To be presented at the International
Joint Conference On Artificial
Intelligence, Cambridge, MA,
August 22 - 26, 1977

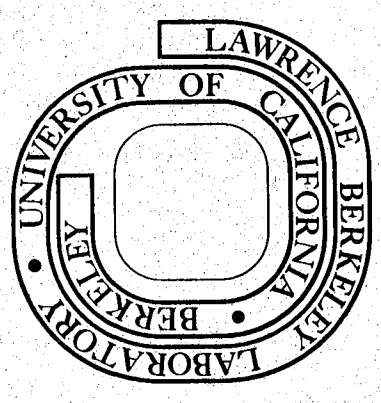
LBL-6164
c.1

AUTOMATIC DOCUMENT CLASSIFICATION BASED ON EXPERT HUMAN DECISIONS

D. F. Cahn and J. J. Herr

March 1, 1977

Prepared for the U. S. Energy Research and
Development Administration under Contract W-7405-ENG-48



LBL-6164
c.1

AUTOMATIC DOCUMENT CLASSIFICATION BASED ON EXPERT HUMAN DECISIONS

D. F. Cahn and J. J. Herr
Information Analysis Group
Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

1 March 1977

Abstract

Machine formation and augmentation of technical information libraries is a provocative and immediately useful application of artificial intelligence. Automatic entry, organization, and retrieval of bibliographic references, and the translation en masse of entire collections of such references from one classification system to another, pose a well defined problem calling for associative processing of contextually linked quanta of information. Measured human expertise in these tasks provides both a yardstick for machine performance and an initial core upon which a dynamic knowledge base can be built for the machine, using incremental learning techniques.

We have constructed several variations of conditional probabilistic and discriminant based automatic document classification programs and compared their performance against manual human document entry over several technical information data bases in the energy field in order to identify parameters affecting this type of classification task. With statistical correspondence between machine and manually generated classifications over fairly extensive document sets as the prime performance yardstick, in a 35 category selection environment, systems we have developed to date have emulated human classification decisions with 60-94% accuracy; in binary decisions, a simulation accuracy of up to 97% was achieved. Through analysis of the algorithms that performed at various levels within this range, and, in particular, through analysis of common characteristics of the document cases in which correspondence failed, it has been possible to isolate several key parameters that regulate performance in systems of this sort and to gain some understanding of the mechanisms used by human experts in performing these types of classification task.

Introduction

A bibliographic database may be considered to be a collection of representations of documents (e.g., journal articles). While these representations have many components, our principal interest is in the components that represent the content or subjects of the documents. This representation condenses the contents of papers in order to facilitate the mechanized manipulation of the information, for instance, in a retrieval system. The condensation is a mapping from the document space into the space of the information retrieval system [1]. We are interested in the general question of mapping from document space to retrieval-system space in order to optimize the process and to automate those aspects of the mapping that can be reliably performed by machine.

There are a large number of bibliographic databases, each designed for a different community of users; because of the different user requirements, each defines its own content-representation space, or indexing language. Our immediate goal is mapping from one indexing-language space to another, that is, translating bibliographic databases into a target indexing language and vocabulary, in order to develop a system in which energy-related subsets of existing bibliographic databases are used to produce a bibliographic database for use by the Energy Research and Development Adminstration's (ERDA) on-going programs. The target indexing language is a combination of a hierarchical mission-oriented classification scheme and a controlled (i.e., limited) indexing vocabulary made up of multi-word terms (called "descriptors"), both of which were developed for ERDA's Energy Data Base (EDB), a bibliographic database being produced using conventional manual content analysis for use in conventional information retrieval systems based on Boolean operations.

In our preliminary investigations, designed to define the scope of the system, we have concentrated on techniques for making use of human experience and judgement in building a system that will classify documents. This approach permits the utilization of a large existing knowledge base in the classification system and provides a method for judging the effectiveness of various algorithms by comparing their results with the manual classifications. Unsupervised clustering and its relationship to experience-based classifiers is also of interest, but will involve developing retrieval methods and techniques for evaluating them.

This paper discusses two approaches to classification of documents based on pattern-recognition techniques. Both make use of a training set made up of documents that have representations in the target indexing language. The two techniques are:

(1) Conditional probabilistic, in which the proportional number of references made by a descriptor to a given category throughout the training set is taken as incremental evidence for inclusion of a test document in that category, and the test document is classified by the accumulation of such evidence provided by its attached group of descriptors.

(2) Discriminant analysis, in which Fisher's linear discriminant function is used as a binary classifier.

By use of these two methods, we have investigated a number of the parameters of an automatic classifier, including breadth vs depth decision trees; selection of classifications in cases in which more than one can be assigned; feature selection (reduction of dimensionality); and detection of ambiguous results.

In the studies reported here, we utilized ERDA's Energy Information Data Base (EDB) for much of our sample data, and Water Resources Abstracts (WRA) for the remainder. EDB subsets for Geothermal Energy, at the most specific (E3G) and intermediate specificity (E2G) hierarchical levels, and Solar Energy, at the most specific level (E3S), containing roughly 4000 document records each, were used in the studies on internal content organization. A 225-document sample from WRA provided the data for use in mapping between databases; for each document in the WRA sample, an EDB record for the same document was identified and the EDB categories for the document were associated with the WRA indexing.

