The Sunflower Visual Metaphor, A New Paradigm for Dimensional Compression

Stuart J. Rose*
Department of Management Information Systems
The University of Arizona

ABSTRACT
This paper introduces the Sunflower visual metaphor for information visualization. The visual metaphor is presented as an alternative to current techniques of dimensional compression and the visualization tools that employ them. The paper discusses the motivation for the Sunflower paradigm, its implementation and critical factors for producing an effective visualization. A primary driver in this research effort has been to develop a visualization tool that facilitates browsing, knowledge discovery, and that supports learning through sense making and integration of new information.

KEYWORDS: information visualization, text visualization, visualization, knowledge management, information retrieval

1. INTRODUCTION
A common problem with extensive collections of information is locating information relevant to the user's interests. Attempts to fulfill the search for a relevant subset of information consist primarily of searching for documents that are similar to an 'ideal' type based on an individual's query. These search results are often displayed as a ranked list, with the items most similar to the query located at the top of the list.

The ranked list is an effective visual format for the individual actively seeking a specific item from a larger collection. However, the ranked list does not promote effective exploration of unfamiliar knowledge domains as it does not facilitate discovery of inter-relationships among elements in the new collection or show how unfamiliar knowledge domains interrelate with familiar ones. Exploring new domains is hampered by the difficulty of formulating a query that acts as a proper guide for conducting the search [9]. This query must be specific enough to return useful information yet broad enough to avoid restricting the user's acquisition of related, and potentially useful, material.

In order to support browsing and exploration of large collections of information, the Sunflower visual metaphor displays each information elements' membership in the knowledge domains of a collection. When applied to a collection of elements, the Sunflower visual metaphor directly conveys the degree of specificity of each information element and shows inter-relationships among domains in the collection.

2. LITERATURE REVIEW
"A primary role of visualization is to shift the activity of information processing from the lexical to the spatial realm in order to enable users to make full use of their innate capabilities to acquire information more efficiently" [6].

Information visualization displays abstract relationships among information elements. Visually representing inter-relationships among the elements in a collection aids exploration of that collection by a user [6]. This is based on the belief that cognitive operations are the essential ingredients of perception [2]. Each element is evaluated across one or more attributes in order to show interrelations among elements in the collection.

*stuart.rose@acm.org

2.1 Continuous Measures
Continuous measures are often used to indicate the degree to which an information element possesses an attribute [11]. Using continuous measures requires that each attribute have a dimensional space in which the element is represented. An element is placed within this dimensional space (dimension or axis) according to its measure for that attribute. Differences among attribute measures for elements are considered to reflect relationships between those elements. The difference among elements' attribute measures is often referred to as distance between elements when comparing across more than one attribute. Elements with similar measures across attributes will consequently be separated by less distance in the multi-dimensional space, appearing in closer proximity to one another [4]. The implication that proximity is associated with similarity is exemplified in visualization tools such as Self Organizing Maps (SOMs), those that employ Multi-Dimensional Scaling (MDS) such as SPIRE, and in social gatherings among people and animals.

Evaluation of elements across additional attributes creates more information for comparison, but requires more dimensions than can reasonably be represented in an informative display. Representation of each detail of elements' inter-relationships is restricted by the ability to represent each of the dimensions for comparison, in a 2-D display [8].

Any technique that attempts to compress a higher dimensional information space into a lower dimensional space must select appropriate details or representation schemes in order to provide an effective visualization of the multi-dimensional artifact. Capable artists, when translating the physical location in space of a 3-D object to a 2-D representation, integrate multiple views into a single 2-D display. The success of an information display is dependent on the choice of details from the available collection of views [10]. Several information visualization tools that support this activity through application of MDS and SOMs include the Cosmic Tumbleweed, SPIRE, DEPICT, as well as OOHAY from the University of Arizona [4, 7]. In addition Parallel coordinates allows the representation of many dimensions in a display but is often challenged by the number of elements it can represent in a single display.

2.2 Discrete Measures
When attributes are not measured on a continuous scale, information elements can be evaluated with a discrete measure for each attribute, i.e., elements are considered to either possess the attribute or not. Previously, when plotting a continuous measure, we needed an axis, or dimension, in which to plot values for a single attribute. When using discrete measures, each attribute can be represented as a domain in a dimensional space. Any element located within an attribute's domain space is understood to possess that attribute. When the domain spaces are mutually exclusive, elements are limited to representing one attribute for each dimension in the display. In the case of a 2-D display, an element located at the intersection of two lines is understood to possess the attributes from which those lines originate [3].

