Notes and references on early automatic classification work

Karen Sparck Jones
Computer Laboratory, University of Cambridge
February 1991

The final version of this paper appeared in *ACM SIGIR Forum*, 25(2), 1991, 10-17.

1 Introduction

This informal note was prompted by discussions and questions at the 1990 AAAI Spring Symposium on Text-Based Intelligent Systems (cf Jacobs 1990). There is a growing interest in access to, and the use of, large scale full-text databases for a variety of purposes, and in the application of classification methods to organise the mass of data involved (see e.g. Church and Hanks 1990). A good deal of work has been done in this field in the past, but it is little known, and some of the early research literature is not very accessible. Classification is an area in which it is easy to make plausible but mistaken assumptions, and as this certainly holds for classification in retrieval, there is a good deal that can be usefully learnt from past experience, most of which was hard won from careful thought and grinding experiment. This paper is intended as an introduction to this initial work on automatic classification, to help those now becoming interested in classification to avoid unnecessarily repeating heavy effort or, more especially, reinventing square wheels. It should also be noted that automatic classification and related (e.g. seriation) methods have been extensively developed for biological applications in particular, but have been more variously applied, and that much of this work may be relevant in the broad area of machine learning.

It must be emphasised that as this paper is focussed on early work on automatic classification, particularly for information retrieval, and is designed primarily to lead into this research and its literature, it does not attempt a critical evaluation of the overall results established by now, or of the current state of the art. However it should be pointed out that in the retrieval context in general, as opposed to the wider one of classification as a whole, there has been comparatively little work since the seventies, largely for the reasons indicated in the paper. More recent work in any case refers heavily to earlier research, so this note can be taken as an entry point to the research of the last decade for which some references are given at the end of the note.

2 Automatic classification research in general

Research on automatic classification got going before 1960, in direct response to the opportunities offered by computers for handling large-scale and/or complex data fast and consistently. The research included both work applying already-existing statistical techniques and work seeking to develop new approaches (e.g. the theory of clumps), and as a natural consequence of computation was as much concerned with class-finding algorithms and procedures as with class definitions. It covered a wide range of theoretical perspectives and issues, of practical problems, and of application possibilities. Thus the theoretical research addressed hierarchical or non-hierarchical, and exclusive or overlapping classification, as well as quasi-classificatory structures given by methods like scaling and the very loose structures represented by associative networks; it was also concerned with underpinning definitions of similarity, with the status and properties of feature sets, and with the consequences of (e.g. sparse) feature distributions over data. A good deal of effort was put into practical matters like the manipulation of large arrays and matrices, but the close relationship between theory and practice was recognised in, for example, discussions of order-dependence in class formation and of divisive versus agglomerative techniques. The research of the sixties considered a wide range of applications including not only many biological ones of very different kinds and granularities, but also e.g. anthroplogical and archaeological ones and a variety of applications based on linguistic information including, but not confined to, information retrieval. The biological applications included both those with phylogenetic implications or parallels and those e.g. in medicine, where the derivational history of classes was irrelevant. Within the area as a whole classification was sometimes approached as a primarily descriptive activity and sometimes as a functionally motivated one: in retrieval, for example, classes were good if they retrieved relevant documents, whether or not they were linguistically intuitive or suggestive; indeed the fact that classes were objectively constructed without human participation raised many questions about the motivation for choices of primitive property and of similarity and class definition, and about the criteria for evaluating classifications both where 'independent' grounds like evolution or archaeological stratification, or alternatively hard functional purpose (as in retrieval), were available, and where they were not.

