Lawrence Berkeley National Laboratory

Recent Work

Title

AUTOMATIC DATABASE MAPPING AND TRANSLATION METHODS

Permalink

<https://escholarship.org/uc/item/51c0d5zp>

Author

Cahn, D.F.

Publication Date

1978-04-01

To be presented at the 4th International
Conference of Cybernetics and Systems,
Amsterdam, The Netherlands,
August 21 - 25, 1978

LBL-6782

C.2

AUTOMATIC DATABASE MAPPING AND
TRANSLATION METHODS

D. F. Cahn and J. J. Herr

April 1978

Prepared for the U. S. Department of Energy
under Contract W-7405-ENG-48

TWO-WEEK LOAN COPY

*This is a Library Circulating Copy
which may be borrowed for two weeks.
For a personal retention copy, call
Tech. Info. Division, Ext. 6782*

RECEIVED
LAWRENCE
BERKELEY LABORATORY

JUN 6 1978

LIBRARY AND
DOCUMENTS SECTION

LBL-6782

C.2

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

AUTOMATIC DATABASE MAPPING
AND TRANSLATION METHODS

D.F. Cahn* and J.J. Herr**

*Employee & Information Services Division
Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

**BALLOTS Center
Stanford University
Stanford, California 94305

23 April 1978

Abstract

Automatic methods for content-directed translation and information mapping among machine-readable databases are considered. Multicategory decision algorithms, tested against the classification decisions of human experts in a bibliographic document classification task, matched human performance to the 88% level and showed several key parametric sensitivities. The algorithms tested were based on prior state (experiential) descriptor and category usage and can be configured to provide incremental adaptation or learning. Techniques used emanate from the pattern recognition and artificial intelligence fields.

Introduction

Computer searching and retrieval of bibliographic citation databases has developed in the last several years into a powerful tool for scientific literature survey. Topical (or subject-centered) searches may be performed rapidly, and may be guided interactively using Boolean operators to narrow or broaden the search within any desired bounds [1].

A major problem area in this field concerns the original entry of citations into the databases. A diversity of databases has evolved, each keyed to the needs and desires of users from a specific scientific discipline, and each with unique classification category and descriptor (*keyword*) schemes. As subsets of many databases overlap the interest of workers in fields beyond that which generated the database, it is desirable to be able to map, under computer control, the contents of all or part of one database into the representation space of another. This allows rapid entry of potentially hundreds of thousands of citations to the available search space of the target database, as well as continued automatic updating from the knowledge field of the source. Because of the unique descriptor and category usages of the databases, there is no guarantee (1) that a given descriptor or category will occur in both source and target, or (2) even if it does, that it will have the same meaning in both databases. Thus a one-to-one conformal mapping between databases generally does not exist, and more elaborate means must be utilized to translate between the databases [2].

Our approach has been to model the internal structural content organization of both source and target; test algorithms are used to classify citations in the descriptor and category space, and their decision performance is compared, over a known citation sample set, against that of the human experts who performed the original citation entry. The intent is to isolate the parameters of maximum sensitivity in this type of classification decision-making, both specific to individual databases and in general. Cross-classification algorithms designed to take advantage of these parameters presumably should have the greatest probability of successfully performing the mapping. It is possible, in many cases, to examine on the same basis sample sets of documents that have, in fact, been manually entered in both source and target, and here the technique yields direct data on the translation links among the categories and descriptors. To date, in a 35 category environment, we have investigated multicategory decision procedures and have recorded up to 88% correspondence between machine and human decision-making over 400 citation samples.

Methods

In the studies reported here, the classification scheme of the U.S. Department of Energy's Energy Information Data Base (EIDB) was universally the target. Descriptor source data, and, by a jack-knifing procedure, test data, came from 4000 citation samples of each of EIDB's Geothermal (GEO) and Solar (SOL) subsets, and from a 400 citation sample of Water Resources Abstracts (WRA).