Conditional Probabilistic Classification

For each document sample, records were prepared indicating the categories in which the documents had been classified by human abstractors and the descriptor terms from the indexing vocabulary that had been considered relevent. 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 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 E3G sample. The columns of the matrix represent the 35 numerical EDB categories covering Geothermal Energy, and the rows represent about 50 of the 1525 descriptors utilized for the E3G 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 E3G, 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:

$$\text{EXCITATION}(j) = \sum_{i} \{p[\text{CATEGORY}(j):\text{DESCRIPTOR}(i)]\}$$

Using this summed EXCITATION 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. Sample output pages showing individual record classifications and the cumulative statistics appear in Figures 2 and 3.

In the first selection criterion (UNALTERED), the human abstractor chosen categories were compared with a descending ordered ranking based on normalized-LINK excitation; the number of automatically selected categories with excitation greater than that of the category selected manually was reported by the program as MISSes. A 'DIRECT HIT' indicated correspondence between the category receiving the greatest EXCITATION and the manually selected one. Since more than one category could be selected manually, there may be MISSes of several ORDERs: ORDER 1 gives the number of MISSes before the highest corresponding category, ORDER 2 the number between the first and second, and so on. As it is unclear, when a document is manually classified into several categories, which category is its primary referent, we established a benchmark case in which only documents with single manual classifications were entered and compared statistically with machine performance.

The second classification algorithm was a generalization (GENLIZ) operator. The highest excitation at the most specific hierarchical level--for example, in categories 150100, 150101, and 150102--was moved to the intermediate specificity level (1501), and the other two nulled. If any of the subcategories had been manually selected, the selection was tagged to the generalized category as well (in this case, 1501). Descending ranked correspondence was then tested as before.

The third algorithm was a strength-grouping operator (STRGP). After descent-ordering, the difference between succeeding category excitations was tested; if it exceeded the average slope for the profile by a fixed threshold factor, then all succeeding entries were zeroed. Thus a strength 'cliff' (see Figure 4) separating a strongly excited group from a weakly excited group limited the correspondence search. A correspondence anywhere within the retained group was considered a hit. The height of the cliff was output as CONFIDENCE factor, since it represented the certainty with which the groups could be separated and hence the certainty that the hit had been correctly

retained or dropped.

The fourth algorithm was an application of the STRGP operator to the GENLIZed data.

As it appeared in early experiments that the manually-selected category was often the second entry in the descending-ordered ranking used by the machine for classification, and as a miss in which the correct category was the second most excited indicated far mor deterministic selection performance than one in which the correct category was farther down the ranking, we became interested to see what fraction of the cases actually ranked the manually selected category within the TOP TWO, thus presenting a fifth algorithm.

A sixth algorithm that appeared worth trying as well operated similarly to the STRGP selection criterion, but retained PEAKS in the category excitation waveform if they surpassed in magnitude a threshold predicated on the average excitation over all categories for that document. Thus the cliff was based on relative magnitude rather than slope, as it had been for STRGP.

Statistical summaries (Figure 3) were printed as the document sample was processed, indicating percentages of direct hits, average confidence, and average number of misses of orders 1,2, and 3. Thus we had a direct and cumulative measure of the accuracy with which each of the algorithms modeled human classification. The assumption was that a system that classifies reasonably closely to human experts is performing classifications that will be effective for retrieval. In the case of each document for which a direct hit was not found, an entry was made onto a 'failure list' giving the numerical index of the document and the algorithm(s) under which it failed; this allowed us to perform various types of analysis on the failure cases, and to scrutinize them for common features of the cases that failed.

Experimental Results

In the first group of experiments, five distinct variations on the classification algorithms described above were applied to several data sets, and the results tabulated for comparison. The data sets selected--400 documents from the EDB Geothermal Subset (E3G), 400 documents from a Level 2 generalization of E3G (noted as E2G), 225 documents from Water Resources Abstracts (WRA) that had also been manually classified into EDB and 400 documents from the Solar Subset of EDB-- provided preliminary data on the generalizability of the algorithms and, in the case of the WRA data, directly on translation performance.

The variations tested were as follows. Case 0, the STANDARD, was configured as described above. In case 1, LO TRUNCation, descriptors with total usage frequencies less than 5 (before normalization) were not used in calculating category excitation; the reasoning here was that with such a low total occurrence, they made an undue and skewed contribution to the categories they excited, and, furthurmore, were statistically unreliable. In case 2, HI-LO TRUNCation, an upper usage

frequency bound of 100 was added as well, on the assumption that an overused term had little power for discrimination. A considerable computational advantage was realized with the truncation operators by their size reduction of the evaluated descriptor lists. In case 3, CATFREQ, an additional normalization was introduced to compensate for varying usage of the classification categories in the document samples. Under this operator, the previously computed category EXCITATIONs were divided by the overall usage frequency of the category:

$$\text{EXCITATION}'(j) = \frac{\text{EXCITATION}(j)}{\text{CATFREQ}} = \frac{\sum_i \{p[\text{CATEGORY}(j):\text{DESCRIPTOR}(i)]\}}{\text{CATFREQ}}$$

where CATFREQ = sum of the j-th column of the LINK matrix

Beyond its immediate interpretation as category-usage balanced excitation, EXCITATION' represents the sum over all descriptors of the ratio of the observed co-occurrence of each descriptor and the category to the frequency that would be expected if the descriptor and category were independent.

