E. BIBLIOGRAPHICAL SERVICES

https://doi.org/10.1017/S0252921100082841 Published online by Cambridge University Press

PROSPECTS FOR AUTOMATED SOLUTION OF THE SUBJECT CHARACTERIZATION PROBLEM IN THE BIBLIOGRAPHIC SERVICES

S. Schiminovich
American Institute of Physics, 335 East 45 Street,
New York, NY 10017, USA

ABSTRACT.-Evidence is presented on the feasibility of an automated generation of classification schemes and the subsequent indexing of papers with the automatically generated subject headings. The output generated for Astronomy and Astrophysics compares favorably with the schemes used by the services and is not inconsistent with recently proposed classifications. Implementation of the automatic techniques seems not only feasible but desirable. However, it would require that the services include for each paper in their database its citations to other papers.

1. INTRODUCTION

Notwithstanding the continuing search for new techniques in information retrieval and the increasing role of computerization, the intellectual steps that go into the preparation of subject indexes by bibliographic services still follow traditional procedures which are more than half a century old. Computerization has had an impact in the clerical aspects of the task; the construction of a working classification scheme for the subject characterization of scientific journal literature and the indexing process itself are intellectual processes that are still carried out manually.

Mounting costs in the processing of the literature would alone justify looking for a computerized alternative, even without an increase in the overall level of quality performance. In this paper we will show that the present state of the art indeed offers such a computerized alternative; that both, the generation of a subject classification and the assignment of papers to the relevant subject categories are tasks that can be performed automatically at a quite improved level of quality compared with the results of the traditional procedures.

After a brief discussion of the algorithm for automatic classification, we will describe its actual implementation in the generation of the first such output ever generated in the fields of Astronomy and Astrophysics.

C. Jaschek and W. Heintz (eds.), Automated Data Retrieval in Astronomy, 147–154.

We will then examine the output of the algorithm and see that it is not inconsistent with some recent proposals for classification schemes, and could be a useful tool if used as supporting material for their development. The discussion will also help explain some shortcomings of the manual, traditional procedures. Mechanization of the classification procedure is not only technologically feasible, it is desirable.

We will end our presentation with an assessment of the likelihood that a bibliographic service will adopt such computerized techniques and the directions in which research should proceed to improve this likelihood

2. THE ALGORITHM FOR AUTOMATIC CLASSIFICATION

We have used the BPDA (Bibliographic Pattern Discovery Algorithm), introduced and described a number of years ago in references 1 and 2.

The procedure, graphically represented in fig. 1, is one of successive approximations. At a given step (n) in the procedure, bibliographies $Bj^{(n-1)}$ generated at the previous (n-1)th step act as triggering files to provide input to the algorithm. A link matrix $Lj^{(n)}$ describing the bibliographic links between the triggering file and the Comparison File is constructed, and the BPDA then determines a set of groups $Gi^{(n)}$ and corresponding bibliographies $Bi^{(n)}$ for the (n)th step in the approximate procedure.

The $Gi^{(n)}$ groups are a partitioning or classification of the Comparison File to n-th order in the appoximation procedure, the corresponding bibliographies $Bi^{(n)}$ can act as kernels for operators to retrieve items relevant to the corresponding subject areas from files similar in structure to the Comparison File.

The available database usually acts as the Comparison File, and its structure determines which links are feasible for the construction of the Link Matrix. Up to now the BPDA has been tested on databases in which the inclusion of references to other papers (citations) are a prominent feature of the database (as opposed to abstracts). Therefore, the links are essentially the measures of biliographic coupling originally introduced by M. Kessler, and in this sense this work is an extension of his pioneering work.[3]

Few new features were introduced into the original BPDA to increase its efficiency. They rely on a symmetry property of the BPDA:

$$G \rightarrow B, \quad B \rightarrow G, \quad L \rightarrow L^T,$$

where L^T is the transpose of the Link Matrix. This symmetry transformation corresponds in the Comparison File to a "time reversal", a resorting of the Comparison File in which citant takes the place of cited and viceversa.

https://doi.org/10.1017/S0252921100082841 Published online by Cambridge University Press

This inherent symmetry of the clustering procedure over the network of citation links proves that too much has been made of the so called "co-citation coupling". The concept is just the time reversed of the citation coupling originally introduced by Kessler, and rather than adding anything new, reflects the fact that for whatever practical reasons, a transposed Comparison File is being used.

