Claudio Gnoli
University of Pavia, Italy

Alberto Cheti
Municipality of Fucecchio (FI), Italy

Sorting documents by base theme with synthetic classification: the double query method

Abstract: Classification offers a unique power in allowing for systematic sorting of information items, thus playing an important role in the visualization of document contents and their relationships in the process of information retrieval. The majority of documents are about combinations of more than one concept. Therefore, classification notation representing the content has to be synthesized. As a classifier combines two or more classes from the schedules, the citation order of the notation elements affects the position of the document in sorted display. Among the concepts discussed in any document, a base theme and several particular themes can be identified. A general rule is that the notation representing the base theme should be cited first, thus producing a "helpful sequence" of compound classmarks. We propose a general method of information retrieval based on a double query combined with an appropriate systematic result display: classmarks starting with the searched concept should be displayed before those having it as an inner part. This principle is discussed on the examples of a simple interface for browsing digital assets currently being developed at the University of Pavia, and of ISKO's Knowledge Organization Literature online search interface.

Keywords: browsing; ordering; citation order; base theme; particular themes

1. The forgotten art of sorting

Many types of knowledge organization systems (KOS) are nowadays available: to the traditional ones developed for libraries – classification schemes and subject heading lists – more recent ones have been added, including keywords, terminologies, thesauri, taxonomies for websites, folksonomies (“tags”), topic maps, conceptual graphs, and formalized ontologies. Their developers, however, do not seem to be concerned with systematic sorting of items. Indeed, in these systems and their online interfaces sorting issues are managed mostly by alphabetical order, by popularity, by some relevance algorithm, or simply manually.

Still, such information units as spatial regions, chronological periods, stages in historical developments, logical steps in reasoning or in processes, and many others clearly need to be presented and organized in a systematic order: it would be inconvenient to see Friday before Thursday in a timetable, or divorce before marriage in a website providing community services, just because they file earlier in the alphabetical order (Gnoli, 2012*a*). It is also well known that systematic presentation, like in the table of contents of a handbook or on the shelves of a library, can act as an intellectual guide to their users by providing them with a general “map” of knowledge, through which they can navigate with

greater awareness and efficacy in order to find relevant information. Especially in need of meaningful order for being exploited in optimal ways are the vast amounts of digital collections. It must not be by chance that the original German term for knowledge organization, *Wissensordnung* that is literally “knowledge ordering” (Dahlberg, 1974), implies a notion of sorting that has got lost in its translations. Sorting of items in browsable systematic lists, in both menus of content options and display of search results, is thus one important component of information visualization.

Classification still offers a unique power for this function. Indeed, Ranganathan (1967) defines classification, first of all, as "helpful sequence" of classes. Systematic sorting is made possible in classification by a notational system: what is arranged alphabetically are not the verbal labels of concepts, depending on the hazards of lexicon in a particular language, but their notational equivalents, conceived to produce a meaningful order of concepts. This is achieved through a variety of technical devices using letters, digits and other characters as symbols for classes of increasing specificity and in various relations between them (Vickery, 1952-1958).

Notation lends itself very well to be recorded and managed by digital tools. This, however, has not always been implemented at its full power, even in existing library catalogues already containing large amounts of classified records (Bland & Stoffan, 2008; Rozman, 2009; Casson, Fabbrizzi & Slavic, 2011). Such a paradoxical situation appears to be due to a lack of collaboration between interface producers and information professionals using classification systems. Indeed, classification requires awareness of its inherent logic and complexity in order to be displayed in appropriate ways (Slavic, 2006). In this paper we will focus on the properties of compound classmarks and on reflecting their semantic richness as faithfully as possible when displaying classified lists.

2. Displaying compound classes

Most real documents do not deal only with a single concept, but with combinations of several ones: subjects of the kind of "the corrosion of tinplate by acid fruit products" are more numerous – and more in need of indexing – than general monographic ones like "fruit" (Foskett, 1958).

Notation for combined concepts has to be synthesized by taking its elements from different parts of classification schedules, according to the syntactic rules of the particular system adopted. This is described in literature as an analytico-synthetic process, as it involves analysis of the subject matter of a document into its conceptual components, then recombination of them in a single classmark. In this paper we will speak generically of synthetic classifications, irrespective of the particular syntactical devices that each of them adopts to rule synthesis (variously called common subdivisions, auxiliaries, facets, roles, links, phase relationships, etc.). Therefore, our discussion can be applied to any synthetic classification system.