In case 4, HI-LO TRUNC/CATFREQ, the category frequency normalization was performed on the bandpassed data. In this first group of experiments, multiple classifications were allowed and the TOP TWO and PEAKS criteria were not yet included.

Table I is a summary of the first group of experiments run on the automatic classifier. All matrix values are given as 'percent direct hits', and the axes indicate the experimental algorithms and the document groups on which they were tested. It is apparant (by comparing entries within the rows) that there is little if any degradation in performance for various input document sets, even in the face of considerable performance variation over the various tests and algorithms.

In each case, the source database listed is the source both for the test documents and the term list, and the LINK matrix is calculated between the source terms and the EDB categories. The E3G source case serves as a baseline for Level 3 classification. Here, the LINK matrix was formulated between E3G terms and E3G categories, and the results hence represent the effects of unit translation. In the E2G case, one level of broadening generalization was imposed on the category structure in calculating the LINK matrix from E3G terms, and the resulting improvement is indicative of the broadened acceptance band available at a shallower hierarchical level.

The WRA source case constitutes a first cut translation test. Here, Water Resources Abstracts terms and categories were used as descriptors, and a LINK matrix calculated between them and the E3G categories. The improvement over the E3G-E3G baseline has been preliminarily attributed (based on analysis of the failure cases) to differences in the term 'richness' between E3G and WRA, with the

richer structure of WRA leading to better-defined classification.

The E3Solar case yields a baseline test in an area of EDB other than Geothermal; the lack of degradation here implies that generalization of the techniques is not an unrealistic expectation.

The first four results in the Group II column of Table I represent the single entry 'benchmark' case mentioned previously; they are directly comparable with their multientry equivalents in the E3G Standard case of Group I, and indicate only minor variation between the single and multientry cases. The TOP TWO and PEAKS selection criteria were tested under single entry conditions on the Standard data set. Random guess expected value for any category is 1/35, or about 3%. Thus the UNALT case is operating 65% above random guess. In 68% of the cases tested, the machine picked the same category as the human. TOP TWO demonstrates that 84% of the time, the 'correct' category was one of the two assigned highest scores by the system. Thus the automatic classifier is at least close 84% of the time: a choice between the two categories it finds most excited may not be clear, but its decisions are largely (81%, by differencing measure) deterministic. PEAKS, which detects cliffs in the excitation profile and selects categories with super-cliff excitations, represents, at the very least, a viable mechanism for a fairly drastic, but nonetheless confident dimensionality reduction; automatic classification by this technique maintains 94% simulation accuracy while reducing the number of candidate categories from 35 to 3 or less, a number that may either prove directly viable for classification or may be further reducible by high confidence secondary decisions.

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 (2). Both factors improve discriminability.

It was at first surprising that a performance degradation resulted from the CATFREQ operator. Analysis of some of the failure cases prompted a closer look at the skewness of the LINK data and it became apparant that, even with 4027 documents in the sample, full development of all the categories had not taken place. Some categories had been greatly used and others hardly at all, so a normalization like CATFREQ, even though it did balance category usage, emphasized categories that were only sparsely defined and hence not reliable. The same fault affects the TOTREF normalization, in this case emphasizing underdeveloped descriptors; this explains some of the degradation accompanying the TRUNCation operators. The immediate solution, monitoring the LINK knowledge base to assure full and relatively even development, will be practised as much as practical in future experiments.

Discriminant Analysis

Fisher's discriminant analysis is a multivariate statistical classification technique developed in 1936. Although it can be used for multiple-group classification, to simplify the detailed interpretation of results, we have used it only as a binary classifier. In Fisher's linear discriminant,

$$y = \sum_{i=1}^{N} \{w(i)x(i)\}, \quad \text{where}$$

x(i) is the value of the ith parameter,
N is the number of parameters, and
w(i) is a constant, or weight, associated with the ith parameter,

the coefficients w(i) are derived by maximizing the between-class variance (scatter) of the discriminant scores, y, while minimizing the within-class variances [2, 3].

In these studies of the use of discriminant analysis for classifying documents, the indexer-supplied descriptors were used as the variables. Each document is represented by a vector with N (the total number of terms or parameters used in the analysis) components. A component of the vector has the value of 1 if the term is present for the document and a value of 0 if not present. For each run, a single linear discriminant function was used (permitting discrimination between two groups) and the cases consisted of records from EDB having one of two specified categories. Only items with a single manually assigned category (in the EDB system more than one category can be assigned) were used in the preliminary work to simplify the analysis of results.

These exploratory studies were facilitated by the availability of the SPSS (Statistical Package for the Social Sciences) package [4]. For each analysis, the data were extracted from an LBL-developed data management system and cast into a form acceptable to SPSS. Because a maximum of 100 variables can be handled by the SPSS Discriminant Analysis procedure, the preparation of data for SPSS manipulation included selecting 100 or fewer terms to be used as parameters. The selection criterion used was frequency of use within the two categories being used for the run. Usage within the categories alone is a fairly arbitrary selection criterion since it has little to do with discriminating ability; however, it is practical since terms with low usage in the past have little probability of occurring in the future, and it is a computationally simple technique. As illustrated below, even this arbitrary method produced very promising results. Feature selection is discussed in greater detail later.

