SEMINAR ON THESAURUS (1975). Paper BA

AN AUTOMATICALLY PRODUCED THESAURUS

GERHARD SCHWANHAUSSER, Thesaurus Hochschulforschung Hochschulbau, Institut fur Hochschulbau, Sonderforschungsbereich 63, Stuttgart, Federal Republic of Germany

A procedure for the computer generation of a thesaurus from a set of descriptors, manually assigned to the documents in a library, is described. Recognises only a quasi-associative relationship among descrip - tors. The specific advantages of the thesaurus as an open-ended one, the keywords derived from actual documents being the most helpful in retrieving the documents, and as an aid in information search in the collection are pointed out. Computational procedures for generating the thesaurus include keyword statistics, matrix inversion, calculation of similarity matrix using Tanimoto coefficient, automatic cluster analysis using minimal-tree procedure, and compilation of groups and main groups of descriptors are given. An algorithm for the graphic display of the main groups of descriptors has been formulated. The main disadvantage of the procedure is that only a limited number of keywords can be processed within a reasonable computer CPU time. Points out that the procedure can be applied to a library of 10, 000 to 20, 000 documents with a keyword base of 1, 000 words using about 3 hours of computing time.

1 GENERAL CONSIDERATIONS

The basic prerequisite for the establishment of an automatically constructed thesaurus is the existence of an indexed corpus of documents. In the case of the library of the Zentralarchiv fur Hochschulbau, Stuttgart, there has been available a corpus of 5, 900 manually indexed documents (1). We suppose that, even if the indexing has been done carefully it is still a rather subjective work and we should use the minimum of further assumptions to produce a thesaurus, which we call, on account of its iater-indexer inconsistency, a pragmatic thesaurus (3), Defining thesaurus as an interrelated list of key words of the area of knowledge which is under consideration, the question can be raised how to limit the area of knowledge and haw to decide which key words do belong to this area, since new branches of science develop their vocabularies fairly rapidly and even changes of meaning occur as time goes on. A pragmatic thesaurus such as ours implies some basic assumptions. Firstly, we assumed every document entered into the library is relevant £or this area of knowledge- Secondly, the manually assigned key words for every document are taken to be the most relevant ones. Thirdly, there is only one type of interrelation between the key words, namely a quasi-associative relation of every key word of a document to all other key words of that document. Having adopted these assumptions to a special library you need but to do some machine calculations and a thesaurus is almost ready, A procedure such as this facilitates the up-dating of a thesaurus, whenever sufficient additional documents have been entered into the library. A detailed specification of the area of interest is not necessary, since the indexed document collection can be understood as a description of it. In addition to that the list of key words need not - should not - be fixed in advance. It should rather be an open ended list, adding new key words whenever it is felt necessary. As to the meaning of the key words, it can be taken that every key word is defined by the content of the document or documents to which it has been assigned in the indexing process.

The main advantage of a pragmatic thesaurus derived from actual documents is that it will be a principal retrieval aid for those documents, and it will also aid the user to find appropriate combination of key words which would give satisfactory retrieval results which he expects from the system.

Further, a pragmatic thesaurus can be a very good discussion basis for any conference interested in establishing a thesaurus on a similar area of research. There will be no need to argue by means of arbitrary examples made offhand by the participants of the conference, because a pragmatic thesaurus will contain an adequate number of different examples along with the frequency counts from which one may draw conclusions more easily than from ficti-

tious examples. The main restriction of the approach described here, is the limited number of key words which can be processed within reasonable computer CPU time. We guess that the procedures can he applied to a library of some 10, 000 - 20, 000 documents with a key word repertoir of 1000 words using not more than 3 hours of calculation time.

COMPUTATIONAL PROCEDURES

21 Key Word Statistics and Matrix Inversion

The library of the Zentralarchiv fur Hochschulbau keeps all its documents with a nummerus currens for identification (5, 7). For every key word, there is a peek-a-boo card carrying the numbers of the documents to which this key word has been assigned during the manual indexing process. The information on the peek-a-boo cards have been key punched coding the key words with a number. The statistical evaluation came up with two tables concerning the key word frequency and the indexing depth. The graphic display shows that most key words have been used between 20 and 400 times. The indexing depth shows that on the average a document has been indexed with 4 to 12 descriptors. Because the representation of all document numbers per key word is a matrix it was necessary to make a matrix inversion to get all key words per document.

22 Associate Relationship

It is assumed that there is an associative relation between every two descriptors (= Key words), whenever they have been assigned to one or more documents The reduction of all possible types of relations to a single one, not distinguishing between them at all, is the reason why we call our thesaurus a pragmatic one. In spite of the fact that true semantic analysis would be a very useful thing, there are up to now no computerized systems known, which would do this for a real library. Even intellectual thesaurus construction is a very hard and time consuming task, so that many authors have not specified the type of relation between key words for manually compiled thesuari (8). We checked our procedure against the recommendations for the establishment of monolingual thesauri (3) and found it to be within the rules. The associative relation between two descriptors is understood to be independent of the sequence and of the context of possible other descriptors. All pairs of descriptors have been checked even if the combination of some descriptors do not make sense.

23 Calculation of Similarity Matrix

The similarity matrix has been calculated on the basis of the Tanimoto coefficient for which is a function of the frequency of the descriptor A and the frequency of descriptor B and of the frequency of the two descriptors taken together C.

TANIMOTO COEFFICIENT

$$f = \frac{C}{A + B - C}$$

A = FREQUENCY OF THE DESCRIPTOR *A*
B=FREQUENCY OF THE DESCRIPTOR B
C = FREQUENCY OF THE DESCRIPTORS A AND B TOGETHER

The calculation of the similarity matrix was the most time consuming one, because 450, 000 combinations had to he checked whether there is a common frequency C different from zero and some 45, 000 Tanimoto coefficients had to be calculated. Since the necessary computer time of one and a half hour was not availa ble for a single job, the symmetry of the similarity matrix has not been used, to make a restart of the programme easier without passing all the results calculated up to the breaking point. The input data, namely the inverted file could be kept within the core memory putting always three integer numbers into one word using a MASK/SHIFT subroutine. The complete similarity matrix has been printed out, the coded key words written in plain text along with their absolute and relative frequency counts. According to suggestions of Ivanova (4) only those descriptors should be assumed to be related whose Tanimoto coefficients sum up to half of all coefficients of the same line. We did use an approximation to the proposal of Ivanova putting all those descriptors into one group, which have a Tanimoto coefficient greater than the mean value of that line. The result of this procedure was a list of as many overlapping groups as there were descriptors. A manipula tion was made concerning the very rare and the very frequent descriptors. All thirtyeight descriptors appearing once and six descriptors appearing more than 500 times have been deleate from the analysis, such that there were left 710 descriptors A similar recommendation has been made by Salton for the SMART system. The list of the resulting 710 descriptor groups corresponds to the alphabetical lists in traditional thesauri, except that this list does not distinguish different types of relations.

24 Automatic Cluster Analysis

In general, there is a systematic part and an alphabetical part in every thesaurus. The systematic part describes the area under consideration by means of some subsets of the keywords. The automatic procedure for cluster analysis can be divided into those which produce separated clusters and those which produce overlapping clusters. Since in our case the alphabetical list of descriptor groups is already overlapping we decided to use the minimal-tree procedure which divides all keywords into exactly separated groups The name of the minimal-tree procedure derives from the fact that a matrix of distances has been used in the original work (6), choosing the minimal distance between the words. In our case we had the similarity matrix as a basis, so it is obviously the maximum of the similarity which corresponds to the same procedure. We have kept the name inspite of using the opposite measure.

