Workshop on
Information Resource Management
13th-15th March,, 2002
DRTC, Bangalore

Paper: CG

Organizing the Web: A Faceted Approach

Dimple Patel

Documentation and Research Training Centre

Indian Statistical Institute

Bangalore-560 059

email: *dimple@isibang.ac.in*

Abstract

Today, the Internet is a popular media for information communication. Like a library it is a storehouse of information albeit in electronic format. But the amount of information available on the Internet is vast and unlike the library, the information is disorganized, chaotic and without systematic and semantic structure whatsoever. Due to which the users often end up with irrelevant results for their queries. Classifying and cataloguing the Internet would be a step forward in organizing it. This paper examines the need for classifying the Web resources; looks at the pros and cons of using library classification, as an organizational tool for Web resources; and the current usage of library classification schemes to organize the Internet (eg. BUBL Link). It further investigates the suitability of freely faceted classification scheme like Colon Classification to the 'dynamic' nature of the Internet compared to widely used enumerative schemes like DDC or LCC.

1 Introduction

"Classification is an uncovering of the thought-content of a written or expressed unit of thought...The reference librarian....applies the classification scheme in the ultimate stage of library service which is effecting contact between the right reader and the right unit of thought in a personal way." (Ranganathan, 1951)

Man has an innate need to organize (Taylor, 1999). The need to organize information lead to the development of various organizational tools like classification schemes, catalogues, indexes, etc. The idea behind developing these tools for storing information was to help users in easy access and quick retrieval of relevant documents in the library. There has been a rapid change in the way information is generated, stored and accessed in the last decade. Internet allows information to be generated and accessed in the digital form. Though likened to a library, the information available on the Internet is overwhelming, about 200,000,000 web pages (Bradley, 1999). And it is growing by the second. To find relevant information within this huge amount is not an easy task. Attempts are on to classify and catalogue the Internet resources to bring about some order to improve access and retrieval efficiency.

Taylor defines classification as the *"...the process of determining where an information package fits into a given hierarchy and then assigning the notation associated with the appropriate level of the hierarchy."* In practical terms, classification is about clustering information objects to make them easier for the user to find, and to gather together similar objects. Most of the websites follow some kind of classification to organize their resources. Most of them use their own 'home-grown' classification scheme (eg. Yahoo!), but a few classify their resources by using traditional classification schemes like DDC, LCC, UDC, etc. (eg. GERHARD - the German academic Web index classifies all documents using the UDC classification in three languages)

2 Need for classifying Internet documents

"Regardless of the nature of the information resource, the need to express its content, describe its format, facilitate its access, and enable its use remains constant" (Dillon & Jul, 1996).

The body of knowledge on Internet lacks a systematic structure. Many of the search engines use sophisticated indexing techniques to organize their resources. Search engines such as Google and Altavista may provide advanced Boolean searching, but the fact is that most users simply type in a few words to describe a concept and sort through the results. Some of the search engines including Yahoo! and Google, organize websites by subject categories, based on their own organizational schemes. Since each search engine has its own vocabulary and organizational scheme, a search with a keyword may result in too many hits using one search engine, while the same search using another search engine may give no hits at all. Classification brings systematic order and control to a collection so that an information package can be retrieved according to a particular aspect of its character. Hence, classification is as important for Internet resources as it is for books and other media.

3 Pros and Cons of classifying Internet resources

"The advantage of using library classification is that it can create cohesion across diverse information stores by establishing a shared conceptual context" (Albrechtsen & Jacob, 1998).

3.1 PROS

A site that organises knowledge with a classification scheme demonstrates several advantages compared to other sites, which do not (Svenonius 1983):

- **Browsing:** Classified subject lists are easy to browse in an online environment, particularly for inexperienced users or for users not familiar with a subject and its structure and terminology.
- **Classification notation does not even need to be displayed** on the screen so an inexperienced user can have the advantage of using a hierarchical scheme without the distraction of the notation itself.
- **Broadening and narrowing searches:** Classification schemes are hierarchical and therefore can be used to broaden (i.e. for improved recall) or narrow a search when required.
- **Potential to permit multilingual access to a collection:** Since classification systems often use notations independent from a specific language, indices in different languages can offer multilingual access to the same resources without any further changes to the collection. The classification system can be used as a switching language to retrieve resources in any given language on the subject.

3.2 CONS

There has been criticism regarding classifying Internet documents. Those against classifying Internet documents state the following reasons:

- **The division of logical collections of material:** classification schemes often split up collections of related material. But this can be partly overcome with good cross-references.
- **Adapting to new areas of interest:** classification schemes, since they are usually updated through formal processes by organized bodies, often reveal difficulty in adapting to rapidly emerging new areas of study.

These problems can be overcome by providing good cross-references, by modifying library classification schemes to suit online documents and by using automatic classification to update the classification schemes without delay.