For each document sample, records were prepared indicating the categories in which the documents had been classified by human indexers and the descriptor terms from the indexing vocabulary that had been considered relevant. From these records, it was possible to determine, over the document sample, the frequency with which each descriptor referenced each category. This reference frequency formed a LINK or bond between the descriptor and category that could be normalized by dividing by the total number of references (TOTREF) of the descriptor to all categories; the normalized LINK then represented the conditional probability of occurrence [3] of the category whenever that descriptor occurred in the entry of a document:

$$p[\text{CATEGORY}(j) | \text{DESCRIPTOR}(i)] = \frac{\text{LINK}(i,j)}{\text{TOTREF}(i)}$$

Figure 1 shows one sample page of a matrix of LINK values calculated in this manner for the GEO sample. The columns of the matrix represent the 35 numerical EDB categories covering Geothermal Energy, and the rows represent about 30 of the 1525 descriptors utilized for the GEO sample. Each matrix entry gives the number of times (in a 4027 document sample) the (row) descriptor was associated with a document classified in a particular (column) category. The total number of (row) references of a given descriptor, presented in the TOTREF column, gives an indication of the overall frequency of usage of the descriptor within the training set.

For a given document in the data sample, several descriptors were generally referenced (the average in GEO, for example, is 8 descriptors/document), and the total of their individual excitations of each category represented the total excitation of that category upon presentation of the document:

$$EVIDENCE(j) = \sum_i (p[CATEGORY(j):DESCRIPTOR(i)])$$

Using this summed EVIDENCE as a measure of the probability that the document being examined belonged in a given category, several candidate selection criteria were applied to choose the optimum categories in which to classify the document; the resulting machine selections were then compared to the categories that had been chosen by the human abstractors and data accumulated in statistical files as to the correspondence of the manual and automatic classifications.

Results

Table I shows percent correspondence between automatically and manually assigned categories over each of the data sets, five decision algorithms, and three forms of feature enhancement. The EDB geothermal (GEO) and EDB solar (SOL) data sets contained citation records based on EDB descriptors and categories; in each set, about 4000 such records were used to establish the experience base, and 400 records were then tested against the decision algorithms to obtain the performance measure shown in the table. The Water Resources Abstracts

(WRA) records (400 in evidence base, 225 tested) referenced documents that had been manually cross-classified in the geothermal subset of EDB; thus they linked WRA descriptors to EDB categories. The GEO and SQL samples provided two measures of category predictability within EDB based on the database's own descriptors (intra-database), and the WRA sample provided a measure of actual interdatabase mapping.

The category selection algorithms tested evidenced considerable performance variation. Best performance was achieved with a selector (PEAKS) that detected categories with evidence greater than three times the mean; if the manually selected category fell within this group, a hit was counted. A similar mechanism (STRGP), that used an arithmetic difference in category stimulation (instead of a geometric one) for selection generally did not perform as well. Selection of the most stimulated category (TOP ONLY), the most rigorous choice, gave generally the worst performance. Selection at the second hierarchical level of the category scheme instead of the third (GENLIZ) provided intermediate performance, implying that the decisions were at least a weak function of category specificity.

Our early attempts at feature enhancement showed descriptor utilization to be an important variable. Elimination of descriptors with low overall usage in the training set (HIGHPASS) made little performance difference. However, a BANDPASSED case decreased performance, showing that descriptors thought to be overutilized were indeed contributing to distinguishability among the categories. In each case, the descriptor filter was a binary step function that

simply eliminated descriptors beyond the acceptance bounds from consideration.

In Table II, the results of variation of the filtering thresholds are shown, as tested on the GEO sample. Decision performance is relatively insensitive (2-4%) to variation of the lower bound regardless of the selection criterion used. However, consistent performance variations of 15-18% are evident as the upper bound is varied, and, in fact, a performance saturation is evident for a threshold frequency of 400, implying that descriptors with higher utilization frequencies contribute little to the decision. In correspondence with Salton [4], descriptors with utilization frequencies of 1-10% over the experience set appear to provide the best evidence on which to base categorization decisions. Restriction of the descriptor set to the 1-10% band (40 to 400 references of 4000) greatly reduces the number of descriptors in the evidence file and affords a saving in processing time per citation with no resultant performance penalty; in the production environment that is our ultimate goal, this saving is of significant utility.

Discussion

Several parameters have emerged as significant determinants of simulation performance. The number of categories considered acceptable has a major influence, largely because of the performance constraint relaxation that accompanies its increase. As noted, the acceptance of

coincidence in either of the TOP TWO categories begets a 16% improvement in recorded direct hits. Some of the improvement noted in the STRGP and PEAKS cases is then due to their acceptance, on the average, of hits in more than one category, but, in these cases, it is legitimate to claim a performance improvement since the categories are selected according to a major breakpoint in their excitation; while it may not be possible to isolate a single clearly optimal selection in each case, a strong decision can be made concerning the correctness of the group, and, in a practical system, this disjunctive result may be sufficient either for enhanced secondary separation or for ultimate classification and recall.