The minimal-tree procedure is started with an arbitrary descriptor. A second descriptor is picked up out of the group of descriptors of the first one, so that the second one has maximum similarity to the first. The third descriptor has to have the maximum value of similarly to one of the predecessors, and is linked correspondingly The procedure has to be continued until all descriptors have entered the tree exactly once. The tree obtained by this procedure is independent from the starting point as to the sum of all similarities which have been entered into the tree. To cut the minimal-tree into pieces it is recommended to use an arbitrary level of similarity which should be surpassed unless a branch of the tree is not to be cut off (2). A series of trials showed that an appropriate level of similarity is hard to define, since too low a level does not break the tree into a sufficient number of fractions, some being very large, others being very small. The number of fractions increases as the level of similarity increases at which the branches stand the cutting trials, but unfortunately the size of the fractions decreases more rapidly so that a number of single descriptors were found to form clusters of their own To finish a first version of our thesaurus we decided to cut the minimal tree manually using the mean value of all similarities as decision aid. I will return to this point at the end. The arbitrary division of the minimal tree could be made such that all fractions are almost of the same size. We have 23 so called main groups each containing some JO descriptors, which is a quickly scannable size. The definition of our groups and main groups is analogous to the facet and semantic classes in traditional thesaurus work.

3 COMPILATION OF GROUPS AND MAIN GROUPS

This part of the total procedure is mainly for output purposes. It produces lists of all groups of descriptors indicating the number of the main groups within which a descriptor can be found. This scheme of reference among the descriptors will be of great use to one who is trying to retrieve documents on the basis of key words but has not found the proper key word combination to get the relevant documents. Using the references he can change his key word combination step by step, always checking the statistical neighbourhood of any key word. The list of main groups has been printed along with two selected key words which may be some sort of intuitive representatives for the main groups, This method too has to be replaced by some better, objective method when a revision of the thesaurus is made. To make the clusters of the main groups easier to read a graphic representation has been produced by the program described in the next section.

4 GRAPHIC DISPLAY OF THE MAIN GROUPS

The graphic display was not a principal aim of this thesaurus project, but the practical value of such a thing is well recognised. Therefore, the program necessary to do this has been designed to be as easy as possible, not striving for the gain of computer and plotter time. The algorithm is based on an equidistant grid with limitation in the direction of the one axis. The difficulty was to find a free place for every point which has to be entered into the tree such that a line could be drawn not creating ambiguity concerning previously drawn lines. This has been solved by the selection of 32 points which could never produce confusion. Whenever there was a big cluster there has been an auxiliary routine to escape from that part of the area all filled up with points. This escape could be made checking the vertical lines at left and at right for the nearest free point. Obviously this drawing procedure could not represent the similarity by the geometric distance. But the measure of similarity stronger than the mean value has been made visible by double lines between the descriptors. The main groups have been drawn in separate pictures as well as all to-

gether, representing the whole minimal tree

5 CONCLUSIONS

* The procedures described here have enabled the automatic construction of a pragmatic thesaurus for a given medium sized special library. This thesaurus will be a valuable retrieval aid for the same document collection from which it has been derived, Since the collection has been manually indexed, this thesaurus cannot be an indexing instrument except for erroneous misuse of key words.

The procedures for the construction allow for some manual steering of parameters and cut-off values. Besides they are also objective procedures which can be repeated whenever necessary, producing an exact mapping of all the meanings of all the documents within the library. The basic assumption that manual indexing can be a good starting point is assured through the fact, that it is easier for an expert to say this is a valuable book for that library and it should be assigned these descriptors, which can be read in the title or the table of contents or the headings, than to ask an expert or expert conference "Which are the key words of the area of knowledge which you are experts in" ?. The automatic indexing of documents can be expected in the future, but there is still a considerable amount of work to be done on both sides, linguistic and computer research. We, therefore, hope that automatic thesaurus construction is an aid for people looking into libraries to find the information they need.

6 BIBLIOGRAPHICAL REFERENCES

1 Sec 1 BOUCHE (Reinhard), KIND (Friedbert) and SCHWANHAUSSER (Gerhard). Gesamtbibliographie, Zentralarchiv fur Hochschulbau. Teil 1 und 2. Hrsg . : Zentralarchiv fur Hochschulbau. Stuttgart : Zentralarchiv fur Hochschulbau 1973. 436, 891 S.

2 Sec 24 DEICHSEL (Guntram). Verfahren der Automatischen Klassifikation durch Cluster analyse und ihre Anwendung bei morphologischen Untersuchungen an Amoben. Zulassungsarbeit zum Staatsexamen fur das Lehramt an Gymnasien. Stuttgart : Universitat, Irtstitut fur Informatik 1972. Getr.

Sec 1 22 DIN 1463 Vornorm. Richtlinien fur die Erstellung und Wetter* entwicklung deutschsprachiger Thesauri. Berlin, Koln: Beuth Vertr. 1972, 11 S

Sec 23 IVANOVA (NS). Probleme der automatisierten Thesaurusbildung. In : Informatik, Berling 17 (1970) H. 1, S. 25-28.

Sec 21 KIND (Friedbert) and SCHWANHAUSSER (Gerhard). Thesaurus Hochschulplanung. Stuttgart : Zentratarchiv fur Hochschulbau 1973. S. 33-38. (Information. 6/25)

Sec 24 KNODEL (Walter). Graphentheoretische Methoden und ihm Anwendungen. Berlin, Heidelberg, New York: Springer 1969. VIII, 111 S. (Okonometrie und Unternehmensforschung. 13)

Sec 21 RISCHKOWSKY (Franziska). Thesaurus Hochschulplanung Ein aufgabenbezogener Thesaurus fur die Literaturdokumentation der Hochschul-Infornmations-System BmbH (HIS) Hannover. Hrsg. : HIS Hannova Munchen-Pullach : Verlag Dokumentation 1973. IX, 2055

Sec 22 THESAURUS BILDUNGSFOKSCHUNG. Verzeichnis der Deskriptoren und Nichtdeskrv ptoren in der Literaturdokumentation des Max-Planck-Institute fur Bildungs-forsdii Hrsg : Rolf Neuhaus. Mitarbs E. Guhde, B. Hegelheimer, M Rick, E Voswinkel. Mun chen-Pullach: Verlag Dokumentation 1972. XVI, 471 S