As a classifier combines two or more classes from the schedules, their citation order becomes a relevant question. This is so especially as it will affect the item position in sorted displays: using UDC examples, while `1:34` "philosophy in relation to law" will be listed together with other items in philosophy, `34:1` "law

in relation to philosophy" will be listed in law (although, in a well-designed information system, it should also be retrievable by searching for philosophy).

Principles for identifying the concepts to be cited first within a classmark have not been discussed widely in classification literature, apart from the case of rigid facet formulas prescribed within disciplinary classes. Important insights, however, come from research on verbal subject indexing. Drawing on the lessons of PRECIS (Austin, 1984), a tradition of methodology for combining concepts to form subject strings has developed within the Research Group on Subject Indexing (GRIS) of the Italian Library Association (AIB) (Tartaglia, 1994; Cheti, 1996; Bultrini & Cheti, 2008; Cheti, 2008). As the combination mechanisms are basically the same in verbal indexing and in classification (Foskett, 1996), it is our thesis that these theoretical acquisitions can be fruitfully extended to classification (Gnoli, 2010).

3. Base theme and particular themes

GRIS method assumes that the subject matter of any document can be analyzed to identify a set of *themes*, the phenomena about which discussion is developed in the document, and of *rhemes*, the new information or ideas that are given in the document concerning the themes.

These notions come from text linguistics: documents can be viewed as texts structured in a series of conceptual statements (Beaugrande & Dressler, 1981; Cheti, 1996). Document titles summarizing their content usually are themes, like “Diet of wolves in northern Apennines”; occasionally they also include rhemes, like in “Wild ungulate abundance affects wolf diet in northern Apennines”. Classmarks or verbal subject headings will usually translate all contents into the form of a thematic phrase, of the kind of “wolf diet as affected by abundance of wild ungulates in northern Apennines”. This can then be translated into notation.

Among the various themes touched upon in a document, it is possible to further distinguish between a *base theme* and several *particular themes*. Base theme is the focus of the document discussion, like wolves and their diet in our example; hence it also is the most relevant concept to be considered when indexing that particular document. Notation (or term) representing the base theme, therefore, should be cited first. It will be followed by notation of particular themes, connected to the former by means of syntactical relationships (typically belonging to roles like property, agent, instrument, space, time, etc.). In our example above, wild ungulate abundance and northern Apennine are particular themes, connected to the base theme respectively by a relationship of cause and by one of place; as subject indexing prefers entities over processes to be taken as headings, it will be wolves to serve as the base theme, while diet can be treated as a property of them.

Logical connections between arguments in the macrostructure of the whole document have thus been translated into syntactical connections between concepts in the microstructure of its subject statement. In such microstructural plane, base theme corresponds to the *key concept*, or *lead*, or *system* as described in literature on subject indexing (Austin, 1969; 1984; Foskett, 1972).

If conceptual analysis of documents into base and particular themes is applied consistently throughout a collection, an optimal "helpful sequence" will result in the systematic index. This can yield benefits especially when browsing large collections of specialized documents.

4. Dealing with themes in information retrieval

In the digital environment, a prominent facility is searching and extracting relevant information from databases. As mentioned above, any concept expressed in synthetic notation can be extracted, be it the main theme or a particular theme, by means of a query looking for any string containing the corresponding notation. GRIS recommends that user interfaces allow this through a two-step search: first the concept of interest should be identified among the available classes or terms of the system, discarding its homographs and other sources of ambiguity; and only after selecting one concept should it be possible to examine all its combinations with other concepts, as either main or particular theme (Casson, Fabbrizzi & Slavic, 2011).

As a matter of fact, display of search results often loses part of the semantic richness of synthesized subjects. Besides compressing the two steps just described into a direct search for the characters input by the user, treated as an unqualified substring, most implementations present all retrieved items together, irrespective of whether the searched concept is the base theme or just a particular theme. Indeed, in systematic sorting, they are usually listed starting by those with a base theme filing before the others in the schedules (`1` philosophy in the UDC example above), which is not necessarily the one searched for by the user (e.g. `34` law). Rather, the latter will be scattered in several points in the list, according to both its filing value and its position in the occurring combinations.

In order to optimize display of retrieved items in a way reflecting the logic of synthetic classification, we then propose a general method of information retrieval based on a double query.

Given a search for class `X` input by the user, the system should perform:

1) a query for all notations starting with `X`, i.e. for documents having `X` as their base theme;
2) a query for all compound notations having `X` as an inner part, i.e. for documents having

x as a particular theme.

