

# Using a Classification-Based Information Space

**Lorraine F. Normore**

OCLC Online Computer Library Center, Inc.  
Office of Research  
Dublin, OH 43017 USA  
+1 614 761-5263  
normorel@oclc.org

**Mark Bendig**

OCLC Online Computer Library Center, Inc.  
Office of Research  
Dublin, OH 43017 USA  
+1 614 764-6072  
mbendig@oclc.org

## ABSTRACT

Currently, visualization is being used to aid understanding of text-based information searches by members of both the information retrieval (IR) and the human-computer interaction (HCI) communities. The dominant visualization model, its origins and issues raised by its use, are explored in this paper. An alternate model based on a systematic classification system is proposed and an application based on this model is described.

## INTRODUCTION

Scientific data visualization was developed to explore ways to aggregate and summarize large quantities of numbers to aid data interpretation. Graphic representations were sought because, as Tufte [1] says, “of all methods for analyzing and communicating statistical information, well-designed data graphics are usually the simplest and at the same time the most powerful” (*Introduction*).

In scientific fields, the inherent data metric underlying the visualization is often apparent. For quantitative data, frequency of occurrence of a dependent variable is plotted against an independent variable whose values are arrayed along a measurable, regularly varying base. For example, temperature changes may be plotted against time to show patterns in the data. (See [2] for many good examples of these visualizations). In other cases, data of interest may be shown against a surface that has an obvious mapping to the real objects with which those data are associated. Thus, in a home-finding application, homes that meet specified parameters were mapped onto a spatial map [3] that clearly showed the distribution of locations in the target area. Similarly, in Maya Design’s powerful Visage application [4], map displays are used as a background metric to show the spatial deployment of military forces. When such well-known background metrics are available, the resulting visualizations are often intuitive and have great impact on the viewer.

The field of information visualization grew out of the field of data visualization. As interest in using the power of visualization to assist in understanding textual data has grown, the need for an underlying metric has become manifest. A single model for “information space” has become widely adopted. It derives from the confluence of work in the 1960’s in information retrieval [5], cognitive

psychology [6] and multidimensional scaling [7]. Because this model dominates the field of text visualization, it is important to identify its working assumptions and potential issues for its unquestioned adoption.

## THE INFORMATION SPACE MODEL

This model depicts a multidimensional information space in which words/concepts are arrayed. To create the space, to place items in it, and to establish relationships among items, variants on similarity-based clustering and multidimensional scaling are commonly used. Typically, the co-occurrence of items within some pre-defined organizational unit (a sentence, paragraph, section, entire article) is taken to indicate a bond of similarity between the units. This is taken as a justification for placing them in a common cluster and the process is repeated for the set of items. Clusters are often reiterated to maximize the extent of overlap among the items, which the component clusters contain. To create a space, a scaling solution places clusters that have a greater degree of overlap closer together while clusters with less overlap are placed further apart. Sometimes relationships are shown as arcs connecting related entities. Sometimes distance alone conveys the strength of the relationship.

## Issues for the Information Space Model

A major benefit of the information space procedure is its ability to provide an ad hoc analysis of unstructured collections. Thus, it has been widely used to provide a framework for showing relationships among unstructured full-text sources, including web pages.

A persistent problem for the information space model, however, is the lack of a communicable dimensionality. There is no effective scheme for telling users about the space; there are no supporting semantics. Clustering can group documents, but it does not automatically provide informative labels for the groupings so produced. Users must infer the characteristics of a cluster by examining the contents of the cluster. The process of multidimensional scaling produces an n-dimensional space that can be reduced to a limited number of dimensions mathematically. However, as practitioners of the art know, the semantic quality of the space is determined by looking at the way in which category exemplars distribute over the space and by then inferring the nature of the dimension from that

distribution. Both processes are challenging cognitive tasks. In both cases, users must take on the load of determining the nature of the space and of the units within it.

## **A CLASSIFICATION-BASED SPACE**

### **Library Collections**

Library collections differ in significant ways from unstructured document collections. Library catalogs contain structured metadata about the documents they include. Among the metadata normally available for cataloged items are intellectually assigned controlled subject headings and classification numbers. These structured metadata elements have been formally developed to create a rich semantic framework for the information they describe. Clustering could be used to group similar library catalog records using assigned subject terms. Because the assigned subject terms have been regularized, clustering them would provide increased homogeneity in the derived space. However, as is true for clustering unstructured text in general, users must still take on the load of understanding the results of the clustering and scaling analyses. Therefore, the approach proposed by this paper is to use the assigned classification numbers, which are highly structured and semantically rich, to provide an underlying metric for the information space.

### **Benefits of classification systems**

Although dividing library content into broad subject categories is commonly traced as far back as the *Pinakes* of Callimachus in the third century B.C.E., the relatively small size of document collections in the following millenium made an extensively divided categorical scheme unnecessary to adequately segment the limited content. However, beginning in the 16<sup>th</sup> century with the growth of the literature enabled by the printing press, many different classification schemes were developed, as Taylor [8] points out. The most commonly used classification systems today in North America, the Dewey Decimal Classification (DDC™)[9] and the Library of Congress Classification (LCC) System [10], began in the last quarter of the 19<sup>th</sup> century.