APPENDIX

Fig 1 Document Titles with Keywords
Fig 2 Register of Keywords in Context
Fig 3 Indexing Depth
Fig 4 Descriptor Frequency
Fig 5 Example of List of Groups
Fig 6 List of Keywords
Fig 7 Example of Main Group
Fig 8 Part of the Total Minimal Tree
Fig 9 Selection scheme for Drawing of Main Groups
Fig 10 Interrelation among Main Grow
Fig 11 Flowchart

```
56    PRES.:ENGL.
60    ZA-BX 69-2996
71    02
72    BIBLIOTHEKEN:FORSCHUNG PROJEKTE*NUTZER BEFRA-
      GUNG*PLANUNG GRUNDLAGEN;
      GROSSBRITANNIEN
73    ZA BIBL., 02 RES


10          GX-2047-69
30    /STOLTENBERG: DUND MUSS HOCHSCHULGESETZ VER-
      LOESCHIDEN/.
      EINE DER SCHWIERIGSTEN AUFGABEN DES NAECHSTEN PAR-
      LAMENTS, FUER BESCHLEUNIGTEN AUSBAU VON FACH-
      HOCHSCHULEN.
51    IN: /FRANKFURTER ALLGEMEINE ZEITUNG/,1969/182,
      9.8.1969, S.1 UND 4
56    PRES.:DEUTSCH
72    FACH HOCHSCHULEN:ERWEITERUNG*FINANZIERUNG;
      HOCHSCHULE GESETZ:BUNDESREPUBLIK
73    A 28/8 BILDUNGS- UND HOCHSCHULWESEN


          GX-2048-69
      BEER,BRIGITTE
      /SONDERSTATUS FUER MEDIZINISCHE FAKULTAETEN/.
      DEPARTMENTS AN STELLE VON FACHBEREICHEN.
      IN: /FRANKFURTER ALLGEMEINE ZEITUNG/,1969/182,
      9.8.1969, S.2
      PRES.:DEUTSCH
72    ORGANISATION VORSCHLAEGE:MEDIZIN FAKULTAETEN;
      BUNDESR    LIK
73    A 28/8     UNGS-UND HOCHSCHULWESEN

2049

10          GX-2049-69
20    PFEIFFER,ERNST FRIEDRICH
30    /AUSLAENDISCHE ERFAHRUNGEN MIT NEUHEITEN
      DEUTSCHER HOCHSCHULREFORM/.
      MITGEBRACHT VON EINEM INTERNATIONALEN MEDIZINER-
      KONGRESS IN WASHINGTON.
51    IN: /FRANKFURTER ALLGEMEINE ZEITUNG/, 1969/202,
      2.9.1969, S.2
56    PRES.:DEUTSCH
72    REFORMMEDIZIN STUDIUM;
      INTERNATIONAL
73    A 28/8 INTERNATIONAL


          GX-2050-69
      DIE /AUSSENSTATION DER EIDGENOESSISCHEN TECH-
      NISCHEN HOCHSCHULE HOENGGERBERG/.
      IN: /CHEMISCHE RUNDSCHAU/. 21/1968/47, 20.11.1968,
      S.859-861
      ABB.3, PL.1 (LAGEPL.)
      PRES.:DEUTSCH
      AUSLAGERUNG:PHYSIK ABTEILUNG; ERWEITERUNG:TECH-
      NISCHE HOCHSCHULE;
      SCHWEIZ:ZUERICH
      A 28/8 SCHWEIZ ZUERICH


10          GX-2051-69
30    DIE /EMPFEHLUNGEN DES WISSENSCHAFTSRATES ZUR
      HOCHSCHULREFORM/.
      AUSZUEGE AUS DEN VORSCHLAEGEN ZUR NEUORDNUNG DER
      STRUKTUR UND VERWALTUNGSORGANISATION DER UNIVER-
      SITAETEN.
51    IN: /STUTTGARTER ZEITUNG/.1969/289, 14.12.69,
      S.63
56    PRES.:DEUTSCH
72    EMPFEHLUNGEN WISSENSCHAFTSRAT:HOCHSCHULREFORM;
      FACHBEREICHE;
      BUNDESREPUBLIK
73    A 29/8 WISSENSCHAFTSRAT


10          GX-2052-69
20    LUTTERBECK,F.
30    /DOKUMENTATION HEUTE/.
      UEBERALL IN DER BUNDESREPUBLIK DEUTSCHLAND
      ZUGAENGLICHE INFORMATIONSSPEICHER.
51    IN: /VDI-NACHRICHTEN/,23/1969/33, 23.8.1969,
      S.15
56    PRES.:DEUTSCH
72    DOKUMENTATION:BEDARF*ENTWICKLUNG;
      BUNDESREPUBLIK
73    A 28/8 DOKUMENTATION


10          GX-2053-69
20    SCHATZ,JOACHIM
30    /SYSTEMMANAGEMENT - INFORMATION ALS ENTSCHEI-
      DUNGSHILFE/.
      AUFBAU UND WIRKUNGSWEISE VON MANAGEMENT-INFOR-
      MATIONS-SYSTEMEN.
51    IN: /VDI-NACHRICHTEN/,23/1969/34, 20.8.1969,
      S.7
55    GPH.5
56    PRES.:DEUTSCH
72    MANAGEMENT INFORMATION SYSTEM:ARBEITSWEISE


      *ORGANISATION; UNTERNEHMENSFORSCHUNG
      A 28/8 INFORMATION


10          GX-2054-69
20    LINDER,WICHARD
30    /WIE ENTSTEHT EIN FACHAUFSATZ/?
      VORSCHLAEGE ZUR PRAXIS DER STOFFSAMMLUNG,
      GLIEDERUNG UND ABFASSUNG EINES MANUSKRIPTES.
      IN: /VDI-NACHRICHTEN/, 23/1969/31, 30.7.1969,
      S.9
56    PRES.:DEUTSCH
72    ARBEIT METHODIK:WISSENSCHAFT REFERAT
73    A 29/8 WISSENSCHAFTLICHES ARBEITEN


10          X-2055-69
30    /PLANSPIEL USP/.
40    BISKU ,MANFRED (MITARB.)) NEEF,RUEDIGER (MITAR
      OSWAL ,GERD (MITARB.)) SCHAEFER,STEFFI (MITARB.
      MARKE IN,ANTERO (HRSG.)
      PROJEKTGRUPPE (ALTERNATIVE ZUR PLANUNG) (HRSG.)
      STUTTART: UNIVERSITAET, ABTEILUNG FUER ARCHI-
      TEKTU , FACHBEREICH GRL/UM 1969/.0.Z.
      (VERVIELFAELTIGUNG)
      BIBL.  TAB., GPH., DOK.
      PRES.:DEUTSCH
72    SIMULATION MODELL; STADTPLANUNG:ABLAUF MODELL
      *ENTSCHEIDUNG VERFAHREN
73    A 28/


10          GX-2056-69
20    KUEHLMEYER,M.
30    /MATHEMATISCHE GRUNDLAGEN ZUR DER
      BEDARFSNACHFRAGE UND AUSBILDUNGS-[illegible]
      ANWENDUNG AUF DEN MEDIZINISCHEN [illegible]
      1. ENTWURF. (ZUM INTERNEN DIENSTGEBRAUCH)
      ARBEITSGRUPPE MEDIZINISCHER LEHR- UND FORSCH-
      STAETTEN, SEKTION THEORETISCHE MEDIZIN,
      HEIDELBERG:ENTWUERFE/1967/.23 BL. (VERVIELF.)
      PRES.:DEUTSCH
      MEDIZINAUSBILDUNG KAPAZITAET*BEDARF PROGNOSE;
      BUNDESREPUBLIK
      A 29/8 MED. LEHR- U.AUSBILDUNGSSTAETTEN


          GX-2057-69
      DUTTO,U.
      /STUDIENPLAN UND LEHRMENGEN DER FACHGRUPPE
      ALLGEMEINE MEDIZIN/.
      1. ENTWURF. (ZUM INTERNEN DIENSTGEBRAUCH)
      ARBEITSGRUPPE MEDIZINISCHER LEHR- UND FORSCH-
      STAETTEN, SEKTION THEORETISCHE MEDIZIN,
      HEIDELBERG:ENTWUERFE/1967/.6 BL.,ANH.11 BL.
      (VERVIELF.)
55    TAB.
56    PRES.:DEUTSCH
72    MEDIZIN STUDIUM:STUDIENPLAN*STUNDENPLAN;
      BUNDESREPUBLIK
      A 29/8 MED. LEHR-U.AUSBILDUNGSSTAETTEN


10          GX-2058-69
20    SCHNEIDER,G.
30    /PROSPEKTIVE ENTWICKLUNG DER STELLENVOLUMINA DER
      THEORETISCH-MEDIZINISCHEN FAECHER IN HEIDELBERG/.
      1. ENTWURF. (ZUM INTERNEN DIENSTGEBRAUCH)
      ARBEITSGRUPPE MEDIZINISCHER LEHR- UND FORSCHUNGS-
      STAETTEN, SEKTION THEORETISCHE MEDIZIN,
      HEIDELBERG:ENTWUERFE/1967/.10 BL. (VERVIELF.)
55    TAB.
56    PRES.:DEUTSCH
72    MEDIZIN FAKULTAET:STELLENPLAN;
      BUNDESREPUBLIK:HEIDELBERG
      A 29/8 MED.LEHR-U.AUSBILDUNGSSTAETTEN


10          ZX-2059-69
20    LADNYGA,FRANZ
30    /PFLEGEBEREICH DES ALLGEMEINEN KRANKENHAUSES/.
      AUS DEM INSTITUT FUER KRANKENHAUSBAU DER
      TECHNISCHEN UNIVERSITAET BERLIN.
51    IN: /BAUMEISTER/,66/1969/9,S.1092-1099 UND 1154
55    BIBL., ABB., PL. (GRUNDR.), GPH.
      PRES.:DEUTSCH
72    KRANKENHAUS PLANUNG; KRANKENPFLEGE:GRUNDRISS
      ORGANISATION
73    ZA BIBL., 2EITSCH,RIFT


10          ZX-2060-69
20    DESCHNITT,JOACHIM; [illegible]
30    /UNTERSUCHUNG UND BEHANDLUNG/.
      AUS DEM INSTITUT FUER KRANKENHAUSBAU DER
      TECHNISCHEN UNIVERSITAET BERLIN.
51    IN: /BAUMEISTER/,66/,1969/9,S.1100-1111
55    PL. (GRUNDR.)
      PRES.:DEUTSCH
      KRANKENHAUS PLANUNG:FUNKTION BEZIEHUNGEN
      *ORGANISATION
```