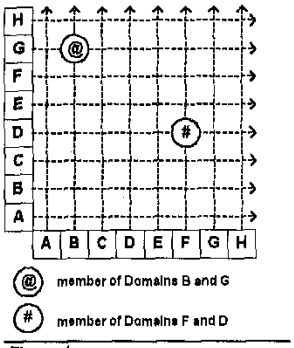

Figure 1

In Figure 1, each attribute along the axis has a corresponding line extending perpendicularly from the axis at its location. An element located along this line is understood to possess that attribute. Adding a second axis places the element in a 2D space, allowing a point to represent any two attributes that the element possesses. There is no effective limit to the number of attributes that can be placed along an axis. However, each element can only be represented as possessing one attribute for each dimension in the display. A feasible limit on the number of dimensions that this type of display can support is three dimensions. The intent of the Sunflower visual metaphor is to clearly represent elements as belonging to more than two or three domains within a 2-D display.

3. RESEARCH QUESTIONS

Consider that a single axis is an order of attributes in one dimension. A line drawn from one of these attributes represents the domain space (domain) of that attribute. The selection of parallel lines as domain shapes results in domains that do not intersect with other domains until a second axis, or dimension, is brought into the display (see figure 1). The Sunflower changes domain shapes from parallel, 1-D lines to consist of higher dimensional shapes that intersect with one another (see figure 2). By displaying a structure of intersecting domains, it is possible to represent many attributes for each element within the collection. We are no longer limited to two or three attributes per element.

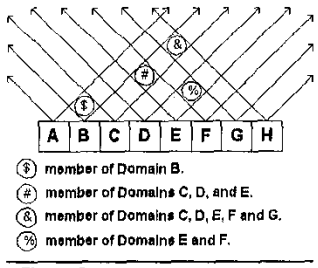

Figure 2

The Sunflower represents multiple attributes on a 2D surface by displaying each attribute as a bounded domain shape. Elements that possess an attribute are considered members of that attribute's domain. Likewise, an element that is a member of domain A is understood to possess attribute A. The task of representing multiple attributes in one element is accomplished by placing the element at an intersection of domains of which the element is a member. Given any number of attributes in an element, we can represent the element at an intersection of those attributes' domain shapes.

Representing the location of one element at an intersection of domains is a simple matter. However in order to support browsing a significant number of elements must be represented in the display. Adding more elements to the visualization may cause a loss of precision in the placement of elements. A measure of precision can be computed for each element displayed in the structure, and is defined as

$$\text{placement precision} = \frac{\text{\# of domains element belongs to}}{\text{\# of domains represented in segment}}$$

The opportunity for loss of precision in the location of an element is a tradeoff with the fact that there is no longer a limitation of representing only two or three attributes per element.

Domain ordering is critical to an effective visualization with the Sunflower visual metaphor. When a compromise is made in the order of domains, there is a greater likelihood that an element will have a reduction in the precision of its placement. Consider an example in which two domains have a high degree of co-occurrence. Placing those two domains in a position where they are not adjacent results in a loss of placement precision for those elements that are members of only those two domains. Since we are primarily interested in displaying large collections of documents in the Sunflower, managing this loss of placement precision is critical for an effective display. In the following section, we discuss the Sunflower's method of construction and methods for maximizing precision.

4. SYSTEM DESIGN

The Sunflower Visualization Tool (SVT) implements the Sunflower visual metaphor by acquiring a list of domains and determining their pattern of occurrence within a collection of information elements. For each element, the SVT determines the domains of which the element is a member and stores the information in an attribute array for the element. From the information of each element's attributes, a co-occurrence matrix is built which is used as the basis for determining an effective order for the display.

The co-occurrence matrix stores measures of co-occurrence between domains. The degree of co-occurrence between two domains is a count of the number of elements that share membership with those two domains. Any algorithm that is considering each of the domains uses the co-occurrence matrix as a source of information on relationships between domains.

4.1 Algorithm

The dynamics of ordering the domains can be understood by considering the analogy of seating people at a round dinner table. Consider that each person at the dinner table wants to sit closest to the people with which they have the most in common. In order to maximize social harmony, seating at the table should be ordered so that people are next to others with which they have the most in common. Arranging the order according to the preferences of one person is straightforward but will result in a compromise for others at the table. The seating arrangement should be ordered so that the need for compromise by each person is minimized.

In similar fashion, ordering domains in the Sunflower should place frequently co-occurring domains close to one another. The priority of domain ordering is to place domains adjacent to one another in the domain order based on their relative degree of co-occurrence. The method of evaluating domain orders consists of calculating the burden for each domain within an order. We can calculate a measure of burden to evaluate the appropriateness of each domain's location in the domain order and develop a measure of total burden for any domain order. The total order burden is the sum of each domain's burden in the order. Each domain's burden is calculated by summing the burden between itself and the other domains in the order. The burden between domains is defined as the degree of co-occurrence between two domains multiplied by the distance separating those two domains in the order.