Classification systems attach meaningful labels to categories they include. These labels (captions) are often enhanced by the semantic information in notes and associated relative index terms. Work by Vizine-Goetz and her colleagues (e.g., [11]) have provided additional semantic information by creating mappings between subject term-based metadata and the class numbers of the DDC. These can be used to add semantically meaningful information for users.

Another advantage of using classification systems for library collections is the support they have. The DDC and LCC are continually under review by trained editors and continue to evolve. Thus, a knowledge view derived from their structures remains viable over time and will reflect changes in the structure of knowledge as changes occur. In

addition, both are widely used in North America in public and academic libraries. The DDC has also been translated into many languages and is a system of choice around the world.

## **NAVIGATION AND CONTEXT MAINTENANCE**

A second area of concern for any information space model is providing methods of dealing with the quantity and complexity of material in the space so that users can maintain a sense of where they are in the space and so that they can successfully navigate within the space.

In the cluster-based “information space model”, there is neither an underlying metric nor labels to define the dimensionality. This makes it difficult for users to comprehend and navigate successfully through dense information spaces. Theorists using information space models have created a number of methods to overcome these deficiencies. The following two applications show different methods for dealing with the navigation problem. SemNet [12] was intended to be used as a general-purpose tool for arbitrary knowledge bases. To reduce the amount of information displayed in the space, a variety of techniques were used, the most famous of which is the Generalized Fisheye View which displays detail near a focal point and only “important landmarks” further away. The second illustration is found in the work of Pirolli and his colleagues at Xerox PARC. They created a process called Scatter/Gather to browse large text collections [13] which used document clustering to define the space. To reduce the complexity, the same technique is then applied recursively to the resulting groups, producing a cluster hierarchy. Users could then work their way through the hierarchy, revealing detail upon demand.

Classification-based spaces, like the DDC and LCC, were created as hierarchical information structures. The DDC, in particular, has features that support its use as a metric for describing information spaces. It is deeply hierarchical and systematically divides categories into regular decimal subunits. The ten top level classes are each divided into ten sections which in turn are divided into ten sections and so on, as long as divisions are required to describe the content. The decimal notation that the DDC uses is comparatively easy to understand and directly reflects the categorical structure of the classification system. For example, the DDC general category Technology/Applied Sciences is 600, while kinds of applied sciences vary at the “decimal” level (e.g., Medical science at 610; Engineering at 620). Not only is this structure repeated throughout the system, the DDC also creates parallel structures across the whole classification. So, for example, the study of the French language is 440 while French literature is 840. In either classification system, the hierarchy inherent in the structure can be used to group related materials into high level categories and then to subdivide those categories using the pre-established order created by the classification theorists.



In Demo mode, FullView's 3D visualization system is used to display the results of any one of five queries issued against the WorldCat database using OCLC's FirstSearch system. The queries, chosen to represent different areas of the database, were all subject ("su") searches. The topics were: (1) french and cook\*; (2) garden\* and flower+; (3) pet+ and disease+; (4) gas and product\*; and (5) history and revolution\* and europe\*. Including only those records that contained DDC numbers produced results sets varying in size between 446 and 4,452 records. The results sets were saved in an intermediate data file. Data extraction software took selected fields (e.g. DDC number, title, etc.) from each record. The data extraction software produced a second set of files in an appropriate format for import into tables contained in an Access database. It is these database tables that are processed and displayed by the FullView Demo prototype.

### **FullView Application**

The Client uses JavaScript and VRML to produce the HTML page in which the application is displayed. The Server includes Active Server Page (ASP) scripts that provide most of the application's functionality. Although the server could be local or networked, Microsoft's Personal Web Server (PWS) has been used for *FullView* development. The final major element of the *FullView* application is the Database Engine. Microsoft's JET engine is used to access the databases, under the control of *FullView's* ASP scripts. The JET engine also provides database functionality for Microsoft's *Access*, which was used to define and load the database.

Although the ASP scripts that run on the Server provide most of the application's functionality, two significant client-side application components are downloaded to the Client and subsequently operate in that environment. These distributed components are described in the next section, which discusses *FullView* operation.

### **FullView Operation**

A *FullView* session begins when the user requests the application's Main Page from the Server. The Main Page display shown in Figure 1 contains three HTML frames which allow the user to enter a query, to view the 3D graphical display representing the search results, and to display lists of database records associated with particular components of the 3D display. Embedded within the HTML code for the Main Page are the two client-side application components mentioned above: (1) a section of JavaScript code, which checks user input for disallowed entries and submits the query to the Server; and (2) a section of VRML code that specifies the resulting 3D "world" seen by the user.