Experimental Results

Three sets of category pairs were chosen for the first experiments. Table II lists information about these three pairs and the results of the analyses. For each of the three runs, the table gives the the definitions of the categories, the number of terms and cases, the centroids (average values of the discriminant scores), and the agreement with the manual classification. The category pairs 150101-150102 and 150201-150202 were chosen because the difference between the two components is geographic. This well-defined differentiation reduces the ambiguity in the analysis of results. The pair 1509-1511 (at the second, rather than third, or more specific, level in the classification hierarchy) was chosen to investigate effectiveness at level 2 in the hierarchy; again, these are fairly well differentiated categories (although not as unambiguously differentiated as are those for which differences are based on geography). Considering the arbitrariness of the feature selection criterion, the uniformly high levels of agreement (92.1%, 92.1%, and 97.3%) with the manually assigned categories indicated that further work with the technique would be useful.

The A sample set (150101-150102) was chosen for an investigation of the effect of decreasing the number of terms. In three additional runs, the terms to be used were selected on the basis of the standardized discriminant coefficients (w(i)) in the first run (labelled A-1). The absolute value of the coefficient indicates the discriminating power of the parameter; its sign, the class to which it contributes. In the first of these additional runs (A-2), the 36 terms with coefficients greater than 0.1 were used as the variables; in the second run, the threshold was 0.15; and in the third, 0.20. Table III indicates that feature selection based on the discriminant coefficents can be used to reduce dimensionality with little degradation in performance (agreement with manual assignments of 92.1, 90.7, 89.1, and 84.2% for 96, 36, 25, and 11 terms, respectively). The two centroids for each run and the distance between centroids are also given to illustrate the degree to which the separation of the two categories decreases (and the amount of overlap increases) as the number of variables is decreased. Because these two categories are

differentiated only geographically (US vs non-US), the number of geographic terms in each of these runs is of interest:

Run No.	No. of Terms	Geographic Terms
A-1	96	18
A-2	36	13
A-3	25	12
A-4	11	11

In the last set of experiments, "jack-knifing" was invoked to investigate the reliability of the procedure for cases that were not a part of the training set. One-quarter of the 150101-150102 (A) sample was used as the prediction set, while the other three-quaters constituted the training set. In the first run, the fourth, eighth, ... items were used as the prediction set, while in the second, the third, seventh, ... formed the test set, and so on, until each item had been used as a test item in one of the four runs. Table IV contains the results for the four jack-knifing runs, along with the original (intact) run. As might be expected, the system performs less well for material that was not a part of the training set (average agreement with the manually assigned categories for the four prediction sets was 80.4% and for the four training sets was 93.2%); however, the performance is still quite respectable and significantly better than the 55% that would result from always predicting the more frequently occurring category (150101).

Discussion

A method of detecting cases likely to be misclassified would permit the routing of ambiguous situations to a more sophisticated (and costly) analysis method, for instance, a human analyst. Discriminant scores in the region of highest overlap (near 0) do seem to indicate ambiguity; this region could be declared a "dead zone." Cases with scores in the "dead zone" would be flagged for additional analysis. Table V uses the 150101-150102 jack-knifing results to illustrate the factors to be considered in establishing such a dual-level classification system. The parameters of interest are the number of cases falling within the dead zone, which would be referred to the more costly analysis process, and the degree of agreement with the manually assigned categories (accuracy) resulting from the combination of classification methods.

Three sources of disagreement between the manually and algorithmically assigned categories were isolated by an analysis of the 29 disagreements in the data for 150101-150102 without jack-knifing (cf Table II). These are:

(1) the geographic term that distinguishes between the two categories was not one of the 96 used for the discriminant analysis, 20 cases;

(2) no geographic term was used in the indexing, 4 cases; and

(3) the manually assigned category was erroneous (only errors between 150101 and 150102, which are readily distinguishable, were considered), 5 cases.

The scores for the 20 cases of Type 1 and for four out of the five Type-2 cases have absolute values below 1.0 and would, therefore, probably fall into a zone considered to be ambiguous.

The Type 1 disagreements are due in part to suboptimal feature selection procedures. For a fixed dimensionality, the discriminant coefficients can be used to replace ineffectual parameters by potentially more useful terms. Decreasing the dimensionality decreases the level of performance (agreement with the manual classification and degree of separation, or unambiguousness), but it is reasonable to assume that as the dimensionality decreases, so do computational effort and cost. (For instance, the run in which the three calculations for 36, 25, and 11 terms were performed cost slightly less than the run in which the single calculation for 96 terms was carried out.) Thus, this approach permits choosing the level of performance as a function of the cost to be expended.

In summary, these studies indicate that discriminant analysis does produce categorizations that agree quite well with manually assigned categories, both for training and prediction sets; the quality of the results appear to be independent of the specific categories involved; the discriminant coefficients provide a metric upon which feature selection can be based; and the discriminant scores are usable as indicators of ambiguity.

While the use of existing software facilitated these exploratory studies, SPSS cannot be tailored to our application. We are currently developing software for the further investigation of the application of discriminant analysis to the classification of documents. Areas of particular interest are refinement of feature selection procedures, extension to multicategory cases (n-ary as opposed to binary decisions), assignment of more than one category per document, and the use of other types of information (for instance, journal titles or title and abstract text).