Fig 1 Document Titles with Keywords

434	ANGEBOT	AUSGABE	ERMOEHUNG	FACHBEREICH	HOCHSCHULE	KOSTEN	LEISTUNG	OUTPUT	ZUGANG	
4359	AUSGABE	EFFEKTIVITAE	ERMOEHUNG	HOCHSCHULWES	MODELL	PLANUNG	PROGNOSE	PROZESS	ZIEL	
4375	AUSGABE	BMBW	ELEMENT	FOERDERUNG	FORSCHUNG	HOCHSCHULE	INSTITUTION	STATISTIK		
4561	AUSGABE	FINANZIERUNG	PROGNOSE	STADTPLANUNG						
4572	AUSGABE	BILDUNG	BMBW	FINANZIERUNG	FORSCHUNG	HAUSHALTSPLA	HOCHSCHULE	MITTELFRISTI	PLANUNG	WISSENSCHAFT
4661	AUSGABE	HAUSHALT	LEISTUNG	OEFFENTLICH	STAAT	THEORIE	VOLKSWIRTSCH			
4663	AUSGABE	BERICHT	BILDUNG	KMK KULTUSMI	KULTUR	LAENDER	POLITIK	TAETIGKEIT		
4712	ALTERNATIVE	AUSGABE	HOCHSCHULWES	STAAT						
4897	ANALYSE	AUSBILDUNG	AUSGABE	FORSCHUNG	MANAGEMENT	OEKONOMIE	SOZIAL	TAETIGKEIT	VERHALTEN	WACHSTUM
4921	ARBEIT	AUSGABE	BILDUNG	FINANZIERUNG	INNOVATION	OEKONOMIE	ORGANISATION	STUDENT	TECHNIK	WISSENSCHAFT
4924	AUSGABE	BUND	ERMOEHUNG	FACHBEREICH	FOERDERUNG	FORSCHUNG	PROGRAMM	WISSENSCHAFT		
4929	ABBRUCH	AUSGABE	BEGRIFF	BUND	MASSNAHME	NUMERUS CLAU	POLITIK	PROGRAMM	SCHNELLBAU	
5156	AUSGABE	BILDUNG	ERGEBNIS	INVESTITION	THEORIE	VOLKSWIRTSCH				
5399	AUSGABE	BMBW	HAUSHALT	VERTEILUNG						
5459	ANWENDUNG	AUSGABE	BIBLIOTHEK	COMPUTER	DATEN	ZENTRALISATI				
5629	AUSBAU	AUSGABE	GESAMTHOCHSC	HOCHSCHULE	LEHRKOERPER	PERSONAL	QUANTITAET	RAUM	REFORM	STUDENT
	STUDIENREFOR	VERWALTUNG	ZAHL							
5699	AUSGABE	BERICHT	BUND	FORSCHUNG	FORSCHUNGSPL	INTERNATIONA	KOOPERATION	POLITIK	PROGRAMM	REGIERUNG
	SCHWERPUNKT	WISSENSCHAFT								
										AUSLAGERUNG
313	AUSLAGERUNG	MENSA	ORGANISATION							
1696	AUSLAGERUNG	AUSSTATTUNG	BEDARF	BESTAND	ERMITTLUNG	GESAMTPLANUN	GESCHICHTE	LAGEPLAN	NATURWISSENS	PERSONAL
	RAUM	UNIVERSITAET								
2050	ABTEILUNG	AUSLAGERUNG	ERWEITERUNG	LAGEPLAN	PHYSIK	TECHNISCHE				
2596	ABLAUF	AUSLAGERUNG	FINANZIERUNG	PLANUNG	PROGNOSE	STUDENT	UNIVERSITAET	WETTBEWERB	ZAHL	
2757	AUSLAGERUNG	GESAMTPLANUN	UNIVERSITAET							
2758	AUSLAGERUNG	GESAMTPLANUN	HOCHSCHULE	LAGEPLAN						
3839	ABTEILUNG	AUSLAGERUNG	BEVOELKERUNG	DIENSTLEISTU	ERWACHSENENB	FERNSEHEN	FUNKTION	PROGRAMM	UNIVERSITAET	UNTERRICHT
	WEITERBILDUN									
4254	AUSLAGERUNG	BEDARF	BEWERTUNG	ERWEITERUNG	FAKTOR	GEBAEUDE	GELAENDE	GESAMTPLANUN	HOCHSCHULE	METHODE
	METHODIK	MIKROSTANDOR	RAUMBEDARF	UNTERSUCHUNG	VERGLEICH	WERT	ZEIT			
										AUSSCHREIBUNG
19	ANSICHT	AUSSCHREIBUN	FLACHBAU	GRUNDLAGE	INGENIEURWIS	KONSTRUKTION	MODELL	NATURWISSENS	NUTZUNG	PLANUNG
	SCHEMA									
260	AUSSCHREIBUN	UNIVERSITAET	WETTBEWERB							
277	AUSSCHREIBUN	STAEDTEBAU	WETTBEWERB							
372	AUSSCHREIBUN	UNIVERSITAET	WETTBEWERB							
954	AUSSCHREIBUN	UNIVERSITAET	WETTBEWERB							
1132	AUSSCHREIBUN	LAGEPLAN	MENSA	RAUMPROGRAMM	STRUKTUR	UNIVERSITAET	WETTBEWERB			
1133	AUSSCHREIBUN	BELASTUNG	ENTWICKLUNGS	GERAET	HOERSAAL	LAGEPLAN	MEDIENVERSOR	MODELL	PROGNOSE	RAUM
	RAUMPROGRAMM	SPORTSTAETTE	STADTPLAN	UNIVERSITAET	VERKEHRSFLAE	VERWALTUNG	WETTBEWERB	ZONE		
1134	AUSSCHREIBUN	BELASTUNG	BEZIEHUNG	BIBLIOTHEK	BIOCHEMIE	BIOLOGIE	DEMONSTRATIO	ENTWICKLUNGS	EXPERIMENTIE	FACHBEREICH
	FUNKTION	HOERSAAL	KONSTRUKTION	LABOR	LAGEPLAN	LAGERRAUM	LEHRSTUHL	MATHEMATIK	MEDIENVERSOR	MODELL
	PHYSIK	PHYSIOLOGIE	PRAKTIKUM	PROGNOSE	RASTER	RAUMPROGRAMM	SCHEMA	STADTPLAN	STELLENPLAN	STRUKTUR
	SYSTEM	THEORETISCHE	UNIVERSITAET	VERKEHRSFLAE	VORKLINIKUM	WERKSTATT	WETTBEWERB	ZAHNMEDIZIN	ZONE	
1152	AUSSCHREIBUN	BIBLIOTHEK	PHILOSOPHIE	RAUMPROGRAMM	STUDENT	TECHNIK	UNIVERSITAET	WETTBEWERB	WOHNHEIM	
1572	AUSSCHREIBUN	STUDIENREFOR	UNIVERSITAET	WETTBEWERB						
2099	AUSSCHREIBUN	UNIVERSITAET	WETTBEWERB							
2100	AUSSCHREIBUN	UNIVERSITAET	WETTBEWERB							
2183	AUSSCHREIBUN	BAUPLANUNG	GRUNDLAGE	RATIONALISIE						
2676	AUSSCHREIBUN	HOCHSCHULBAU	INSTALLATION	JAHRESBERICH	KONSTRUKTION	PLAN	PROGRAMM	STANDARDISIE		
3025	AUSSCHREIBUN	ERGEBNIS	ERWEITERUNG	HOCHSCHULE	MEDIZIN	NATURWISSENS	RAUMPROGRAMM	WETTBEWERB		
3047	AUSSCHREIBUN	BEWERTUNG	ERGEBNIS	LAGEPLAN	SCHULBAU	SCHULWESEN	VERFAHREN	WETTBEWERB		
3404	AUSSCHREIBUN	GESAMTPLANUN	GRUNDLAGE	HOCHSCHULE	PLANUNG	UNIVERSITAET	VERFLECHTUNG	WETTBEWERB		
3515	AUSSCHREIBUN	BIBLIOTHEK	HOCHSCHULE	INSTITUT	MENSA	NATURWISSENS	RECHENZENTRU	SPORTSTAETTE	STUDENTENHAU	VERWALTUNG
	WETTBEWERB	WOHNEN	ZENTRUM							
4044	ARCHITEKTUR	AUSSCHREIBUN	RAUMPROGRAMM	WETTBEWERB						
4045	AUSSCHREIBUN	HOCHSCHULE	RAUMPROGRAMM	WETTBEWERB						
4317	ANALYSE	ARCHITEKTUR	AUSSCHREIBUN	DIDAKTIK	HOCHSCHULREF	LEHRSTUHL	STUDIUM			
4395	AUSSCHREIBUN	BAUVERWALTUN	BERICHT	DOKUMENTATIO	GUTACHTEN	HAUSHALT	KOMMUNIKATIO	LEITUNG	NEUGRUENDUNG	ORGANISATION