In the present implementation of the BPDA, both Comparison Files, the direct and the transposed, have been constructed and the procedure alternates between them from step to step of the approximation, generating in alternative fashion groups Gi$^{(2n+1)}$ and Bibliographies, Bi$^{(2n)}$.

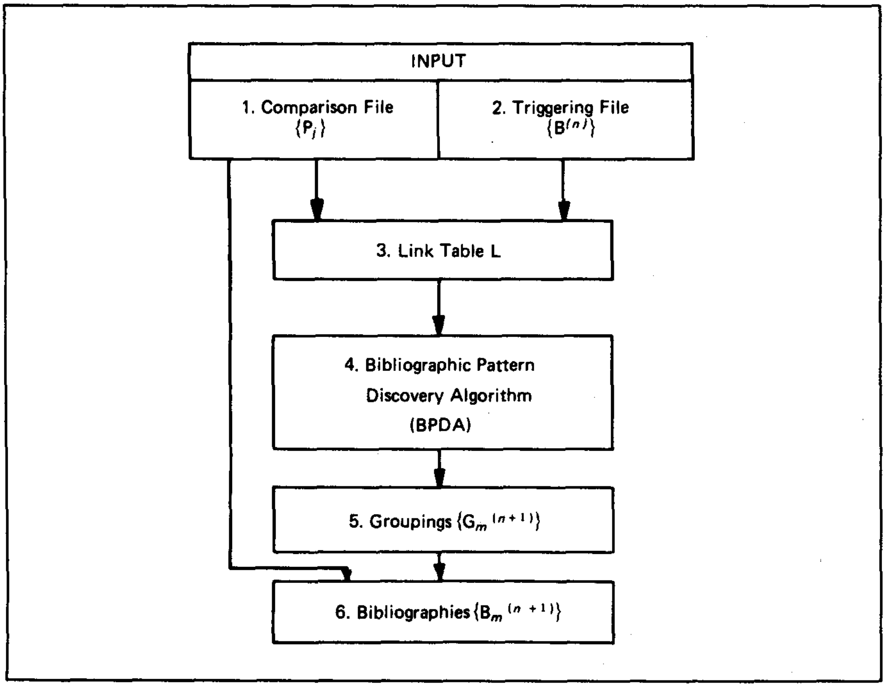
INPUT
1. Comparison File
$\{P_j\}$
2. Triggering File
$\{B^{(n)}\}$
3. Link Table L
4. Bibliographic Pattern Discovery Algorithm (BPDA)
5. Groupings $\{G_m^{(n+1)}\}$
6. Bibliographies $\{B_m^{(n+1)}\}$

Figure 1.

3. IMPLEMENTATION OF THE ALGORITHM

The results presented in this paper stem from the first full scale computerized implementation of the BPDA. The available database was the complete set of 1978 ISI Citation Tapes, comprising the more than 4,000,000 references contained in about 400,000 papers published that year in different areas of science and technology.

The computer configuration used was a Datapoint micoprocessor with 64K RAM attached to a 100MB storage-capacity disk operating system. The database was loaded on the disk storage in the form of two Comparison Files of about 20MB each and an auxiliary file of about 25MB containing bibliographic information to be used for display purposes.

The typical computation time for performing each of the steps of the approximation procedure described in the previous section was approximately 5 to 10 seconds. Four steps of the approximation procedure proved to be adequate for our purposes. Thus, for a given triggering paper chosen from a representative journal in the Comparison File, there were produced a $G^{(1)}$, a $B^{(2)}$, a $G^{(3)}$, and a $B^{(4)}$; the whole process taking, on the average, half a minute of computation time.