Conclusions

While the simulation experiments reported here are only a first step, we feel that they indicate automatic document classification to be an enlightening and achievable goal within artificial intelligence. The cataloguing of parametric sensitivities that has been started is a central aspect of program development in this area as it ultimately leads to codification and understanding of the fundamental governing principles of the problem. The simulation approach is useful from both the engineering and scientific standpoints, as understanding of human expert performance in a well defined contextual classification task can be developed simultaneous with viable automatic classification systems.

The conditional probabilistic and discriminant analysis approaches presented here are but two of many potentially useful methods. In the practical automatic classifiers that are our ultimate goal, they may be useful either independently or in tandem (perhaps at different hierarchical levels of the decision structure), or they may lead us to other principles that prove more effective in the long run. In any case, it is gratifying that they have converged to reasonable simulation performance as quickly as they have, even before refinement, and they have, at the minimum, provided an excellent test vehicle for incipient illumination of the internal connectivity of the problem.

Acknowledgments

The data were provided by C. Giles, F. Hammerling, and J. Owings, Oak Ridge National Laboratory. S. Sorell assisted with the data-management programming and J. Perra with studies of properties of the databases and EDB-WRA correlations.

References

1. G. Salton, **Automatic Information Organization and Retrieval**, McGraw-Hill Book Co., New York, 1968, Chapter 4.

2. R. O. Duda and P. E. Hart, **Pattern Classification and Scene Analysis**, John Wiley and Sons, New York, 1973, pp. 114-21 and 131-4.

3. W. W. Cooley and P. R. Lohnes, **Multivariate Data Analysis**, John Wiley and Sons, New York, 1971, pp. 243-61.

4. N. H. Nie, C. H. Hull, J. G. Jenkins, K. Steinbrenner, and D. H. Bant, **SPSS: Statistical Package for the Social Sciences**, 2nd Ed., McGraw-Hill Book Co., New York, 1975, pp. 434-67.

Work performed under the auspices of the U. S. Energy Research and Development Administration.

Table I. Direct hit percentages for various experiments. Multiple selections allowed in Group I only.

	Group I				Group II
Cases	E3G	E2G	WRA	E3S	E3G Benchmark
Standard:					
unalt	66	78	84	80	68
genliz	77	78	86	83	74
strgp	67	89	80	75	71
strgp/genliz	86	89	88	83	84
top two	-	-	-	-	84
peaks	-	-	-	-	94
Lo trunc:					
unalt	60	76	53	78	
genliz	73	76	59	81	
strgp	64	89	45	74	
strgp/genliz	85	89	68	83	
Hi-lo trunc:					
unalt	50	61	55	68	
genliz	59	61	61	70	
strgp	51	71	55	66	
strgp/genliz	68	71	69	71	
Catfreq:					
unalt	52	67	72	58	
genliz	60	67	75	69	
strgp	45	59	66	45	
strgp/genliz	61	59	72	67	
Catfreq/Hi-lo trunc:					
unalt	32	42	48	33	
genliz	40	42	53	42	
strgp	33	47	40	36	
strgp/genliz	46	47	51	47	

TABLE II: General Results for Discriminant Analysis

A. Categories 150101 and 150102, 96 Terms, 367 Cases

150101 Geothermal Resources and Availability, USA
150102 Geothermal Resources and Availability, Non-USA

Manually Assigned	Centroid	No. of Cases	Predicted 150101	Predicted 150102
150101	-1.36	202	176 (48.0%)	26 (7.1%)
150102	+1.67	165	3 (0.8%)	162 (44.1%)

92.1% agreement with manually assigned category

B. Categories 150201 and 150202, 96 Terms, 680 Cases

150201 Geothermal Site Geology, Hydrology, and Meteorology, USA
150202 Geothermal Site Geology, Hydrology, and Meterology, Non-USA

Manually Assigned	Centroid	No. of Cases	Predicted 150201	Predicted 150202
150201	+2.27	186	151 (22.1%)	35 (5.1%)
150202	-0.85	494	19 (2.8%)	475 (69.9%)

92.1% agreement with manually assigned category

C. Categories 1509 and 1511, 99 Terms, 695 Cases

1509 Geothermal Engineering
1511 Geothermal Data and Theory

Manually Assigned	Centroid	No. of Cases	Predicted 1509	Predicted 1511
1509	-3.61	202	186 (26.8%)	16 (2.3%)
1511	+1.48	493	3 (0.4%)	490 (70.5%)

97.3% agreement with manually assigned category

TABLE III: Influence of Number of Terms on Discriminant Analysis

Categories 150101 and 150102, 367 Cases (150101, 202; 150102, 165)

No. of Terms	Agree 150101	Agree 150102	% Agree	Centroid 150101	Centroid 150102	Centroid Distance
96	176	162	92.1	-1.36	+1.67	3.03
36	172	161	90.7	+1.28	-1.57	2.85
25	164	163	89.1	+1.22	-1.50	2.72
11	144	165	84.2	+1.08	-1.32	2.40

TABLE IV: Discriminant Analysis Jack-knifing Results

Categories 150101 and 150102
T = Training Set; P = Prediction Set

Run No.	Manual 150101	Agree 150101	Manual 150102	Agree 150102	% Agree
0	202	176	165	162	92.1
1T	153	140	123	120	94.2
1P	49	33	42	39	79.1
2T	152	137	123	120	93.5
2P	50	39	42	38	83.7
3T	150	131	125	122	92.0
3P	52	42	40	28	76.1
4T	151	134	124	122	93.1
4P	51	36	41	40	82.6

Average Agreement

93.2% Training Sets (excluding Run 0)
80.4% Prediction Sets
90.0% Training plus Prediction Sets

Run No.	Centroid 150101*	Centroid 150102*	Centroid Distance
0	-1.36	+1.67	3.03
1	-1.54	+1.92	3.46
2	-1.48	+1.83	3.31
3	+1.51	-1.81	3.32
4	-1.48	+1.80	3.28

* For Training Set.