Fig 2 Register of Keywords in Context

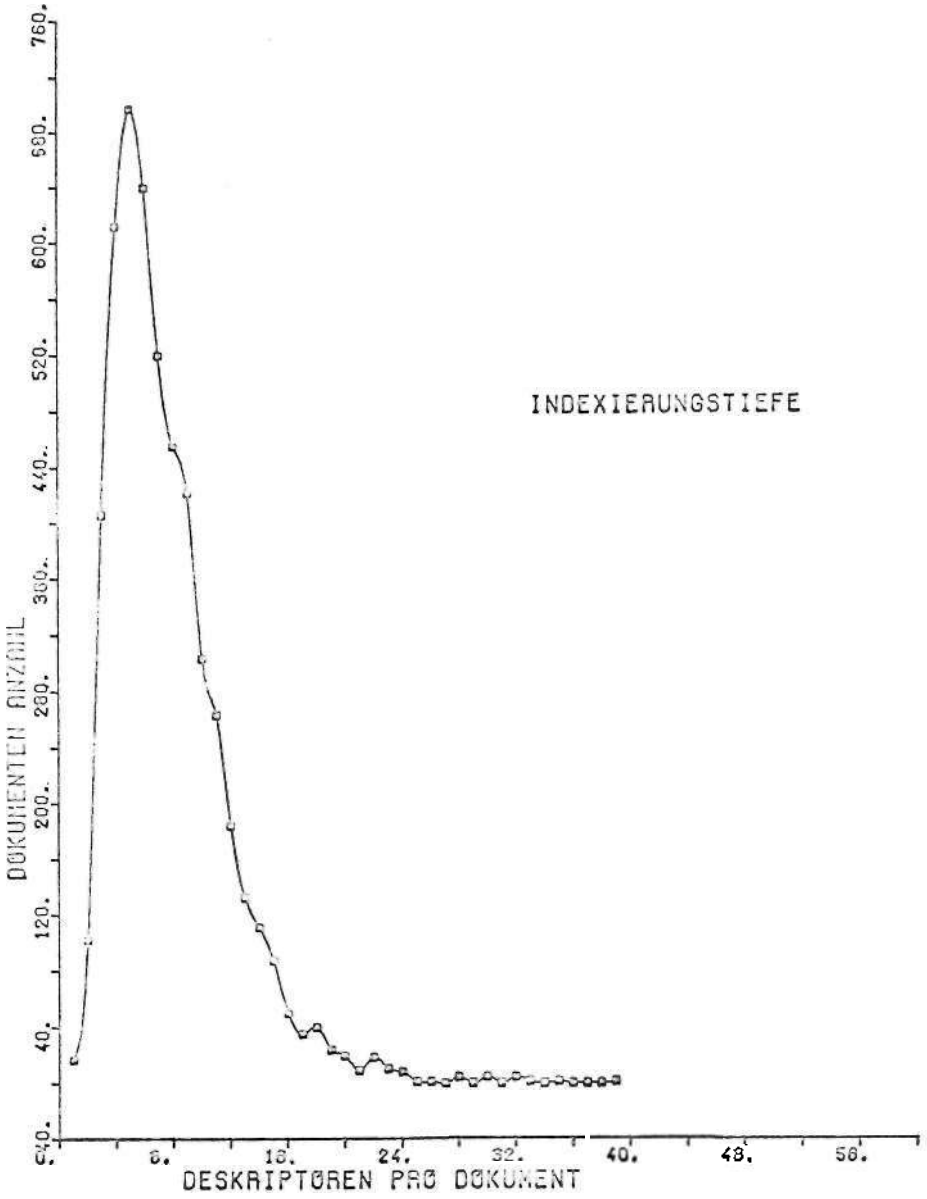


Fig 3 Indexing Depth

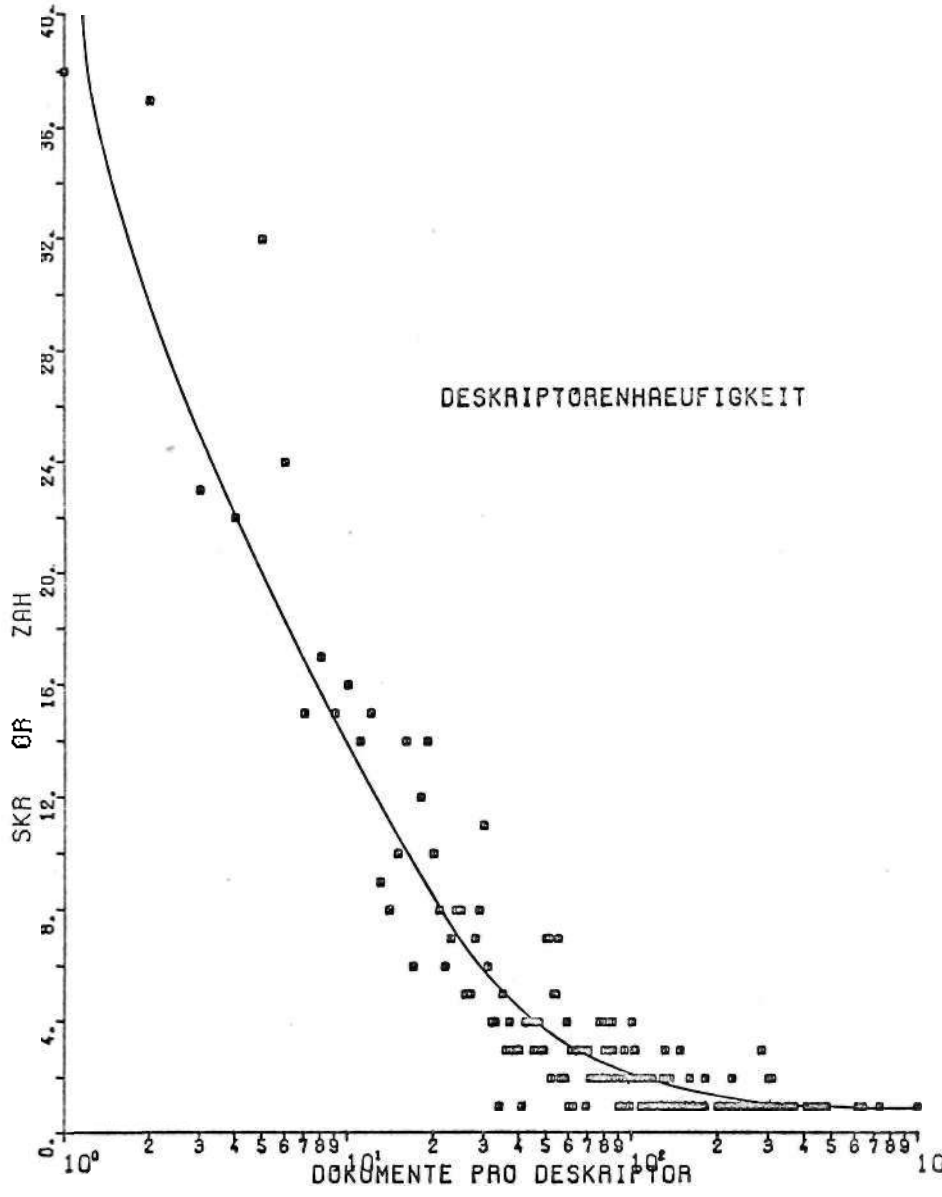


Fig 4 Descriptor Frequency

ABSTAND	5
EINKOMMEN	5
ERHOLUNG	5
LOGIK	5
QUANTITAET	5
VERKEHRSMITTEL	5
VORAUSSETZUNG	5
ZEICHEN	5
ZUSTAND	5

AKUSTIK	HGR. 9
ARBEITSPLATZ	9
BELEUCHTUNG	9
BELICHTUNG	10
BE- UND ENTLUEFTUNG	7
BUERO	9
FLEXIBILITAET	7
GROESSE	9
GROSSRAUM	9
HEIZUNG	7
KLIMATISIERUNG	9
MOEBLIERUNG	7
MUSIKWISSENSCHAFT	9
SCHALLSCHUTZ	9
SCHUTZ	9
WAERME	8

ALLGEMEIN	HGR. 4
BAUSTUFE	10
BEBAUUNG	4
BEISPIEL	3
BELASTUNG	11
BEREICH	14
BESCHREIBUNG	2
BESTIMMUNG	16
BEWERTUNG	4
DARSTELLUNG	10
DETAIL	5
DICHTE	11
KONZEPTION	6
LANDSCHAFT	13
MEDIZINISCHE FORSCHUNGS- UND AU	4
PRIVAT	11
RASTER	7
SIEDLUNG	
SITUATION	
THESAURUS	
URSACHE	
WOHNEN	

ALLOKATION	HGR.
DEMOKRATIE	19
FAKTOR	22
INFORMATIONSSYSTEM	6
INNOVATION	1
KAPITAL	8
KOSTENRECHNUNG	8
MANAGEMENT	19
MOBILITAET	11
OPTIMIERUNG	19
PRODUKTION	7
QUALIFIKATION	22
SIMULATION	20
ZUORDNUNG	8

ALTERNATIVE	HGR. 4
ANALYSE	18
BEISPIEL	3
BESTAND	20
BEWERTUNG	4
BEZIEHUNG	2
DARSTELLUNG	10
EINRICHTUNG	7
ENTSCHEIDUNG	19
ERMITTLUNG	20
ERWEITERUNG	15
FLAECHE	18
FOLGE	5
GEBIET	15
INVESTITION	17
KAPAZITAET	20
KOMMUNIKATION	19
KOSTEN	18
KRITERIUM	4
KRITIK	16
MEDIZINISCHE FORSCHUNGS- UND AU	4
METHODE	18
METHODIK	18
NUTZUNG	18
PROBLEM	21
PROGRAMM	19
PROZESS	19
REGION	15
REGIONALPLANUNG	17
STAAT	3
STADT	15
STANDORT	15
STRUKTUR	19
SYSTEM	19
TRAGWERK	7
TYP	8
UNIVERSITAET * STADT	15

Fig 5 Example of List of Groups

ANALYSE

ANATOMIC

ANFAENGER

ANFORDERUNG (NICHT ALS NACHFRAGE)