When users issue a query or select a results set, the Client sends a query to the Server and an ASP script residing on the Server is executed. The ASP script is a regular HTML page with sections of server-side scripting embedded within

it. The ASP script is responsible for actuating the JET Database Engine and issuing the specified user query against the *FullView* database. Occurrences of the target string are located and a list of the associated DDC numbers is compiled and then sorted into 100 "pockets", numbered 00 through 99, based on the first two digits of the DDC number. Thus, 004.6 would go into pocket 00, 538.84 would go into pocket 53, and so on. The count of DDC numbers in each pocket becomes the basis for the specification of a 3D display portraying a special-format bar graph. In the graph, 100 pillars, representing the 100 pockets, are arranged in a 10 x 10 matrix. The height of each pillar is proportional to the record count for the associated pocket.

The ASP script is also responsible for receiving the search results from the Database Engine and using them to generate the corresponding VRML "world" specification, as described below. This VRML code, generated "on the fly" is then embedded into the HTML page that is sent to the Client in response to the query.

The VRML code is generated by means of a VRML Template File. This Template File is read in and interpreted by the ASP script discussed above when it is time to generate the VRML output. In addition to actual VRML code, the Template File contains formatting commands, which are acted on as they are read in by the ASP script. The Template File also contains data placeholders that are replaced with the appropriate data at the time the file is interpreted.

Once a 3D graph has been generated, users can click on any of the pillars comprising that graph. In response to the click, users may either Drill Down to the next level of detail for a given column (the default) or choose to List Contents of the chosen column.

The Drill Down option could be thought of as "expanding" the contents of the selected pillar--sorting its contents into 100 sub-pockets based on the third and fourth digits of each DDC number and producing a new 3D display showing the record counts for the resulting 100 sub-pockets. Thus, drilling down into the pillar representing the 530's in the original graph results in a display with pillars for 530.0, 530.1, and so on through 539.9. This process can be repeated to the resolution limit of the database.

The List Contents option displays the list of DDC numbers contained in the pillar and their captions (if any). This display appears in a separate browser frame, which is independently scrollable. Once the DDC numbers are displayed, users may choose to see associated LCSH headings by clicking on that option in the radio buttons at the top of the frame containing the list output. In Demo mode, users can also request a display of the DDC numbers and captions and titles of the bibliographic records that contain the related DDC number.

## EXPERIENCE TO DATE

The prototype has been demonstrated to groups including the editors of the DDC. Informal testing has also been carried out with colleagues. Response has been positive. Viewers find the visualization interesting, those who use the application tend to explore its features for extended periods of time. Users also seem to be comfortable with navigating the application and find its metaphors intuitive. Although these studies have focussed on the DDC, we believe that similar benefits could be gained from the use of other systematic classification systems. Our immediate goals are to explore ways to embed this approach within real-world applications.

## REFERENCES

1. E.R.Tufte. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press, 1983.
2. P.R. Keller & M.M. Keller. *Visual Cues: Practical Data Visualization*. Piscataway, NJ: IEEE Press, 1993. P. 67.
3. C. Williamson & B. Shneiderman. The dynamic home-finder. *SIGIR Proceedings*, 1983, 339-346.
4. Maya Design Group. Visage base site. Nov. 29, 2000. <http://www.maya.com/Visage/base/>
5. G. Salton & M.E. Lesk. The SMART automatic document retrieval system. *Communications of the ACM*, 1965, 8(6), 391-398.
6. C.E. Osgood, G.J. Suci & P.H. Tannenbaum. *The Measurement of Meaning*. Urbana, IL: University of Illinois Press, 1964.
7. Shepard, R.M. Stimulus and response generalization: tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology*, 1958, 55, 509-523.
8. A.G. Taylor. *The Organization of Information*. Englewood, CO: Libraries Unlimited, 1999.
9. OCLC. Forest Press. *Dewey Decimal Classification*. <http://www.oclc.org/dewey/index.htm>
10. Library of Congress Cataloging Directorate. Home page. <http://lcweb.loc.gov/catdir/>
11. D.Vizine-Goetz. Subject headings for everyone: Popular Library of Congress Subject Headings with Dewey numbers. *OCLC Newsletter*, May/June, 1998. [http://cweb.oclc.org/oclc/new/n233/rsch\\_subj\\_headings\\_everyone.htm](http://cweb.oclc.org/oclc/new/n233/rsch_subj_headings_everyone.htm)
12. K.M. Fairchild, S.E. Poltrock, & G.W. Furnas. SemNet: Three-dimensional representations of large knowledge bases. In: R.Guidon (ed.) *Cognitive Science and its Applications for Human Computer Interaction*. Erlbaum, 1988. 201-233. Reprinted in S.K. Card, J.D. Mackinlay & B. Shneiderman (Eds.) *Readings in Information Visualization*. San Francisco: Morgan Kaufmann, 1999.
13. P. Pirolli, P. Schank, M. Hearst & C.Diehl. Scatter/Gather browsing communicates the topic structure of very large text collections. *Human Factors in Computing Systems (CHI)*, Vancouver, April 13-18, 1996. *Proceedings*. New York: ACM, 1996. 213-220.