Results of query 1 should be displayed before results of query 2, as they match user's search more strictly.

In this way, the user will be informed first about the documents focused on the concept (s)he is interested in (provided any of them exist), and only after that about other documents also related with that concept though not as their primary focus. Clearly, this will fit the user's information needs more closely.

5. Two applications

A simple system of this kind is currently being developed by the University Library System (SIBA) at University of Pavia. Its purpose is allowing users to browse and search a catalogue of bibliographic and factual online databases that can be accessed from the University. More digital assets, including free subject gateways selected by librarians from the Web, are planned to be added as a further step and treated in the same way.

Each resource is indexed by one class of the Dewey Decimal Classification representing its base theme and marked as such, plus other classes representing particular themes. For example PubMed, the database of biomedical references and abstracts, is classified with 610 medicine as its base theme, and 570 biology as a particular theme. At user's selection of a knowledge domain, the system answers by showing first the records having it as their base theme, followed by those having it as a particular theme.

Another example of a synthetic classification using citation order of concepts in a meaningful way can be seen in the Knowledge Organization Literature, recently made available online (Gnoli, 2012*b*; ISKO, 2012). About 3,000 references on literature published in the field in the last two decades are recorded and indexed by the Classification Scheme for Knowledge Organization Literature, originally compiled by Dahlberg (ISKO, 1993). Most references are indexed by more than one class from the scheme. To form a compound classmark, apart from special combinations introduced by dashes or asterisks, classes are combined by listing them separated by a semicolon, having a syntactical value similar to that of colon in UDC.

Although default citation order can follow the same order of schedules, where the base theme of the indexed document is represented by a number that would file after those of particular themes, it is promoted to the first position. For example, the paper “The use of facets in Web search engines” by E. Milonas has 325 facet analysis as a particular theme and 757 search engines as the base theme: it is then classified as 757;325.

Users browsing the scheme and selecting class 757 launch a PHP page performing the following double MySQL query of the Literature table, that is part of the database:

```
$queryA = "SELECT * FROM `literature` WHERE `classmark` REGEXP
'^757*' ORDER BY classmark";
$queryB = "SELECT * FROM `literature` WHERE `classmark` REGEXP
';757*' ORDER BY classmark";
```

The first query searches for the regular expression ^757*, that is for all classmarks starting with 757 as their first characters, followed by any others. The second query searches for the regular expression ;757*, that is for all classmarks including 757 only after a semicolon separating it from a previous class. Results of each query are sorted by classmark.

In the results page, records matching the first query are displayed before records matching the second query. As the mentioned paper has 757 search engines as its base theme, it will be displayed in the first array of results, which are the most relevant for this search. On the other hand, if the searched class were 325 facet analysis, it would instead be displayed in the second array.

6. Conclusion

We have shown how principles of subject indexing, concerning the representation of base and particular themes of a document, can be extended from verbal subject indexing, where they have been already formulated in literature with some detail, to classification. In the latter context they have been hardly acknowledged in these terms until now, although often adopted implicitly in the application manuals of classification schemes, where these prescribe to cite notation for the class representing a focus of discussion in the document before the other components of a compound.

We have then shown how these principles can be easily applied to search interfaces and display of results, basically by a script including a double query. Although in our example we used PHP and MySQL, our approach can be implemented in any other programming language allowing to operate on string functions. These resources are widely available at no additional cost. However, they are generally not used in ways fitting the principles of classification, mainly due to poor communication between Web professionals and classification professionals.

Particular implementations using such principles in implicit, empirical ways may well have existed here and there already. However, we believe that their explicit formulation and conscious application can be of great utility to exploit the semantic power of classification on a larger scale.

References

Austin, D. (1969). Prospects for a new general classification. *Journal of Librarianship*, 1 (3), pp. 149-169.

Austin, D. (1984). *PRECIS: a manual of concept analysis and subject indexing*, 2nd ed. London: British Library.

Beaugrande, R.A. de; Dressler, W.U. (1981). *Introduction to text linguistics*. London-New York: Longman. Also available at: http://beaugrande.com/introduction_to_text_linguistics.htm

Bland, R.N.; Stoffan, M.A. (2008). Returning classification to the catalog. *Information Technology and Libraries*, 27 (3), pp. 55-60.

Bultrini, L.; Cheti, A. (2008). The Italian model. Presentation at session Conceptual models of aboutness, 10th International ISKO Conference, Montreal, 5-8 August 2008.