* ANFORDERUNG
NACHPRAGE

ANGEBOT

* ANGESTELLTER
PERSONAL

* ANLEITUNG
EINFUEHRUNG

ANORDNUNG

ANORGANISCHE CHEHIE

ANPASSUNG

ANSIGHT

ANTEIL

ANWENDUNG

APOTHEKE

ARBEIT

* ARBEITSPEREICH

TAETIGKEIT BEMESSUNG

ARBEITSKRAFT

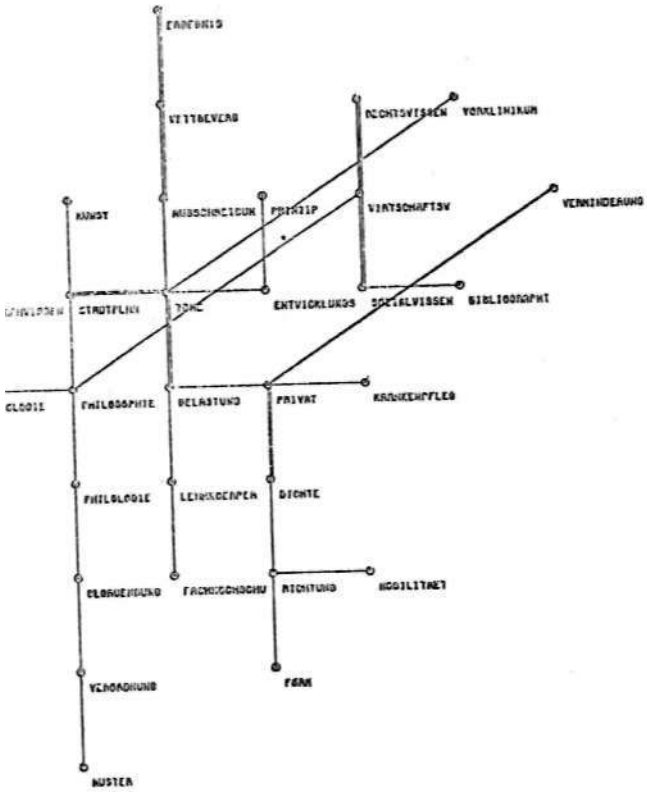
* ARBEITSKREIS
ARBEIT GRUPPE

ARBEITSMARKT

ARBEITSMITTEL

ARBEITSPLATZ

Fig 6 List of Keywords



(WETTBEWERB - ZONE

AUSSCHREIBUNG
BEGRUENDUNG
BELASTUNG
BIBUDGRAPHIE
DICHTE
ENTWICKLUNGSPLAN
ERGEBNIS
FACHHOCHSCHULE
FORM
KRANKENPFLEGE
KUNST
LOHPKOERPER
MOBILITAET
MUSTER
PHILOLOGIE
PHILOSOPHIE
PRINZIP
PRIVAT
RECHTSWISSENSCHAFT
RICHTUNG
SOZIALWISSENSCHAFT
5PRACHWISSENSCHAFT
STADTPtAN
THEOLUGIE
VERMINDERUNG
VERORDNUNG
VORKLINIKUM
WETTBEWERE
WIRTSCHAFTSWISSENSCID:
ZONE

