Abstract: The paper looks at ways in which traditional classification and indexing tools have dealt with the relationships between constituent terms; variations in these are examined and compared with the methods used in machine searching, particularly of the Internet and World Wide Web. Apparent correspondences with features of index languages are identified, and further methods of applying classification and indexing theory to machine retrieval are proposed.

There are various ways in which indexing and retrieval systems, both conventional and electronic, deal with the problem of searching for documents on a subject basis, and various approaches to the analysis and processing of a query. There appear to be three basic models; the taxonomic or structural system, in which the user is offered a map of the 'universe of knowledge'; the language based system, which offers a vocabulary of the subject and a grammar for dealing with compound statements; and the mathematical model using the language of symbolic logic or the algebra of set theory.

1. Enumerative Classification Schemes; their Structural Basis and Searchability

Traditional systems of knowledge organization (the major schemes of classification in particular) have represented knowledge as a coherent and mappable field, with fixed and stable relationships between constituent parts. Older schemes, such as the Library of Congress and (earlier editions of) the Dewey Decimal Classification treat knowledge as an infinitely divisible corpus with the emphasis largely on genus-species and other hierarchical relationships; very often the classification is based on a theoretical model of the totality of knowledge with fixed relationships between the constituent classes. Such structures generally avoid dealing with problems posed by complex inter-relationships between topics and the difficulties of placing compound subjects. The collection as organized by such a system is fixed by the classification and the shelf order is a mirror of the schedule.

2. Faceted Schemes: their Structures and Concept of the Nature of Information

Modern schemes with faceted structure such as the Bliss Bibliographic Classification 2\textsuperscript{nd} edition (Mills and Broughton, 1977-), recent revisions of the Universal Decimal Classification (UDC, 1993)\textsuperscript{\textdagger} and many special schemes such as the British Catalogue of Music Classification (Coates, 1960), also deal with different sorts of non-hierarchical relationships between terms. These systems function in a more dynamic manner, operating in much the same way as a natural language, utilising the constituent terms to express complex concepts in the form of 'subject statements', equivalent in many respects to structured sentences with both a vocabulary and syntactic structure. Individual terms/concepts are arranged in categories largely corresponding to functional roles. It is generally not assumed that any pre-determined relationship exists between the elementary terms, or concepts, of
which the scheme consists, until they are brought together in the shape of a document of some sort except of course for taxonomic relationships of a ‘whole-part’ or thing-kind’ nature, which are essentially permanent). In different documents the relationship can vary, and the same selection of terms can potentially give rise to a number of different compound subjects.

In respect of its appropriateness for dealing with large unorganized databases, it is important to emphasise the ‘unconnectedness’ of the terms in the faceted classification. Theoretically any term can be combined with any others and a very large number of potential compounds of varying complexity is possible. These compounds do not appear in the schedule, and there is no pre-determined grouping of subjects. Scanning the schedules of a classification like BC2 will not give the enquirer any precise idea of what might be contained in the classified file of a collection organized by it.

3. Forms of Syntactic Relationships Exhibited by Faceted Systems

Whilst all faceted systems feature a vocabulary which is controlled by similar types of conceptual and functional analysis, the syntax of the individual systems can vary quite considerably. It is not difficult to find analogies with various forms of natural language structures in the way in which these index languages deal with inter-term relations, which are more complex and subtle than those acknowledged by hierarchical schemes.

A significant feature of faceted schemes is that relationships are displayed, or implied, that do not exist in the traditional classification scheme. In addition to structural and semantic links, such systems also display functional relationships between terms. These may be notated or encoded or symbolised in some way, or they may be implicit in the ordering of terms.

In a similar way to the use of case and verb endings in inflected natural languages, some classifications express the functional dimension or role of the individual term notationally, regarding it as an additional element in the subject summary. Very detailed work was carried out in the fifties and sixties by many theorists such as Pages (1959), Farradane (1961) and others in attempting to analyse the nature of inter-term relations, or operators, and to define all possible examples.

The operator in a case like this defines the role of the term to which it is attached in just the way that case endings define the function of terms in a Latin sentence; the order of words is irrelevant. A survey by Perrault (1965) records and tabulates the variety of these, and attempts to provide one comprehensive and unitary list where all the various analyses are reconciled.