Casson, E.; Fabbrizzi, A.; Slavic, A. (2011). Subject search in Italian OPACs: an opportunity in waiting? In *Subject access: preparing for the future*. Edited by P. Landry, L. Bultrini, O.T. O'Neill, S.K Roe. Berlin-Boston: De Gruyter Saur, pp. 37-50.

Cheti, A. (1996). Testo e contesto nell'analisi concettuale dei documenti [Text and context in the conceptual analysis of documents]. In *Il linguaggio della biblioteca: scritti in onore di Diego Maltese*. A cura di M. Guerrini. Milano: Editrice Bibliografica, pp. 833-855.

Cheti, A., (2008), Il punto di vista del GRIS sulla "relazione di soggetto" in FRBR [GRIS's viewpoint on the “subject relationship” in FRBR]. In *Principi di catalogazione internazionali: una piattaforma europea? Considerazioni sull'IME ICC di Francoforte e Buenos Aires*, atti del convegno internazionale, Roma, Bibliocom - 51o Congresso AIB, 27 ottobre 2004. A cura di M. Guerrini. Roma: AIB, pp. 91-100. Also available at: http://www.aib.it/aib/congr/c51/chetint.htm

Dahlberg, I. (1974), *Grundlagen universaler Wissensordnung*. Munich: Verlag Dokumentation.

Foskett, A.C. (1996). *The subject approach to information.* London: Library Association.

Foskett, D.J. (1958). Library classification and the field of knowledge. London: Chaucer House.

Foskett, D.J. (1972). Information and general systems theory. *Journal of Librarianship*, 4 (3), pp. 205-209.

Gnoli, C. (2010). Themes and citation order in free classification. *IASLIC Bulletin*, 55 (1), pp. 13-19. Also available in *DLIST*, http://arizona.openrepository.com/arizona/handle/10150/111813

Gnoli, C. (2012*a*). L'arte dimenticata dell'ordinamento: perché non tutto deve andare dalla A alla Z [The forgotten art of sorting: why not everything should go from A to Z]. In: *6o Summit italiano di architettura dell'informazione, Sesto San Giovanni 10-11 maggio 2012*. Available at: http://mate.unipv.it/gnoli/ordinamento.pdf

Gnoli, C. (2012*b*). KO Literature now searchable online, Knowledge organization, 39, n. 4, p. 304.

ISKO (1993). Classification Scheme for Knowledge Organization Literature. *Knowledge Organization*, 20 (4), pp. 211-222.

ISKO (2012). Knowledge Organization Literature. Literature editor Hur-li Lee, database editor C. Gnoli. Available at: http://www.isko.org/lit.html

Ranganathan, S.R. (1967). *Prolegomena to Library Classification*, 3rd ed., Part F. Bangalore: SRELS. Also available in *DLIST*: http://arizona.openrepository.com/arizona/bitstream/10150/106370/7/ProlegomenaF.pdf

Rozman, D. (2009). The practical value of classification summaries in information management and integration. Paper presented at Classification at a crossroads: multiple directions to usability, The Hague, 29-30 October 2009. *Extensions & Corrections to the UDC*, 31, pp. 275-284. Also available in *DLIST*: http://arizona.openrepository.com/arizona/handle/10150/199893

Slavic, A. (2006). Interface to classification: some objectives and options. *Extensions & Corrections to the UDC*, 28, pp. 24-45.

Tartaglia, S. (1994). Per una definizione di 'soggetto' [For a definition of 'subject']. In *Il linguaggio della biblioteca: scritti in onore di Diego Maltese*. Firenze: Giunta regionale toscana.

Vickery, B.C. (1952-1958). Notational symbols in classification. *Journal of Documentation*, 8: 1952, pp. 14-32; 12: 1956, pp. 73-87; 13: 1957, pp. 72-77; 14: 1958, pp. 1-11.

About authors

Claudio Gnoli has been working as an academic librarian since 1994. His main research interests are the principles of knowledge organization and their application to digital assets. He is a current vice-president of the International Society for Knowledge Organization and a member of UDC Editorial Board.

Alberto Cheti has been holding technical and managing positions in library and social services for the Municipality of Fucecchio (Tuscany, Italy) since 1979. He has been a leading researcher in the Subject Indexing Research Group (GRIS) of the Italian Library Association (AIB), lecturing in university classes and contributing to the structure of Nuovo Soggettario, the national subject indexing system.