Fig 7 Example of Main Group

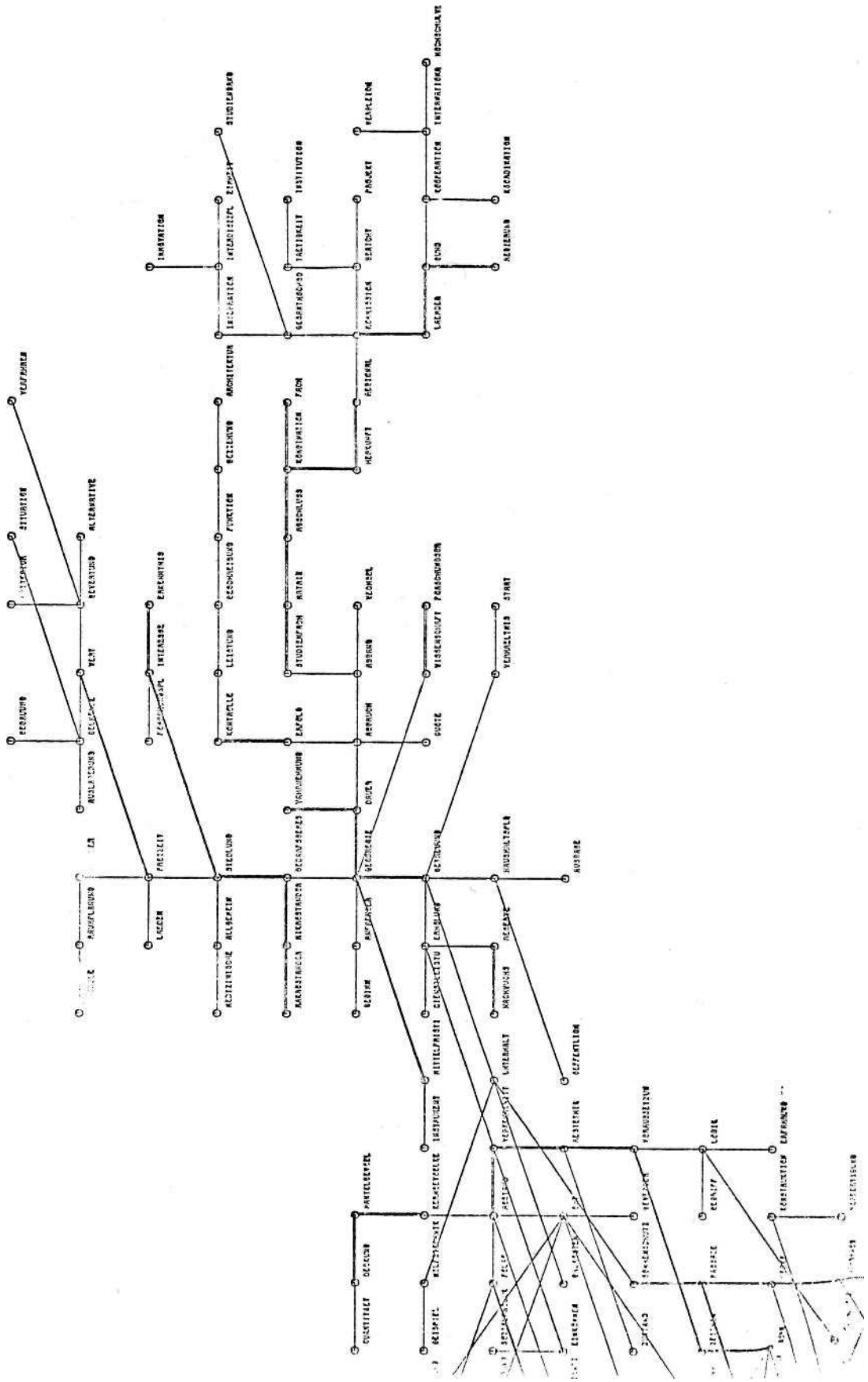


Fig 8 Part of the Total Minimal Tree

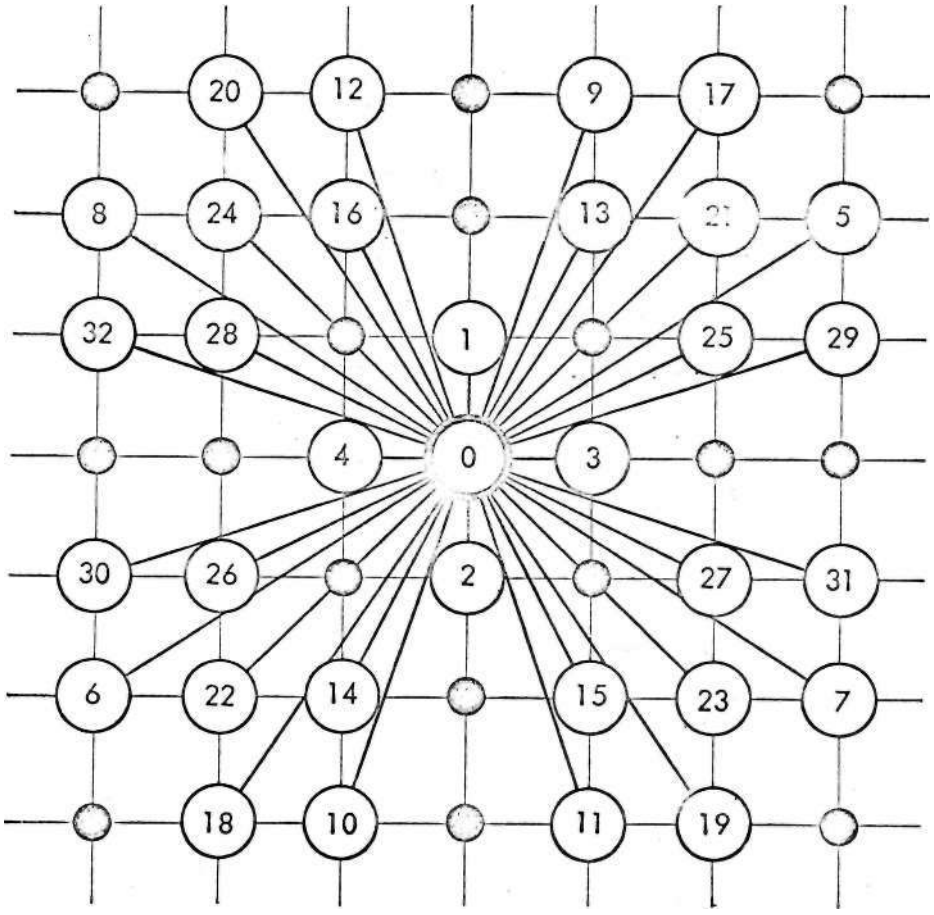