A reduction in burden requires that each domain be in closest proximity to the domains with which it co-occurs most frequently. The problem of establishing an effective domain order is

summarized as a circular 1-D order whose solution is NP-complete [1].

By placing domains in close proximity, we increase the area of intersection (number of segments) between those domains (see figure 2). The greater the number of segments that two domains share, the more opportunity there is to increase the precision of element placement for the elements that are members of those domains.

4.1.1 Close Friends

The Close Friends algorithm begins by locating the domain that has the highest membership from the co-occurrence matrix, the dominant domain. Once the dominant domain is selected, the two domains that most frequently co-occur with the dominant domain are placed in the empty seats flanking the dominant domain. The process is repeated by finding a domain from those remaining that most strongly co-occurs with the domain at the end of the order. Because the order is circular, the algorithm switches from side to side as it fills in locations around the table. An option with this algorithm is to start with a user-selected domain rather than the dominant domain. This gives priority to the user's interest while still providing scope for that domain within the collection and its inter-relationships with other domains.

4.1.2 Equal Burden

In developing an order, the Equal Burden algorithm considers the relationships across the table as well as the relationships on the outside edge. The current implementation employs a genetic algorithm to evaluate and modify a collection of potential domain orders. An option with the equal burden algorithm is to scale the weights in the co-occurrence matrix to one so that each domain has an equal weight in the calculation of burden.

4.2 Implementation

The Sunflower Visualization Tool implements the Sunflower visual metaphor and offers users a choice of using the Close Friends or Equal Burden algorithms. In the current implementation, a list of recommended domains is presented to the user for possible editing. The domains from this list are then used to compare elements in the information collection in order to build the co-occurrence matrix. After the domain order is established, domains are placed in the display based on their location in the domain order.

Domains are represented as similar shapes with their centers placed equidistantly around a central point (see figure 3). In the current implementation of the SVT, circles are used as the domain shapes, which results in a visualization resembling the head of a sunflower. The use of circles results in mutual exclusion of domains directly opposite one another in the circular domain order. The implication for the placement of elements is that none of the segments contains more than half of the total number of domains visible in the display. Other shapes may be selected for domain representation which do not have this mutually exclusive property in the display and in that case, an element can be represented at an intersection of all of the domains in the display.

After the domains are represented and labeled in the display each element is placed in a segment of intersection that best corresponds to the element's attributes. The element's domain array is used to locate the best-fit segment for the element.

Each element is placed in a segment that contains all of the domains that the element contains. The greater the range of domains, the closer to the center of the structure the element will be placed (see figure 3). Since an element may be a member of two domains remotely located in the domain order, the element may be located in a segment at which more than two domains intersect. The best-fit location for an element is the segment that maximizes placement precision.

Each element is represented in the display as a star glyph in which line segments are drawn at angles corresponding to the domains of which the element is a member. The order of domains for each star glyph matches the order of domains for the sunflower. The star glyph is useful as a visual measure of placement precision for each element and allows comparison of elements on a micro level.

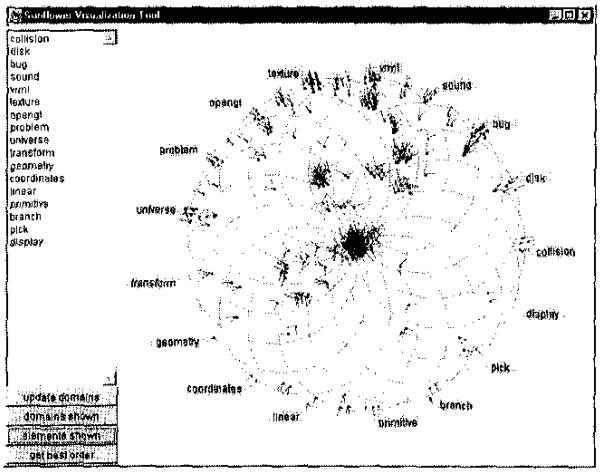

Figure 3

5. AN EXAMPLE

Consider that a document is a collection of themes, topics and terms that can be summarized as attributes of the document. The document can then be described in terms of the attributes that it contains. By visually representing attributes of a document, an information display is created in which a user can grasp those attributes of the document by simply observing the components of the information display [4, 7].

The SVT has currently been employed to explore a collection of 500 email messages received from a list-serve covering the topic of the Java3D API. This collection has the advantage of a specific ontology (the keywords, methods, and classes occurring in the API) that can be used to access documents (email messages) covering aspects of Java3D. Having a specific ontology that is adhered to by each information element aids the appropriate placement of elements in the visualization [3].

