very weak (see the number of internal keywords [6] and internal associations [7] in table 3-1), we consider them as peripheral themes. *Structural Change* is a special case, because it points out a theme with a strong density, i.e. a homogeneous research field, but without any association with other clusters.

SDOC visualizes such thematical networks in the form of maps. In other words, it maps the knowledge embedded in documents (thematic structure), but also the individual agents (authors, institutions) and the way they communicate. By considering the relationships between clusters, their internal structure and the less or more central role they play within a network of themes the importance of a certain thematic aspect for the research field studied can be examined.

4. Conclusion

In the present paper, two possibilities of using the mapping method of SDOC are illustrated. The first one is to give an easy access to distributed database information. In front of the thematic structure of the database content the user can define his own strategy of information search for the problem he has to solve. He may discover relations between themes he would not have thought of; and on this basis he can adjust his query. The second method is to use such Thematic Maps as a means of analysing information. Besides the traditional way of analysing a coword map as a strategic diagram, which reflects only two parameters characterizing the clusters (centrality and density), we have introduced the clusters relationships analysis taking into account further important parameters of the clustering: the saturation threshold, the size of the clusters, and the number of associations. Since this approach avoids some interpretation problems due to the criteria of cluster size, it provides a more adequate interpretation of links between themes.

Our objective was to implement an environment which offers the user a contextual view of the informational space contained in a set of bibliographical references, so that he can locate his demand of information more precisely. Since we are working at a level of indicators, we are not concerned with exactness. A specialist in the field will always have the final say concerning the results of an

automatic information analysis. Our intention is to provide him with a working tool to support his own information discovering process, with the possibility of going beyond his special subject in order to explore neighbouring domains. We believe that such an environment best arms the user to face the growing volume of information.

Acknowledgments: We are very grateful to our INIST and IZ colleagues, and particularly to M. Herfurth (head of the IZ research department), for their valuable comments.

Notes:

1 This method is an alternative to the well known tradition of citation analysis (9) and co-citation analysis (13); see (1) for a comparison of Co-Citation and Co-Word Clustering; see (7) and (4) for an introduction to scientometrics and scientific watch.

2 SDOC differs from LEXIMAPPE concerning technical characteristics: SDOC has been implemented in C under UNIX, in order to allow the treatment of very large data files, whereas LEXIMAPPE is for DOS- and McIntosh-systems. The modules of SDOC rely on a library of C-functions, developed at INIST, specialized in the treatment of any SGML document (8), so that SGML is used by SDOC both as a conversion format for the raw data as input and as pivot format for the intermediary data which are exchanged between the modules.

3 In the following, cluster names are printed in italics and start with an uppercase letter. Keywords are printed in italics, small letter size and lowercase letters.

4 indicated by a star in the example.

5 This is only a label suggested by our program. It may be changed if it is not felt to be appropriate to the cluster.

6 To avoid recovering clusters having similar coordinates on the map, the software also makes it possible to plot the clusters by rank along these two axes.

References:

(1) Braam, R.R., Moed, H.F., Raan, A.F.J.van: Comparison and Combination of Co-Citation and Co-Word Clustering. In: Select Proc. First Int. Workshop on Science and Technology Indicators, Leiden, 14-16 Nov. 1988, p.307-337

(2) Brookes, B.C.: The foundations of information science. Part. IV. Information science: The changing paradigm. J. Inform. Sci. 3 (1981)p.3-12

(3) Callon, M., Courtial, J.P., Turner, W.A., Bauin, S.: From translation to problematic networks: an introduction to co-word analysis. Soc. Sci. Inform. 22(1983)p.191-235.

(4) Callon, M., Coutial,J.P., Penan,H,: La scientométrie. Presses Universitaires de France. Collection: Que sais-je. Paris, 1993.

(5) Callon, M., Law, J., Rip,A. (Eds).: Mapping the dynamics of science and technology, London: Macmillan Press 1986.

(6) Callon, M., Courtial,J.P., Laville,F.: Co-word Analysis as a tool for describing the network of interactions between basics and technological Research: the case of polymer chemistry. Scientometrics 22(1991)No1, p.155-206.

(7) Desvals, H. Dou, H.: La veille technologique. Paris: DUNOD 1992.

(8) Ducloy, J., Charpentier, P., Francois, C., Grivel,L.: Une boite d'outils pour le traitement de l'information scientifique et technique. Génie logiciel et systèmes experts 25(1991)p.80-90, Paris.

(9) Garfield,E.: Citation analysis as a tool in journal evaluation. Science 178 (1972), pp 471-479.

(10) Polanco X., Grivel,L.: Mapping knowledge: the use of co-word analysis techniques for mapping a sociology data file of four publishing countries (France, Germany, United Kingdom and United State of America), 4th Int. Conf. of Bibliometrics, Informetrics and Scientometrics, 11-15 Sept. 1993, Berlin, Germany, (to be published in Informetrics).

(11) Salton,G.: The SMART retrieval system - Experiments in automatic document processing. Englewoods Cliff, NJ: Prentice Hall 1971.

(12) Salton,G.: Automatic text processing: the transformation, analysis and retrieval of information by computer. New York: Addison Wesley 1989.

(13) Small, H.: Co-citation in the scientific literature: A new measure of the relationship between two documents. J. Amer. Soc. Inform. Sci. 24(1973)p.265-269.

Address: Mr.Peter Mutschke, IZ Sozialwissenschaften, Lennéstr. 30, D-53113 Bonn.