Fig 9 Selection Scheme for Drawing up of Main Groups

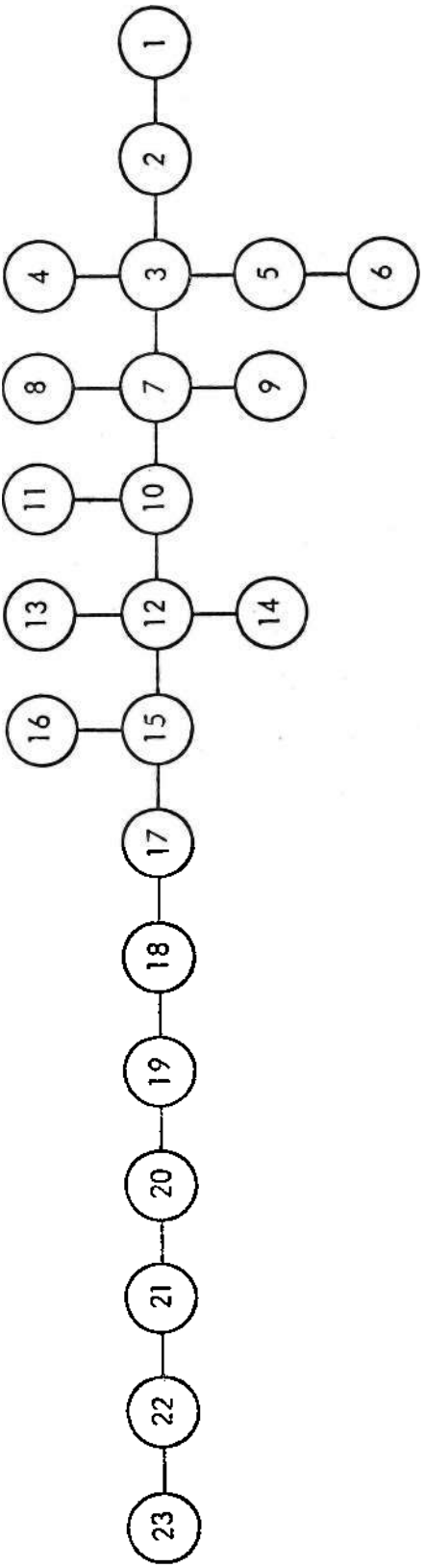


Fig 10 Interrelation among Main Groups

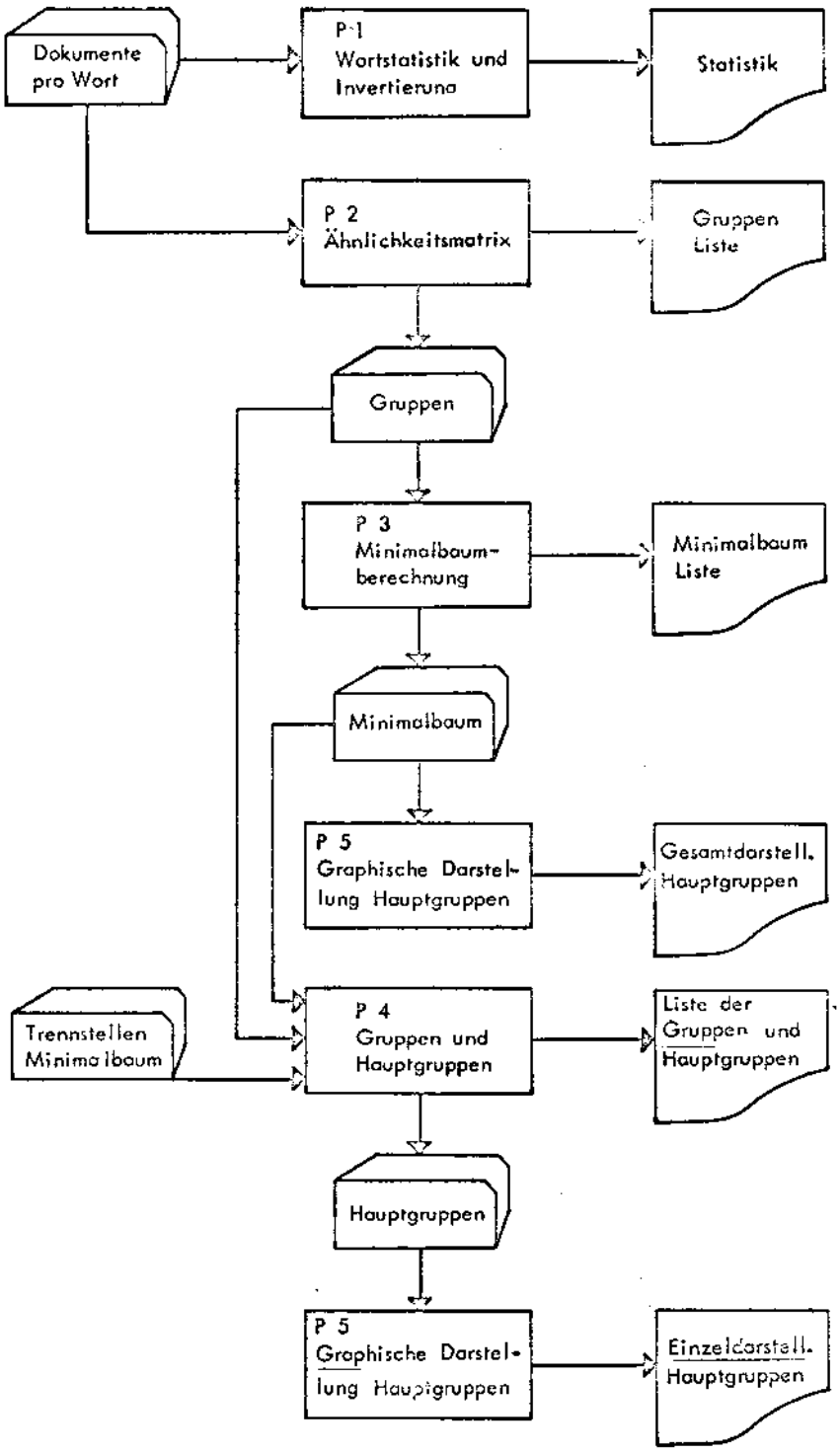


Fig 11. Clowchart