TABLE V: Inclusion of a Dead Zone in Discriminant Analysis

Jack-knifing Data for Categories 150101-150102, 367 Cases

Dead Zone Range: Range of discriminant scores taken as ambiguous
In DZ, 150101/150102: Number of cases manually assigned to 150101/150102 with scores in the DZ range
Above DZ, 150101/150102: Number of cases manually assigned to 150101/150102 with scores above the DZ range; those with manual category 150101 and scores above the DZ are in disagreement
Below DZ, 150101/150102: Number of cases manually assigned to 150101/150102 with scores below the DZ range; those with manual category 150102 and scores below the DZ are in disagreement
Total Inside DZ: Total number of cases with scores in the DZ; the number of cases that would be referred to a more sophisticated analysis procedure; the number following is the percent of all 367 cases that would fall in the DZ
Total Outside DZ: Total number of cases with scores outside the DZ; these are classified by this procedure; the figure below is the percent of the all (367) cases classified by the procedure
Outside and Agree: Number of cases outside the DZ (classified by this procedure) and in agreement with the manual classification; the figure below is the percent of the outside the dead zone that are in agreement with the manual classification
Total Agree: Total number of cases (out of 367) agreeing with the manual classification; the sum of the number outside the dead zone and in agreement plus the number in the dead zone (assuming that the more sophisticated procedure will always agree with the manual assignment; the number following is the percent of the total (367) cases in agreement with the manual classification

Dead Zone Range	In DZ 150101 150102	Above 150101 150102	Below 150101 150102	Total Inside DZ	Total Outside DZ	Outside and Agree	Total Agree
None	0	52	150	0	367	295	295
	0	145	20	0%	100%	80.4%	80.4%
0.0	14	42	146	32	335	278	310
+0.5	18	132	15	8.7%	91.3%	83.0%	84.5%
-0.5	40	23	139	86	281	248	334
+1.0	46	109	10	23.4%	76.6%	88.3%	91.0%
-1.0	65	12	125	138	229	212	350
+1.5	73	87	5	37.6%	62.4%	92.6%	95.4%