Others with an in-built citation order express functional status by the relative position of the concept in the citation string, mimicking the use of word order in non-inflected natural languages. It is not necessary to identify the relationship between terms as a separate entity in the class description, nor to notate the relation in any way. As in English the ‘citation’ order of words defines the meaning: ‘man eats dog’ is clearly quite different in conception to ‘dog eats man’.

Nevertheless, whatever the differences between the varieties of faceted classifications, there is a clear natural language analogy operating in all of them; in this they are quite distinct in their handling of subjects, from the older enumerative systems grounded in a philosophical structural view of knowledge and its inter-relations.

4. Machine Searching Methods

When we look at the methods used for handling information electronically, quite different criteria are employed. The tendency for search and retrieval devices developed by computer scientists is to disregard language based analysis in favour of mathematical relations, such as set membership, and to employ logical operators as constraints on searching rather than linguistic and functional controls within an index.
Here, the notion of relationships between terms is irrelevant (since they are non-existent) except at the point of search, and the only meaningful operation is the mapping of the search onto the database. There is no ‘organized collection’ in the conventional sense, and no correspondence between the search language and the field of search. The correspondences that are sought are those between the individual search terms and the content of the database. No attention is paid to the roles or relative positions of the terms in the search string.

5. The Purpose of Internet Searching

It is commonly assumed by librarians that because the emphasis of their retrieval activities is on ‘quality’ academic material, that this is the raison-d’etre of all searching. In fact the Internet is subjected to all sorts and kinds of searches, besides those simply looking for documents by subject content. It is a misassumption therefore to think that search engines and intelligent agents have been developed primarily with the needs of the academic researcher in mind; rather the commercial user, whose emphasis may be on file and data management or promotional activities, is the dominant end-user. Consequently, the results required from intelligent agents for example3, may be much less sophisticated in nature than for the average subject enquiry presented to a bibliographic database. Such searches may well be looking at the retrieval of personal names, domains, and other less complex properties of documents than intellectual content.


Search engines work by matching the terms in the search to terms on the database (whatever this might consist of (words in the title, words in the page, or words on the whole site) or the index compiled from these). The only relationships that can be exploited are those between the terms and the text of the database; these relationships are essentially mathematical in nature, rather than linguistic. Relationships between terms cannot come into the equation. This is in contrast to a conventional subject description using indexing languages where both syntactic and semantic relationships feature, whether these are categorised as explicit relations (expressed by relational operators or syntactic notation) or implicit relations (contained in a faceted structure with citation order).

7. Search Strategies on Major Engines

A number of preliminary difficulties occur for the searcher using standard available search engines. Normally users are encouraged to enter their terms in natural language and the fact that there is a search mechanism at all is usually not at all obvious. Even if the searcher understands the nature of machine searching, it is very hard to locate information or help on the mechanics of search, and specific descriptions of search criteria are usually well hidden. Additionally, search mechanisms vary from one engine to another (e.g. the default operators may be AND or OR, or various levels of proximity may be implicit in the search), and this is also rarely clear to the searcher. In fact, “the current trend with search engine designers is for each of them to have a rather different and secret way of rating pages.” (Wang et al, 1999)

8. Applications of Knowledge Organization Theory to the WWW

It is clear that these existing methods are not always successful or efficient at retrieval over very large databases such as the World Wide Web and there is increasing evidence of renewed interest in ‘traditional’ methods of subject analysis, as having relevance to new forms of information organization and management. One approach now would be to examine ways in which classificatory techniques of language analysis and control might be applied to supplement the currently available search methodologies.
9. Existing Forms of Classificatory and Thesaural Structures in Search Engines

Searching using a classification in the past has normally meant searching a classified file of some sort (whether this is the physical arrangement of documents, or as an ordered sequence of document substitutes in the form of a card catalogue or classmark index of an automated system). Where classifications have been used as retrieval tools dissociated from their storage function, this has normally been where 'metadata' of some sort has been assigned (whether this is on the Internet or searching on assigned subject descriptors on a library catalogue). But can classification only be used over an encoded database? Are we only concept searching, or can we deal with free text. Can the elements of vocabulary control inherent in an indexing language be built into the search mechanism of an engine, or is it true to say that over an unorganized data mass, searching can only be on the basis of matching of terms.