Generalization level also has a major effect. At higher generalization levels, the decision involves fewer choices, hence there is simultaneously a larger information base (higher development) to specify the decision bounds of each category and a smaller number of bits in the decision [3]. Both factors improve discriminability. As noted above, the bandpass bounds can be optimized to limit the descriptor set and improve processing efficiency with no degradation in performance.

We feel that the performance levels achieved thus far can be dramatically improved and are encouraging for the future of automatic classification programs as an aid in database development and mapping. The initial 88% performance bound is far above random guess (which would be 3% for 35 categories), and presents a good platform from which to begin system refinement and optimization.

References

- [1] DJE/REC2N: Manual for the Dial-Up User, Publication number TID-4587-R1, Office of Technical Information, U.S. Department of Energy, 1977.
- [2] G. Salton, Automatic Information Organization and Retrieval, McGraw-Hill Book Co., New York, 1968, Chapter 4.
- [3] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, New York, 1973, pp. 114-121 and 131-134.
- [4] G. Salton, A. Wong, and C.S. Yang, "A Vector Space Model for Automatic Indexing," Communications of the A.C.M., 18:11, pp. 613-620 (1975).

Work supported by the U. S. Department of Energy.

Table I: Direct hit percentages for various feature enhancements and category selection algorithms over several data sets.

<u>Feature Enhancement</u>		<u>Data Set</u>		
	<u>Selection Algorithm</u>	<u>GEO</u>	<u>SOL</u>	<u>WRA</u>
NONE				
	TOP ONLY	61	74	47
	TOP TWO	77	83	59
	GENLIZ	69	77	55
	STRGP	63	73	43
	PEAKS	86	88	70
HIGHPASS (usage > 5)				
	TOP ONLY	61	76	42
	TOP TWO	76	83	56
	GENLIZ	68	79	49
	STRGP	63	74	36
	PEAKS	85	88	66
BANDPASS (5 < usage < 100)				
	TOP ONLY	45	65	43
	TOP TWO	59	75	59
	GENLIZ	50	67	49
	STRGP	48	65	46
	PEAKS	69	79	70
Citations in training set		4027	4324	400
Citations tested		393	420	225

Table II: Direct hit variation with modification of descriptor usage bandpassing thresholds. (GEO data set, 393 documents per case, table entries in percent.)

<u>Descriptor Usage</u>		<u>Selection Algorithms</u>				
<u>Low Threshold</u>	<u>High Threshold</u>	<u>TOP ONLY</u>	<u>TOP TWO</u>	<u>GENLIZ</u>	<u>STRGP</u>	<u>PEAKS</u>
<u>Variation of lower threshold:</u>						
5	9999	61%	76%	68%	63%	85%
10	9999	61	75	68	63	84
20	9999	61	75	69	62	83
30	9999	60	74	68	61	83
40	9999	<u>59</u>	<u>74</u>	<u>66</u>	<u>59</u>	<u>83</u>
Variation Range:		2%	2%	3%	4%	2%
<u>Variation of upper threshold:</u>						
5	100	45%	59%	50%	48%	69%
5	200	52	70	59	58	79
5	300	56	74	63	61	82
5	400	60	77	66	65	87
5	500	<u>60</u>	<u>76</u>	<u>66</u>	<u>63</u>	<u>87</u>
Variation Range:		15%	18%	16%	17%	18%

Cahn/Herr:

Automatic Database Mapping Methods

(Fourth Intl. Congress Cyb. and Sys.)

FIGURES

Captions:

Figure 1. LINK matrix sample page for GEO data.
(Full matrix: 1525 descriptor terms X 35 categories;
'-' implies unit reference.)

EDB Hierarchical Level:

- j categories

[illegible][illegible]

This report was done with support from the Department of Energy. Any conclusions or opinions expressed in this report represent solely those of the author(s) and not necessarily those of The Regents of the University of California, the Lawrence Berkeley Laboratory or the Department of Energy.

TECHNICAL INFORMATION DEPARTMENT
LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720