4 Use of library classification schemes to organize the Internet

A number of projects have applied LCC, DDC, UDC and other schemes to the Web. 'Beyond Bookmarks: Schemes for Organizing the Web' (*http://www.iastate.edu/~CYBERSTACKS/CTW.htm*) and 'Classification on the Net' (*http://web.simmons.edu/~schwartz/myclass.html*) provide links to the Web sites using classification schemes such as LCC, DDC. Some of the websites that organize web resources using classification scheme like DDC, UDC, LCC, etc are listed in Appendix I. Currently the most widely used scheme is the DDC.

5 Problems in using Enumerative Classification Schemes

"While there are many discipline-based listservs and collections of documents on the Internet, I believe that its primary use is currently, and will continue to be, searching for specific pieces of information." (Weinberg, 1996).

Many websites are using DDC and LCC to organize Web resources. But these schemes have the following disadvantages:

- Enumerative schemes like DDC and LCC attempt to exhaustively map the whole universe of knowledge. Hence these schemes have large schedules.
- Because of the dynamic nature of knowledge, it is impossible list 'all' the subjects in universe of knowledge.
- Due to their rigid hierarchy, these schemes are less hospitable in accommodating new as well as interdisciplinary subjects.
- As stated by Weinberg, the Internet is used to search for specific information. And enumerative schemes are not suitable for classifying very specific topics.
- While using enumerative schemes, the classifier, is often compelled to choose a number to fit the information package. Many important aspects of the subject of the item have to be omitted because the schedules do not list them.
- Due to the hierarchical nature of these schemes, it is hard or inconvenient to reclassify a topic if a mistake is uncovered, or the field evolves in an unexpected

direction, or a better structure is discovered. Many major websites are restructured frequently which result in broken links.

- Some of the classification schemes are prone to subjectivity and cultural bias. For instance, DDC gives an in-depth classification for Christianity whereas other religions of the world are neglected. And LCC gives priority to military and naval science.
- Some of the other disadvantages of schemes like DDC and LCC are
  - may require complex or lengthy notation,
  - are often difficult to use to locate materials,
  - may not provide for enough coordination of terms,
  - may not meet the needs of the individual or special library,
  - not provide enough detail to accurately describe all subjects in all media

6 Using Faceted Classification Scheme for organizing the Internet

6.1 FACETED CLASSIFICATION

Definition

Faceted classification is also called as analytico-synthetic. It is named after the two main processes involved in the composition of a call number. The two processes are:

Analysis: Breaking down each subject into its basic concepts.

Synthesis: Combining the relevant units and concepts to describe the subject matter of the information package in hand.

A facet is a "clearly defined, mutually exclusive, and collectively exhaustive aspects, properties or characteristics of a class or specific subject" (Maple 1997)

Example

Suppose a book is titled *"Research in the cure of Tuberculosis of lungs by X-ray conducted in India in 1950".* The facet analysis of the title of the book using a faceted classification like Colon Classification reveals the different facets of the subject, like:

Medicine,Lungs;Tuberculosis:Treatment;X-ray:Research.India'1950

It covers all the significant aspects of the subject of the item. Such a classification scheme is considered to be "hospitable" to all sorts of complex topics. It is therefore a "dynamic" scheme.

6.2 ADVANTAGES OF USING FACETED CLASSIFICATION

- **Offers flexibility for interdisciplinary subjects:** Since only the basic concepts are given in the schedules of a faceted classification, even if a document has an interdisciplinary approach, the classifier can combine any concept with another to produce a tailored call number for the specific item.
- **Ease of accommodating new concepts:** When a new subject develops, a classifier using an enumerative scheme will have to wait until the scheme provides a term for that particular subject. On the other hand, using faceted classification, since we start with the basic concepts and build towards our topic, it is much easier to combine already-existing terms to form a new subject.
- **Specificity of the subject can be expressed:** The example given above is illustrative of the details that can be expressed by using a faceted classification scheme.
- **Citation order is not essential:** The citation order of the class number is not relevant as the user can retrieve the documents with any combination of the facets. This is especially useful on Internet, where the topics are widely scattered.
- **Extra facets can be added:** Extra facets indicating the scope of the document, it’s target group, language, point of view can be easily added.

6.3 DISADVANTAGES OF USING FACETED CLASSIFICATION

It is not that Faceted Classification is without disadvantages. Some of its weaknesses are:

- Unlike DDC and LCC, they are very unfamiliar to most of the users.

- They are very complex and are difficult to use without experience.
- It is difficult to integrate browsable hierarchy of the schedules.

But these problems can be solved by providing a interactive user interface, which can explain the facets to the user; which will prompt the user with queries to enter one topic at a time (from the broader to the specific); and it can also have different levels of interfaces for different kinds of users (expert search, easy search, etc.).