10. Use of Classificatory Structures by Search Engines

Early search engines did nothing more than match terms in the search to terms in the database. Relevance was assessed purely on numerical criteria such as the prominence of terms in the site in terms of frequency or early occurrence in the text. Results were usually mixed; very large numbers of hits could be guaranteed, but the usefulness of specific sites was very questionable. A survey of relevance assessment carried out for major search engines states that: "The biggest search engine problem is the 'needle in the haystack'. Usually there are too many matches ... and relevant documents can be found a few levels deep ... Thus the ability to refine a query in a sensible way is very important in improving the quality of search engines" (Wang et al., 1999).

11. Directory Based Search Systems

The first attempts to introduce structure into the database came with Yahoo's directory. Many in the field of computing who were unaware of systems such as the DDC, UDC or LCC regarded it as useful and highly innovative, but to the information professional it presented a very outdated picture of the hierarchical structure of knowledge, and one which disregarded modern theoretical developments in the field of classification. The major disadvantage however, was that like conventional library classifications, it was totally dependent on manual data input to make it workable.

Related attempts to improve the usability and quality of web resources (such as the construction of managed gateways, many of which used conventional classifications, or the assignment of metadata to selected sites) also required labour intensive methods; at this stage nobody thought of using traditional indexing methods other than in the traditional manner. Clearly the amount of material on the Web is unmanageably large; methods which rely on manual data collection and indexing (while they are valuable particularly in the educational context), are necessarily highly selective and are not viable options for serious searchers.

Nevertheless, recent initiatives such as the Open Directory Project (Open Directory Project, 2000) (in which volunteers are recruited to index sites) continue to advance this as a legitimate methodology, looking primarily at ways to decentralize (and cut costs of) the indexing task, rather than investigating means of streamlining it.

12. Types of Searching Based on Word Forms

Consequently the second generation of search engines began to consider linguistic rather than structural elements in the search string, and to utilise what appear at first to be 'thesaural' type relationships in the search framing.

'Fuzzy searching' or searching on less precise criteria for matching, in an attempt to retrieve words which are similar to the sought term, may appear at first sight to be
progressive. However the facility to retrieve a truncated word, a word stem, or a similar word using wildcards, is only an extension of the basic search process of specify and match. It may look more sophisticated but in fact widens the search rather than refining it, returning more hits rather than fewer. The search is fuzzy in that it is vaguer, but it does not really retrieve alternative (as opposed to additional) documents.

The ability to search on synonyms (or near synonyms) would represent a move forward and some search engines now appear to offer that function, notably Excite’s ‘concept based searching’ (Glossbrenner 1999). Excite attempts to link search terms with associated terms. It achieves this both by extending the original query (by substituting alternatives) or, via the Search Wizard device, offering the searcher alternative keywords to modify the search. These alternatives have the appearance of having come from a structured index or thesaurus underlying the search engine, but in fact they are derived from statistical analysis of the co-occurrence of terms in sites, and are not necessarily indicative of synonymy or of taxonomic relationships between terms. Specifically..."Excite’s developers avoid... natural language processing. They instead use their statistical concept extraction to improve the effectiveness of keyword queries. They chose to work with keyword queries because search databases for these systems are much easier to build and maintain." (Sonnereich, 1998) Such techniques offer useful strategies to the searcher, and represent an improvement on previous systems, but they do not appear to alter the basic search mechanism of mathematical matching.

13. Methods based on Inter-Subject Relationships

Other recent work on search engines has looked at a similar type of search refinement linked to the relationships between subjects and sites, by following hypertext links. Projects such as HyPursuit (Weiss, et al. 1994) have attempted to build clustering networks based on hypertext referencing, and the classificatory structures thus revealed are of a different quality to those described above. This appears to be an example of a searching system that considers subjects as concepts (rather than words) and examines inter-concept structures in a more extensive fashion. It looks to have much in common with bibliographic citation studies of the past, which examine the internal structure of study and research within a discipline.

14. Relations between Organized/Classified Collections and the Web

It is clear that although the ‘structures’ employed in search engines apparently reflect the structures of conventional indexing languages this is far from the case. Conventional indexing performs two quite distinct operations that are missing from search engine tactics. Firstly, it establishes context, and secondly it takes regard of the relationships between index terms (whether these are structural, as in older schemes or functional as in the case of modern faceted systems). Even when search engines appear to mimic this process, the links they make are usually based on word occurrence, statistical associations and frequencies, rather than logical and scientific links. Not only does this make the search engine manipulable by the cunning, it also prejudices against unlikely, and therefore potentially productive connections, because of their uncommonness. This is particularly unfortunate for academic searching. It seems likely therefore that traditional classificatory techniques can inform search mechanisms, whether of the search engine or of the end-user.