Some of the early work was concerned with constructing classifications for given sets of objects, which might be treated as a one-off activity (as in classifying a set of archaeological pots from a single excavation) or as a starting operation open to modification with new data over a period of time, either continuously or up to some point when a total redo was required: these were strategies in the document area for example. In principle one would expect well-founded methods to allow continuous modification, quite possibly leading to a totally different overall classification structure; but there were many inadequacies in theory and, more importantly, many necessary compromises in practice (e.g. only looking for some of the possible classes), so heuristic strategies naturally followed: some of these strategies allowed adjustment of existing class definitions, in other cases only assignment of new objects to existing classes. Some of the work in the field was indeed primarily concerned with assignment, i.e. with categorisation, for example in indexing where texts were assigned to existing manual subject headings through word-heading correlations. There was little explicit reference to learning (using the word would have been felt to have been claiming too much), but a good deal of the work would nowadays be so characterised, and relevant general issues were certainly recognised (for instance to what extent order-dependence and consequent classificatory biases are formally objectionable but cognitively entirely kosher). There was a good deal of concern with fundamental issues like what constitutes well-foundedness in classification methods, and with defining well-foundedness criteria in such a way that proposed classification methods could be proved to satisfy them: classification stability is an example, where the intuitively reasonable requirement that classification should not be materially affected by small data details has to be given an applicable interpretation. There was a similar concern with establishing generic characterisations of types of classification method, and with providing criteria for determining the appropriate methods to apply to data with given generic properties and for given intended classification uses.

Of course these issues had, and have, been concerns for statisticians. What was felt at the time to have made, and I believe did genuinely make, the difference was three things. The first was the concern with computational procedures and autonomous, large-scale processing. The second was that some conventional statistical techniques, like principal component analysis, were felt to be inappropriate because there was no reason to think that the underlying data had the properties required to ensure the techniques were well grounded; they were also too computationally heavy. But the third and most important factor was the concern with *grouping*. What this meant was finding distinct classes in situations where there were many objects, many properties, and many complex relationships among all of these, so there might be many classes but nevertheless separable (if not exclusive) classes. This treatment of the data was contrasted on the one hand with having just a few partitions or descriptive axes, whether or not these were based on a small number of selected properties or more complex and abstract functions of many, and on the other with continuous orderings. It was justified on the one hand by reference to the manifest complexity of things, and on the other to the equally manifest utility of classes as simplifying devices, treating their members as equivalent and different from the members of other sets. (This view thus allowed both for the possible existence of real natural kinds, out there in the world, and for the empirical construction of classifications designed to impose utilitarian structures on the world.) Basically, the interest in grouping was seen as requiring a balance between plausible simplicity and formal propriety in class definition, matching the need for a handy but well-motivated characterisation of the world. The feeling was that many conventional statistical data reduction techniques were not useful because they did not lead to the right kind of chunking; but there was then of course a problem in evaluating proposed chunking methods and in demonstrating that particular proposed chunkings were useful and reliable.

2.1 Some general references

The terminology in the area is not standard. I have used ”classification” here as a very general term, following earlier practice; ”clustering” was very frequently, but was not systematically, used to refer to hierarchical methods; ”taxonomy” often has biological affiliations. However ”taxonomy” has also been used to refer to theory (or structure) and ”classification” to practice (or process): there are no fixed meanings for the important terms of the area (see the first reference below).

P.H.A. Sneath and R.R. Sokal Numerical taxonomy, San Francisco: Freeman, 1973.

This substantial book gives a very good and comprehensive, if somewhat biologically-oriented, picture of the area as a whole, and is also a very useful point of access into the literature. It is essential reading as an indication of the sophistication and scope of the field. (The book is not a simple updating of Sneath and Sokal's earlier Principles of numerical taxonomy, 1963: the difference reflects the growth in the field in the sixties.)

R.M. Cormack 'A review of classification', Journal of the Royal Statistical Society Series A, 134, 1971, 321-367.

A much shorter, but usefully information-packed introductory review.

P. Macnaughton-Smith 'Some statistical and other numerical techniques for classifying individuals' (Home Office studies in the causes of delinquency and the treatment of offenders 6), London: Her Majesty's Stationery Office, 1965. Good, primarily discursive, presentation of issues.

N. Jardine and R. Sibson Mathematical taxonomy, London: Wiley, 1971.

This emphasises, and considers in detail, well-foundedness in classification, treating a range of problems and approaches from this point of view. Jardine and Sibson's work was notable for demonstrating the formal merits of single-link clustering.

R. Sibson 'Order invariant methods for data analysis', Journal of the Royal Statistical Society Series B, 34, 1972, 311-349.

A useful review focusing on an important general issue in relation to classification and data analysis, especially from a computational point of view.