6.4 COLON CLASSIFICATION AND THE INTERNET

Since the advent of Internet and the WWW there has been bludgeoning of information. The issues involved in organizing this huge amount of information are the emergence of new interdisciplinary subjects, the use of the WWW for searching specific information, etc. To address these problems Colon Classification can be considered for organizing the Web resources. Being a freely faceted classification it is easy to classify the highly scattered Web resources. One of the major disadvantages of the Colon Classification is its lengthy notation, which makes it difficult to put the call number on the spine label of a book, or the user to memorize the whole notation to locate the particular item in the library. However, this problem is not encountered on the Web, since we do not have to worry about the "physical" location of a document as well as notation of the document.

The implementation of the Colon Classification scheme on the Web may be useful in that it provides the website creator, the indexer, and the user a common language to describe and identify the content of the page. Since the Colon Classification is very accommodating to new concepts and interdisciplinary subjects, it is quite appropriate for the fast-growing web environment.

Every query is unique and comes from a specific perspective. What is relevant to one user is different from another. The advantage of using a faceted scheme like Colon Classification in retrieval of information from the Web, is that it is more attentive to the user's need, giving him the freedom to search documents using any combination of facets.

7 Conclusion

The Internet lacks a systematic structure. Classifying and cataloguing Internet documents is a step forward in organizing it. Though, currently DDC is being used widely to classify Web resources, it has many disadvantages like rigid hierarchical structure, bias, lack of specificity, etc. A faceted scheme of classification is more suited to the dynamic nature of the Internet. Due to it's flexibility a faceted scheme can keep pace with the ever increasing information on the Web and also the rapid emergence of new topics. In this regard Colon Classification, which is freely faceted, is a worthy candidate for organizing the Internet and further research, to implement Colon Classification in organizing the Internet, can be undertaken.

8 References

1. **Albrechtsen (Hanne) & Jacob (Elin).** The Dynamics of Classification Systems as Boundary Objects for Cooperation in the Electronic Library. Library Trends, 47 (2), pp.293 – 312, 1998.
2. **Bradley (Phil).** The advanced Internet searchers handbook. London: Library Association Publishing, 1999.
3. **Dillon (Martin) and Jul (Erik).** Cataloging Internet Resources: The Convergence of Libraries and Internet Resources. *In* Cataloging and Classification Quarterly, 22 (3/4), pp.197 – 238, 1996.
4. **Iyer (Hemlata).** Classificatory structures: Concepts, relations and representation. Textbook for Knowledge Organization, Vol. 2. Indeks Verlag, Frankfurt/Main, 1995.
5. **Ranganathan (S. R.).** Classification and Communication. Delhi University Publications, Library Science Series, 3. Delhi: University of Delhi, 1951.
6. **Svenonius (Elaine).** Use of classification in online retrieval. *In* Library Resources and Technical Services, 27(1), pp.76-80, Jan./Mar. 1983.
7. **Taylor (Arlene G).** The Organization of Information. Englewood, Colorado: Libraries Unlimited, 1999.

8. **Vicine-Goetz (Diane).** Online Classification: Implications for Classifying and Document{-like Object} Retrieval. *In* Knowledge Organization and Change: Proceedings of the 4th International ISKO Conference, 15-18 July, 1996, Washington, D.C., Rebecca Green, ed. Frankfurt/Main: Indeks Verlag, 249-53, 1996.
9. **Scorpion**: Automated classification of the Web. *http://www.greencoast.ca/sitez/Scorpion/*
10. **The** role of classification schemes in Internet resource description and discovery *http://www.ukoln.ac.uk/metadata/desire/classification/class_1.htm*

Appendix I

Universal Decimal Classification (UDC)

Directory of Networked Resources: UDC "Shelfmark" Order (NISS Information Gateway)
http://www.niss.ac.uk/subject/index.html

GERHARD: German Harvest Automated Retrieval and Directory
www.gerhard.de/gerold/

Dewey Decimal Classification(DDC)

ADAM: Art, Design, Architecture & Media Information Gateway. Index+ Dewey Search
http://adam.ac.uk/advanced/dsearch.html

Browse LINK by DDC (BUBL)
http://link.bubl.ac.uk/ISC2

Library of Congress Classification (LCC)

Cooperative Online Resource Catalog (CORC)
http://www.oclc.org/news/oclc/corc/index.htm

CyberStacks(sm)
http://www.public.iastate.edu/~CYBERSTACKS/

National Library of Medicine (NLM)

List of Subject Headings, By NLM Code (OMNI - Organising Medical Networked Information)
http://omni.ac.uk/listings/numlist.html

NLM Classified Subject Index (Kuopio University Virtual Library)
http://www.uku.fi/ROADS/subject-listing/Default/numlist.html