Identification of the subject headings corresponding to each grouping had to be done through manual examination, since no words from either titles or abstracts were available in our database. The task proved quite straight forward and unambiguous, given the close link of relatedness between the papers in a grouping. Since the B's are retrospective bibliographies with frequently cited papers in a grouping, a convenient approach was to examine the papers included with highest weight in a $Bi^{(4)}$. These turn out to be those papers from the past which have been seminal in shaping the area of research in question.

The objection may be raised that the manual determination of the names of the groupings vitiates the very automaticity of the whole process. More important than the feasibility of constructing automatically fair approximations to the name (as shown in Reference 1), is in our opinion the fact that

1) The groupings generated by the algorithm can be named in such a way that their membership will produce simultaneous high relevance and recall ratios;
2) The same high relevance and recall ratios will be obtained when retrieval for the generated subject headings is continued on other Comparison Files (or data bases) with retrieval operators that use the B's as their kernels.

In other words, the algorithm has uncovered a structure which is intellectually satisfying and which is stable, that is, one which is not sample-dependent.

4. RESULTS FOR THE FIELD OF ASTROPHYSICS

We have used as a triggering file articles in our own database from the journals "Astronomy and Astrophysics", "Solar Physics", and "Astrophysical Journal". Thus far a set of 100 different groupings have been produced and evaluated; the corresponding subject headings being a 15 to 20% sample out of 500 to 800 expected, by extrapolation, for this area of research.

We cannot present here the subject headings, due to a lack of space. To give a flavor of the generated output, however, we will list a dozen, selected from the subfield of Interstellar Matter:

Dynamics of Molecular Clouds.
Molecule Formation in Interstellar Medium.
Molecule Hydrogen in Interstellar Clouds.
Molecular Species in Interstellar Clouds.
Theoretical Identification of Radicals in Interstellar Clouds.
Studies of CO Clouds.
Silicate and Ice Grains in Interstellar Clouds (IR Observations).
Chemical Composition of Hii Regions.
Far Infrared Observations of Hii Regions.
Recombination Lines and Masers in Hi and Hii Regions.
Coronal Gas in the Galaxy (Ovi Absorption Lines).
Supernova Remnants and Galactic Cosmic Rays.

The tenor of these headings, as well as our estimate of their total number shows that the degree of specificity achieved through the automatic procedures goes one level of depth beyond that used by the services in their classification schemes. See, for example, those used by "Astronomy and Astrophysics Abstracts"[4] or sections 95 to 98 of the "ICSU AB International Classification for Physics".[5]

This greater level of specificity is, however, similar to that being proposed for new, revised schemes for Astronomy and Astrophysics as, for example, the joint UDC revision of FID and commission 5 of the International Astronomical Union.[6] Greatest similarity is encountered with a scheme published by VINITI in 1981.[7] Compare, for example, our headings with section 41.25.29, Interstellar Medium (ISM):

Atomic component of ISM; H II regions
Molecular component of ISM; cosmic masers
Dust component of ISM
Compact regions of ISM
Intercloud medium and clouds in the ISM
Gas--dust complexes in ISM
Dynamics and evolution of ISM
Emission sources of unestablished nature
ISM in the solar neighborhood

The possiblity of achieving this greater level of specificity di-

https://doi.org/10.1017/S0252921100082841 Published online by Cambridge University Press

rectly, as raw material from the output of the algorithmic procedure, is of no little importance, since it allows us to build classification schemes upwards, instead of downwards on the levels of a hierarchy, as is customarily the case. Our experience with the task of establishing the ICSU AB Classification System for Physics[8] seems to indicate that within the higher levels of the hierarchy, i.e., within the first two levels of that system, there is a better consensus between scientists and specialists on how to construct the classification than within the third and fourth levels. There, a paradoxical lack of consensus arises between scientists as to the proper intellectual characterization of their fields of endeavor, and this lack of consensus seems to disappear at the deeper level of specificity encountered by our algorithmic procedures. At this level there is a closer identification with the concrete subject meaning of their work.

Several reasons of a practical nature have kept the services from offering the subject characterization of their databases at the very specific level which scientists would in all likelihood prefer. The less specific levels used do present the material with a vocabulary that eludes the specialized user, a fact that may explain his chronic lack of interest or, sometimes, outright animosity towards the problem of subject characterization of his fields of research.