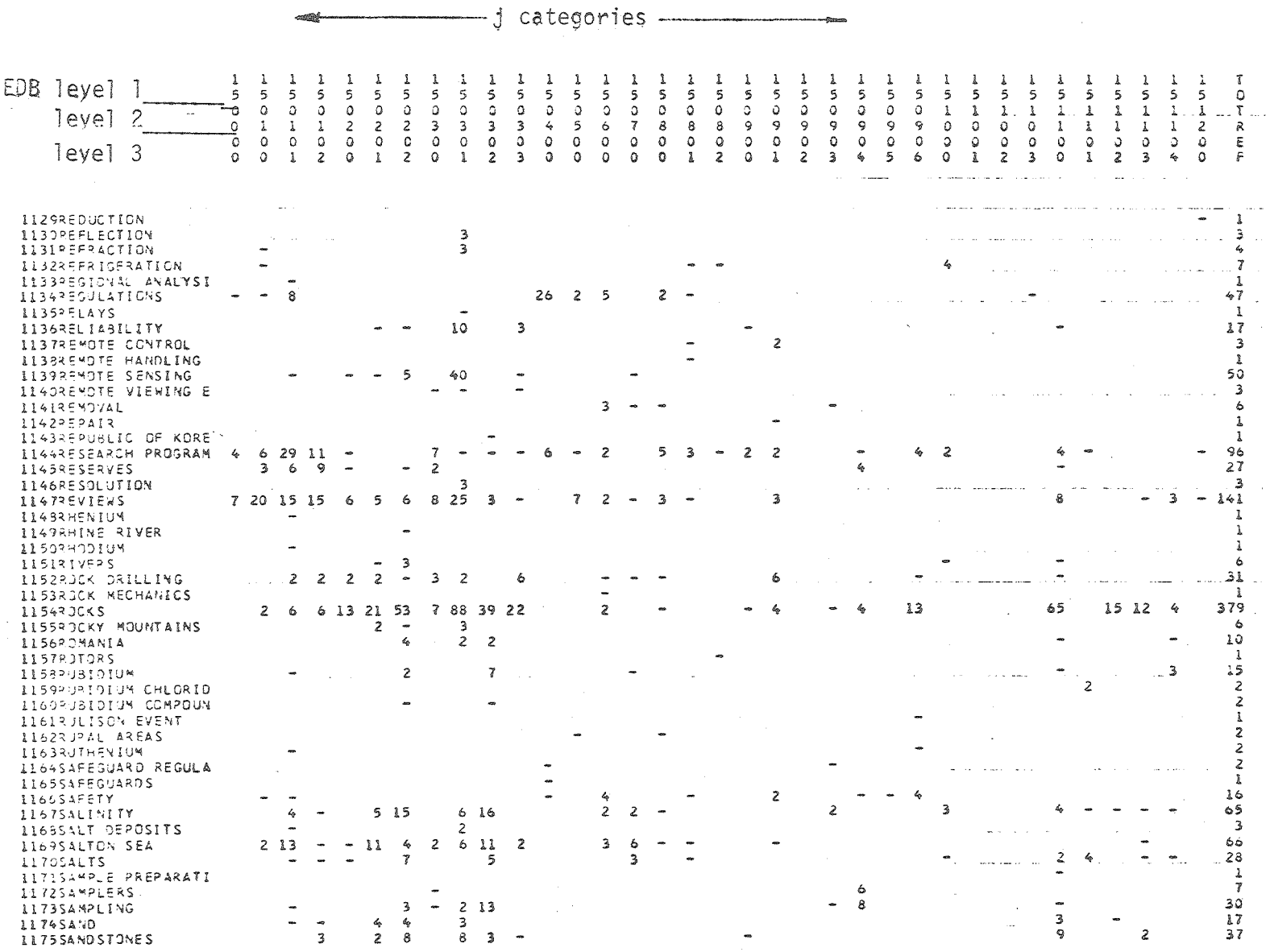


Figure 1. LINK matrix sample page for E3G data.
(Full matrix: 1525 terms X 35 categories; '-' implies unit reference.)

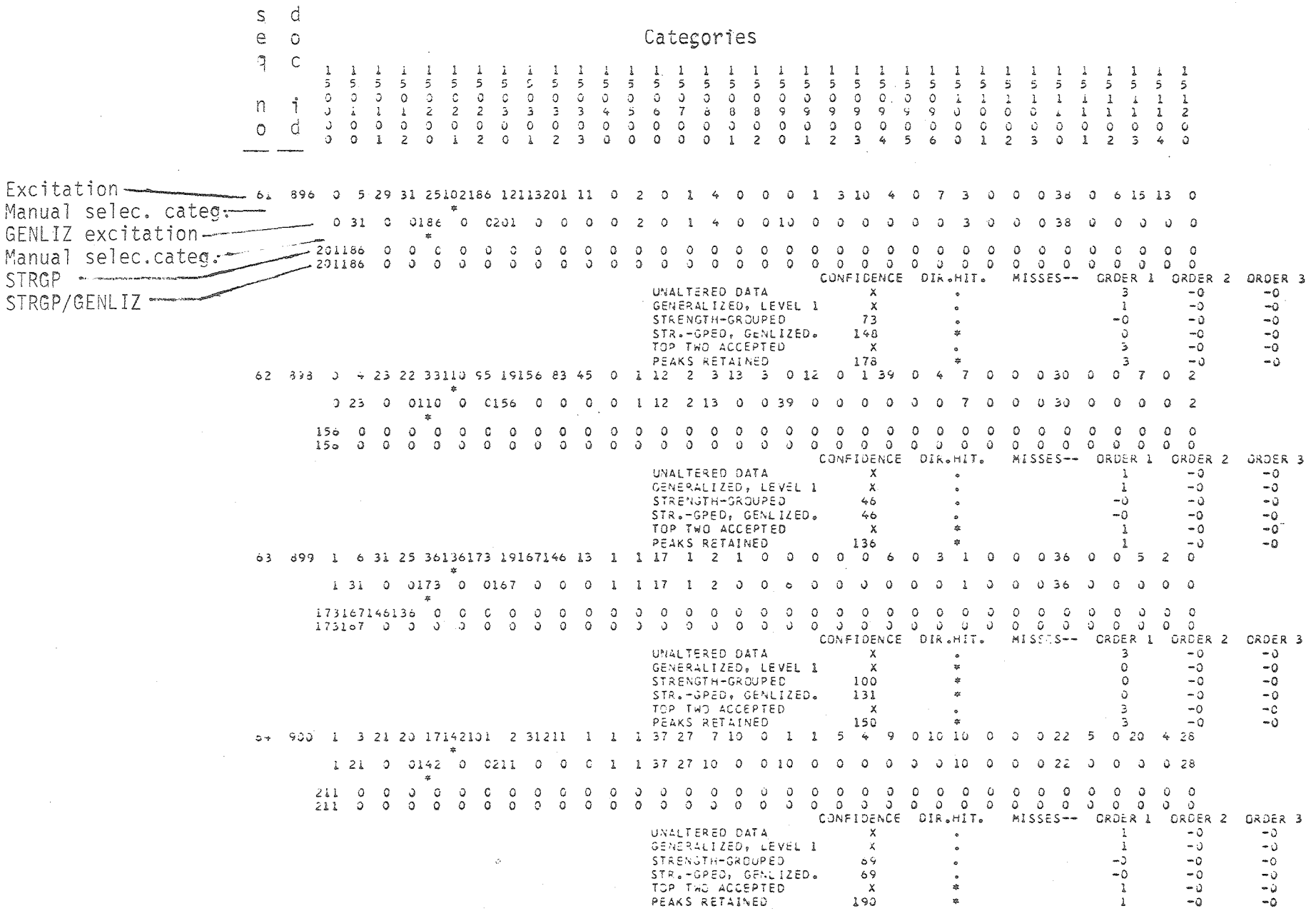


Figure 2. Probabilistic classification sample output. A '*' in 'direct hit' column implies a direct hit for that trial. a '*' in either manual selection row denotes the category chosen manually.

STATISTICAL SUMMARY FOR 380 DOCUMENTS

DIRECT HITS IN PERCENT....
....ALL OTHER ENTRIES IN UNITS/DOCUMENT.