I am not acquainted in any material detail with the very considerable work that has been done in classification and in related areas of statistics and probability since the early seventies. There is a large literature, a specialist Journal of Classification, an International Federation of Classification Societies, and a lot of software, notably the Clustan package (for an early book related to this and discussing classification from a social science perspective, see B. Everitt Cluste analysis, London: Heinemann (for the Social Science Research Council), 1974; for a useful introductory recent text on the salient form of classification, namely clustering defined as exclusive classification, both hierarchical and non-hierarchical, see A.K. Jain and R.C. Dubes Algorithms for clustering data, Englewood Cliffs NJ: Prentice Hall, 1988). But in spite of the extent to which classification techniques have become established in the last two decades, it is worth noting that even with a lot more computing power available than there was for the early research, there are still substantial challenges in operating on a large scale.

3 Automatic classification relating to information retrieval

The work here was concerned with both word (term) classification and text (document or request) classification, and as in the area in general, stretched from aggressive partitioning to the construction and use of networks. Term classification research covered every kind of activity under the broad heading of thesaurus formation and use, exploiting term relationships of all sorts in all kinds of ways: thus term links can be manipulated to promote recall or precision. Analogously, document classification can be treated both as a device for reducing search effort and as a device for enhancing relevant retrieval through concentration. The natural complementarity of terms (occurring in documents) and documents (having terms) also allows a range of combined classifications.

Thus to illustrate the possibilities, if we have classes of words based on shared word distribution patterns in documents, we can treat these classes as substitution groups defining generic concepts with the same function as conventional thesaurus descriptors, so that a request cointaining one word in a class can match documents containing any of the others. This promotes recall. We can alternatively treat a class as a source of associated words to be added to a description, to increase the number of matching items and thus promote precision. When documents are grouped, say hierarchically, by their shared words, each group can be represented by a single derived term description, so searching can be via matching on these group descriptions. This is more efficient than matching against all the member descriptions individually, but also, depending on the definitions of class, group description and matching function, can promote precision or recall by bringing together similar, presumably co-relevant, documents.

One feature of the early investigations of classification for retrieval (as of the research on lexical classification for language processing purposes like machine translation) was the priority given to functional effectiveness: classifications had to work when put to use to provide relevant documents. In the earliest work, retrieval tests in a proper sense were fairly limited: serious attempts at retrieval testing began in the second half of the sixties. They showed it was much more difficult to get performance improvements using associative and classificatory information than had been expected, even if automatic classifications sometimes worked in ways not predicted from manual thesauri, and led in a naturally recursive way to work designed to understand the underlying properties of document and term data and of the conditions for retrieval derived from such factors as the nature of queries and relevance requirements. This research led in turn to much more work on experimental design and evaluation methods. Over the period from 1965-1980 in particular there were major programmes investigating every form and use of classification for retrieval in long series of experiments, using increasingly standard technology in the way of test collections and evaluation measures and so allowing cross- project comparisons. Salton and his group at Cornell investigated both term and document associations and classes, Sparck Jones and van Rijsbergen, both at Cambridge, focused respectively primarily on terms and on documents. The work was mainly within the then paradigm of non-interactive searching, taking given requests, but did extend to some feedback and adaptation, and there was other work in the field envisaging truly interactive searching exploiting associations and classes.

Unfortunately, the main finding in all of this research was that in general, and therefore setting aside rather specific purposes like reducing search effort through document clustering, associative and classificatory structure contributed little to retrieval performance as measured by e.g. recall and precision, a finding in line with other experimental results seeking to improve, by any means, on basic term coordination. The only exception was when associative information was explicitly, and thus post hoc, tied to relevance assessment, as opposed to being implicitly predictive of relevance status. Thus, for example, enlarging an initial request to include other terms associated with the request's starting terms in known relevant documents may be helpful. But association here is defined very simply as the co-presence of terms in relevant documents, and is not computed independently in a way intended to predict future relevant cooccurrence from actual plain cooccurrence, which was the original aim and is required when relevance information is not available. The work on classification for retrieval as a whole did not show that more sophisticated methods, even taking relevance information into account, were of special, or indeed of any general, value. The major positive finding from the experiments of the seventies was that term weighting, especially when based on relevance information, could be much more effective than classification, and as weighting requires very much less effort than classification, there was little apparent point in continuing to study ways of forming and exploiting classifications.