15. Structures of Classifications: Appropriateness of Different Classificatory Structures to Large Unorganized Databases

The essence of the web is that it is a large unstructured database, with no realistic hope of structure being imposed externally. Any search system therefore has to exploit types of relationships other than fixed structural hierarchies. There are two possible ways of doing this:
either to introduce other forms of relationship seeking into the search engine index; or to modify the search string at the search framing stage.

16. Faceted Classification and Search Strategies

We know that search strategies based on numerical criteria, such as word frequency and early occurrence in a text, can easily be manipulated to promote the retrieval of given sites. Search techniques formulated on a linguistic basis might prove to be a more effective instrument in accurate retrieval, since searches framed in this way should bear a closer relationship to the forms of natural language used by the searcher.

The faceted classification functions by imposing a pre-determined sequence on the terms in a search; this gives meaning to the relationship between terms and defines their role. There is no inherent order between the component concepts which are only linked at the time of indexing or searching. It thus appears that this might be an inherently suitable structure for searching over a large unordered database. The faceted classification can also support both of the possible modifications of search strategy described above.

17. Facet-Analytical Features applicable to Web Searching

The faceted structure is particularly appropriate as a basis for the construction of thesauri (Aitchison, 1986), many linguistic and semantic relationships being generated in a mechanical way from the classified structure. Synonym control, hierarchical relationships and many examples of related terms can be dealt with systematically from a faceted base. Semantic factoring of compound terms also becomes largely irrelevant in a facet based thesaurus. A faceted approach could therefore be used to develop a properly structured thesaurus to underlie the search engine, and to prompt appropriate modification of search keywords with synonyms and broader and narrower terms.

The second possibility is to modify the search string, using facet formulae to frame effective searches by ordering search terms. A current research proposal at the School of Librarianship, University College London (SLAIS, 2000), looks to investigate the feasibility of an on-screen help programme, that will offer the end-user assistance with the choice of keywords, and suggest appropriate structuring of the search for optimum effectiveness.

A third possibility, as yet unexplored, would be to investigate the use of interterm operators, or role indicators, in place of Boolean operators.

Notes
1. A detailed discussion of the issues involved in the conversion of UDC to a more fully faceted system is contained in McIlwaine, I.C. and Williamson, N. J. (1993) “Future revision of UDC; progress report on a feasibility study for restructuring.” Extensions & Corrections to the UDC 15 p.11
2. It is worth noting that in Perrault’s analysis of operators he includes the categories of standard citation order as developed in the UK in the 1960s, although it seems to me there is a clear distinction in nature between categories (‘roles’) and operators (‘links’) which in themselves express the relations between isolated or ‘uniterms’ in other systems (such as Farradane’s relational indexing).
3. Etzioni, O. & Weld, D. (1994) “A softbot-based interface to the Internet”. Proceedings of the ACM 37(7) p 72f Describes the development of an intelligent agent which refines searches on the basis of experience of previous similar activities. But there is no sense of analysis of the search, and the search criteria are very limited in range (proper names, domains, etc.). There is no suggestion that intellectual content of the document would be a search criterion, and consequently search modification on a subject basis is not an issue.
4. For example, resources on the NISS and SOSIG gateways are organized by the assignment of UDC classmarks
References


Farradane, J.E.L. (1952) “A scientific theory of indexing” *Journal of Documentation*


National Information Systems and Services at: [http://www.niss.ac.uk](http://www.niss.ac.uk) viewed 9.3.2000


Pages, R. (1959) “L’Analyse codee, technique documentaire en psychologie sociale et en science humaines” *Chiffres* (2) p. 103-22


SLAIS (School of Library, Archive and Information Studies), University College London (2000) *Project to investigate the suitability of dictionaries and classification schemes for interactive pre-processing of Web search queries; (Unpublished proposal for research into the use of classification schemes in framing searches for Internet searching)*

Social Science Information Gateway at: [http://www.sosig.ac.uk](http://www.sosig.ac.uk) viewed 9.3.2000


*Universal Decimal Classification* (1993) The Hague; UDC Consortium