	CONFIDENCE	DIR.HIT.	MISSES--	ORDER 1	ORDER 2	ORDER 3
UNALTERED DATA	0.	.69		.71	0.	0.
GENERALIZED, LEVEL 1	0.	.75		.34	0.	0.
STRENGTH-GROUPED	102.34	.71		0.	0.	0.
STR.-GPED, GENLIZED.	111.93	.85		0.	0.	0.
TOP TWO ACCEPTED	0.	.84		.71	0.	0.
PEAKS RETAINED	191.86	.94		.39	24.13	0.

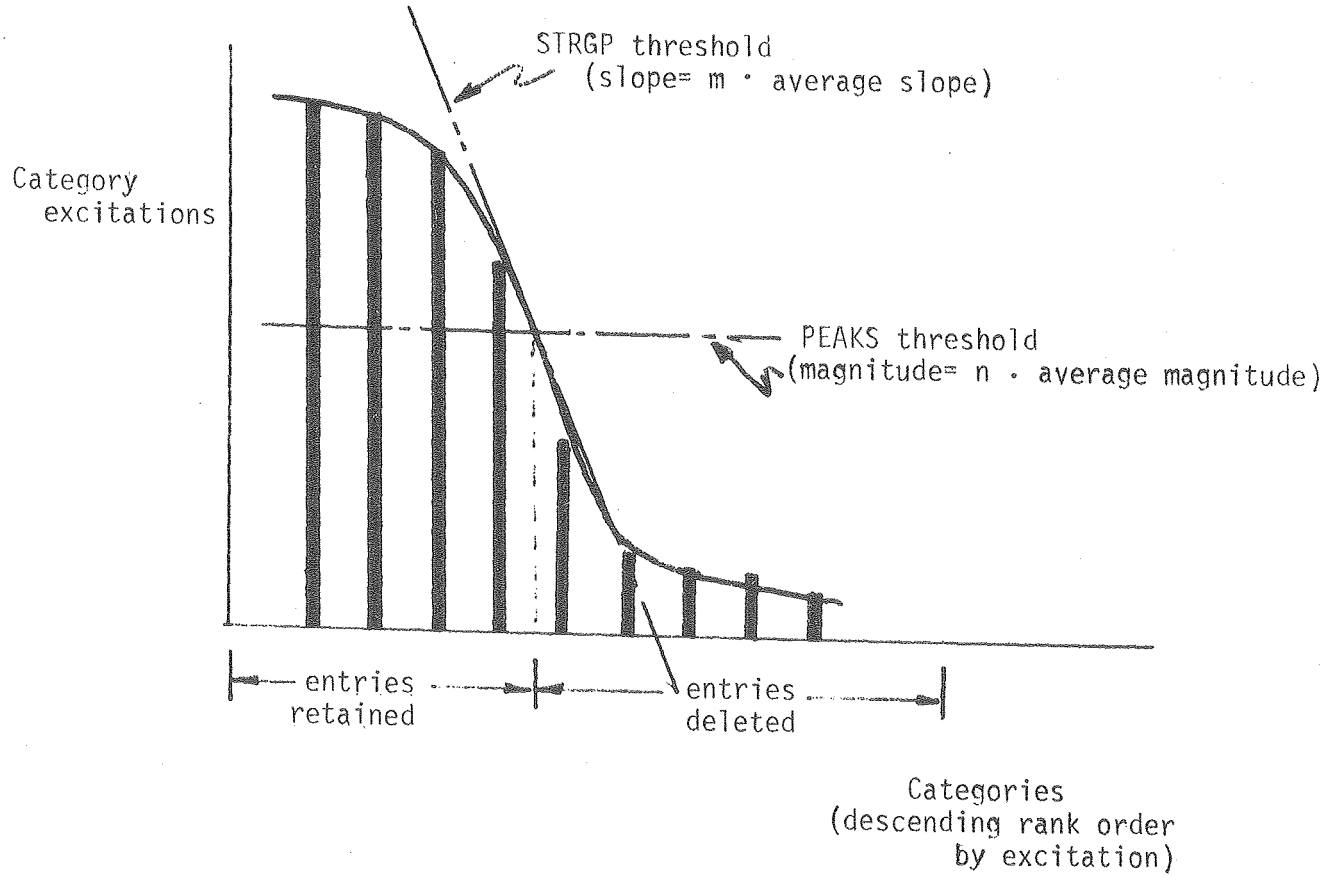
NUMBER OF CATEGORY OCCURRENCES--
5 5 33 13 5 24126 1 20 65 3 0 0 6 1 0 0 0 1 4 1 2 7 0 1 1 0 0 0 19 10 4 16 4 3

TOTAL NUMBER OF CATEGORY OCCURRENCES= 380

NORMALIZED CATEGORY OCCURRENCES, PERCENT--
.01.01.09.03.01.06.33.00.05.17.01.0 .0 .02.00.0 .0 .0 .00.01.00.01.02.0 .00.00.0 .0 .0 .05.03.01.04.01.01

Figure 3. Sample statistical output from automatic document classifier. (Test condition: Standard, single selection benchmark, midstream after 380 of 393 document cases.)

Figure 4. Strength-grouping operator profiles.



This report was done with support from the United States Energy Research and Development Administration. 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 United States Energy Research and Development Administration.