The more recent work done at Cornell, for example, has on the whole confirmed that associative information is most likely to have some utility when sharpened by relevance facts, perhaps, as Croft has suggested, in an environment combining distinct strategies for characterising terms, documents and requests. The procedures for identifying and using association information may however be very much simpler, for example in query expansion, than those studied in the early classification research summarised in this note. In contrast, statistically principled techniques for document clustering, while sometimes leading to specific improvement in precision, have not generally paid their rent. As the literature introduced via the further references section below makes clear, however, while relevance weighting appears generally useful, the results obtained with associative methods, including ones exploiting relevance information, have been very variable. These results are also complex and difficult to interpret, especially given the lack of consistency in experimental methods and the limitations of many of the tests, particularly where collection scale is concerned. Thus even for devices that may be worthwhile, there is still a general problem in providing an adequate characterisation of the environment conditions determining the utility of different indexing and searching strategies. Information-seeking contexts can vary enormously, and characterising them in a manner which leads to the correct choices of strategy has been a problem since retrieval experiments began, and still is. The germane factors can only be determined by systematic study, and since the number of data variables and their values, and of system parameters and their settings, is normally very large, and the effort of doing many comparative experiments over different, big collections is substantial, we are still not in the position of being able to do more than offer very tentative generalisations, for example about the effects of document and request description lengths on device performance. The underlying issue in all of this is how far collections satisfy the Cluster Hypothesis, to the effect that relevant documents are alike, and unlike non-relevant ones: association methods rely on the Hypothesis, but other indexing and retrieval devices do too. It is therefore unfortunate, as has been found with some test collections, that it is not always well satisfied.

3.1 Retrieval references

The references which follow focus on, and provide convenient access to, early research in this area, or supply connecting links indicating the continuity between early and later work. They are NOT intended to be comprehensive, or to establish priorities.

L.B. Doyle 'Semantic road maps for literature searching', Journal of the ACM 8, 1961, 553-578.

A key early proposal.

M.E. Stevens Automatic indexing: a state-of-the-art report, Monograph 91, National Bureau of Standards, Washington DC, 1965, revised edition 1970.

A comprehensive review including classification work, produced when enthusiasm and hope for this area was at its height.

M.E. Stevens, L. Heilprin and V.E. Giuliano (eds) Statistical association methods for mechanised documentation, Symposium proceedings (1964), National Bureau of Standards, Washington DC, 1965.

This is also a 'peak' collection, directly presenting the work being done in the area and showing its variety.

K. Sparck Jones 'Some thoughts on classification for retrieval', Journal of Documentation 26, 1970, 89-101.

A short discussion of key issues, linking the retrieval application with work on automatic classification in general.

G. Salton (ed) The SMART retrieval system: experiments i automatic document processing, Englewood Cliffs NJ: Prentice Hall, 1971.

A valuable collection of key SMART project papers illustrating the range of the work done by the SMART team and showing how early some ideas like relevance associations were tested. This set of papers gives a better flavour of the early SMART work in this area, and feeling for some of the important detail, than Salton's two textbooks of 1968 and 1975.

G. Salton A theory of indexing (Regional conference series in applied mathematics 18), Society for Industrial and Applied Mathematics, Philadelphia, 1975.

Includes some aspects of term association within the framework of a unified approach to characterising terms by their discrimination value.

K. Sparck Jones Automatic keyword classification for informatio retrieval, London: Butterworths, 1971.

A monograph describing the motivation for, and experiments in, automatic retrieval thesaurus construction initiated with Needham's work on the theory of clumps (itself summarised for the retrieval context in K. Sparck Jones 'The theory of clumps' in The encyclopedia of library and information science (ed Kent and Lancour), 1971).

K. Sparck Jones and R.G. Bates 'Research on automatic indexing 1974-1976' 2 vols, British Library R&D Report 5428, and Computer Laboratory, University of Cambridge, 1975.

Describes a whole series of tests with different collections covering a wide range of indexing methods, including term classifications, and showing that term weighting is much more useful than term classification.