Differences in the interpretation of subject categories may explain, in part, poor indexing consistency, which is also a problem with the manual systems. The ICSU AB classification, adopted by four major services for the processing of similar material, provides us with an opportunity for conducting large scale studies of indexing consistency between independent teams of indexers. A pilot study that we have conducted shows that two-way agreement for the classification of a document stands at about 35%, and that three-way agreement drops to about 20%. It is then difficult to see how the manual procedures could match the high relevance and recall ratios for algorithmically generated groupings, shown in previous tests to be, simultaneously, both in the vicinity of 90%.

The tardiness with which the manual system is capable of introducing new subject headings to update itself, in response to developments in the field, is another weakness which can be illustrated with an examination of the output of the algorithmic procedures. The concepts uncovered by our analysis were potentially available to automatic systems since 1978, if not earlier; yet, many of them are just filtering into the manual systems. Artificially imposed delays, such as a rule not to update the system more often than a set number of years, should be avoided for proper delivery of a service. Very generally, an algorithmic system would provide for a service a foundation based on objective measures of the bibliographic properties of the database; manual systems, on the other hand, are much too dependent on the accidents of implementation.

5. PROSPECTS FOR THE FUTURE

Evidence has been given that an automatic option is available to the services for the subject characterization of their databases. We believe that the state of the art makes this option not only possible, but, desirable. However, notwithstanding improvements in the products to be delivered and savings in the operating costs, we believe that the prospects for the adoption of the automatic option are very dim at present. The main obstacle is the requirement to incorporate in their databases the references of each document to other documents, in order to achieve high quality results with the algorithmic techniques.

The problem is not one of economics in the long run, since the input of citations is less expensive than manual classification of a document. The problem is rather one of initial investment and break of inertia to follow customary procedures.

For this reason it would seem worthwhile to pursue further work on algorithmic analysis exploring the possibility of achieving comparable results to those presented in this paper through the use of abstracts instead of citations for the construction of bibliographic links between documents. While preliminary results have not been very encouraging in the past,[9] not all avenues have been exhausted.

In the meantime, we believe that an effort should be mounted to algorithmically map all of science through the use of available citation databases. The massive determination of all the algorithmically possible subject headings, together with the uncovering of interrelationships between corresponding groupings, is interesting and has value of its own, and may become a standard against which to measure further developments in subject classification schemes.

REFERENCES

1) Schiminovich, S.: Automatic classification and retrieval of documents by means of a bibliographic pattern discovery algorithm. Informat. Storage Retrieval, vol. 6; pp. 417-435 (1971)
2) Schiminovich, S.: An Automatic Classification of Bibliographic Data Bases. Biosci. Commun. 1; pp. 24-39 (1975)
3) Kessler, M. M.: Bibliographic coupling between scientific papers. Am. Docum. 14; pp. 10 (1963)
4) "Astronomy and Astrophysics Abstracts", Published for the Astronomisches Rechen-Institut by Springer-Verlag
5) "International Classification for Physics 1977", 2nd ed. (revised and expanded). Paris: International Council of Scientific Unions-Abstracting Board (1977)
6) Draft proposal published by FID in collaboration with Commission 5 of IAU.
7) We are indebted to R. B. Rodman for calling our attention to this scheme published in 1980 in Referativnyi Zhurnal Astronomiya, Geodeziya as well as for providing us with an English translation.

https://doi.org/10.1017/S0252921100082841 Published online by Cambridge University Press

8) Berthelot, A., Clague, P., Schiminovich, S., Zwirner, W.: The ICSU AB International Classification System for Physics: Its History and Future. J. Amer. Soc. Information Science, 30, pp. 344-352 (1979)

9) Feinman, R. D., Kwok, K. L.: Classification of scientific documents by means of self-generated groups employing free language. J. Amer. Soc. Information Science, 24, pp. 382-396 (1973)

https://doi.org/10.1017/S0252921100082841 Published online by Cambridge University Press