N. Jardine and C.J. van Rijsbergen 'The use of hierarchic clustering in information retrieval', Information Storage and Retrieval 7, 1971, 217-240.

Account of early document clustering experiments in context of general theory and motivation for clustering for retrieval.

C.J. van Rijsbergen Information retrieval, 2nd edition, London: Butterworths, 1979.

This gives a coherent account of the whole field, concentrating on fundamental properties of the problem and on principled approaches to it. This second edition is superior to the first as it includes a chapter on probabilistic retrieval: this is important as van Rijsbergen sees probability as the key modelling notion in the whole area, providing a common underpinning for such actitivities as classification and searching and clearly linking them with learning. The book includes some very useful comments on the earlier classification literature.

K. Sparck Jones (ed) Information retrieval experiment London: Butterworths, 1981.

This includes a review chapter on retrieval system tests 1958- 1978 which serves to place work on automatic classification for retrieval in a wider indexing context.

Several theses of the sixties illustrate the sophistication of early classification work. The work reported in the references listed below was focused on information retrieval; but Needham's work in particular was concerned with general methods and was also applied in other areas.

R.M. Needham 'The application of digital computers to classification and grouping', PhD thesis, University of Cambridge, 1961; published as a report under the title 'Research on information retrieval, classification and grouping, 1957- 1961', Cambridge Language Research Unit, 1961.

E.L. Ivie 'Search procedures based on measures of relatedness between documents', PhD thesis, MIT, 1966.

J.L. Rocchio 'Document retrieval system - optimisation and evaluation', PhD thesis, Harvard University, 1965; also as Report ISR-10, Computation Laboratory, Harvard University, 1966.

B. Litofsky 'Utility of automatic classification systems for information storage and retrieval', PhD thesis, University of Pennsylvania, 1969.

(For some early research on automatic classification for general natural language processing purposes see K. Sparck Jones Synonymy and semantic classification (thesis 1964), Edinburgh: Edinburgh University Press, 1986.)

3.2 Further references

Though this note is focussed on early classification research, useful leads into the most recent work can be found in the references which follow. This list is again not intended to be comprehensive, but is designed to round out the paper by providing access from the other end into what has been a quite continuous line of investigation.

W.B. Croft 'A model of cluster searching based on classification', Information systems 5, 1980, 189-195.

W.B. Croft and R.H. Thompson 'I3R: a new approach to the design of document retrieval systems', Journal of the American Society for Information Science 38, 1987, 389-404.

A. Griffiths, H.C. Luckhurst and P. Willett 'Using interdocument similarity information in document retrieval systems', Journal of the American Society for Information Science 37, 1986, 3-11.

H.J. Peat and P. Willett 'The limitations of term co-occurrence data for query expansion in document retreiavl systems', Journal of the American Society for Information Science, in press.

C.J. van Rijsbergen, D.J. Harper and M.F. Porter 'The selection of good search terms', Information Processing and Management, 17, 1981, 77-91.

S.E. Robertson, M.E. Maron and W.S. Cooper 'Probability of relevance: a unification of two competing models for document retrieval', Information Technology: Research and Development 1, 1982, 1-21.

G. Salton and C. Buckley 'Improving retrieval performance by relevance feedback', Journal of the ASIS 41, 1990, 288-29

A.F. Smeaton and C.J. van Rijsbergen 'The retrieval effects of query expansion on a feedback document retrieval system', The Computer Journal 26, 1983, 239-2

K. Sparck Jones, 'A look backwards and a look forwards', Proceedings of the 11th International ACM SIGIR Conference Research and Development in Information Retrieval (ed Chiaramella), Grenoble: Presses Universitaires, 1988, 13-29.

P. Willett 'Recent trends in hierarchic document clustering: a critical review', Information Processing and Management 24, 1988, 577-597.

I am grateful to R.M. Needham and C.J. van Rijsbergen for comments and suggestions.

3.3 Miscellaneous reference

C.W. Church and P. Hanks 'Word association norms, mutual information, and lexcicography', Computational Linguistics 16, 1990, 22-29.

P.S. Jacobs (ed) Text-based intelligent systems: current research in text analysis, information extraction, and retrieval Report 90CRD198, General Electric Research and Development Centre, Schenectady, 1990.