Figure 3 displays output of the SVT analyzing a collection of 500 Java3D email messages. The terms used to evaluate elements and create domains are located in the text area. When the optimal order is calculated, the terms are placed back in the text area in that order. Each document is represented as an icon plot at an appropriate intersection of domains based on the content of that document. By looking at the icon plots for each individual element, we can visually determine the placement precision for that element and assess the effectiveness of the visualization.

At the macro level, we can see that the terms 'bounding', 'compiler', 'collision', and 'disk' occur in only a few documents. In contrast, the terms 'geometry', 'transform', 'universe', 'problem' and 'opengl' are well represented in the collection of documents. Based

on their proximity in the ordering, it is understood that they co-occur frequently with each other. This can be explored further on a micro level by investigating each of the icon plots within the domains to evaluate which terms the document contains. Each document is appropriately represented as an individual element based on its attributes but appearing among neighbors by virtue of similarity in content.

The metaphor provides an interface that assists users in exploring domains correlated with their initial domain of interest. In this case, a user exploring documents about 'texture' will see that 'opengl' and 'vrml' occur in the same documents as 'texture'. An interesting aspect is that the icon plots reinforce the notion of "families" of documents in the collection. Users can easily locate those documents that contain the grouping of terms that fit the user's interests. Direct exploration is promoted through manipulation of the selected domains. Changing the domains that are presented in the display results in a new visualization.

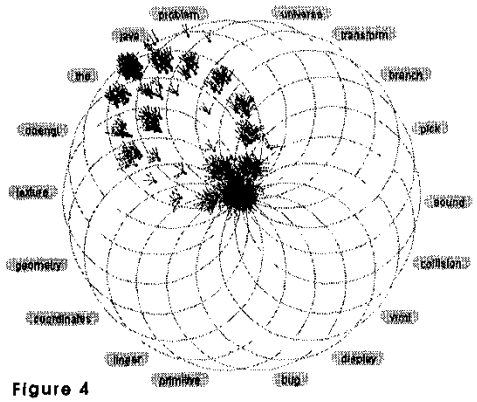

Figure 4

Topics that are dominant in the collection have a tendency to pull all of the messages to its domain, eliminating the usefulness of much of the display (see figure 4). An important point to follow in using the display is to therefore balance the keywords by avoiding words that will appear in every message on Java3D such as 'java' or common stop-words such as 'and' and 'the'.

6. CONCLUSION, FUTURE RESEARCH

Specificity of each information element is conveyed through the location of the document in the Sunflower. The consequence for exploration is that a user familiar with one knowledge domain can enter new domains through elements that belong to both. A major advantage of constructing the visual metaphor with a single perimeter is that one layer of segments (in this case the outer layer) contains elements that belong to only one of the domains displayed. The second layer of segments contains two domains and so on until the center most layer of segments contains those elements that are the most broad in nature.

A user can select which elements to access based on their interest in specific elements or more general elements. A user interested in exploring an unfamiliar domain can look for elements within that domain that are also located in familiar domains. This allows the user to begin exploration of the new domain within a familiar context. In this way, a user can build knowledge of the entire collection by moving from domain to domain, selecting the level of specificity that is desired based on the attributes of the collection.

A critical aspect of producing an effective visualization is the selection of relevant domains from the collection. A domain can represent any attribute. Those attributes that are useful in comparison of elements and which provide useful information about the element are those that should be selected for the visualization. Establishing the domains to include in the Sunflower structure can be accomplished through two methods, gathering user-entered keywords, and using automatic techniques to determine and provide relevant domains.

The Sunflower allows users to be placed in a context appropriate to their search interest, preventing the pervasive disappointment and frustration that accompanies a search through large information collections [5]. Anecdotal evidence suggests that changing the list of domains to create a new display is an interesting, interactive process that encourages exploration but precision can be low as the relevance of some user provided keywords to the collection is questionable.

The Sunflower is intended to disambiguate representation of information elements in a display by representing each one within bounded domain spaces. The critical factors in creating a useful visualization rest in the ability to generate relevant domains from the collection. The use of a thesaurus is useful in the creation of term-groupings when an ontology is not adhered to among all of the elements in a collection.

Further research is ongoing to produce "meta-terms" that represent a group of synonyms occurring in the collection and occupy a single domain in the visualization. Successful implementation would allow the application of the Sunflower visual metaphor to visualize document summarization.

8. REFERENCES

[1] Michael Ankerst, Stefan Berchtold, Daniel A. Keim. Similarity Clustering of Dimensions for an Enhanced Visualization of Multidimensional Data. In Proceedings of IEEE Symposium on Information Visualization, InfoVis '98. 1998. IEEE CS

[2] Rudolf Arnheim. Visual Thinking, University of California Press, 1969

[3] Beth Hetzler. Nancy Miller. Four Critical Elements for Designing Information Exploration Systems. http://www.pnl.gov/infoviz

[4] Beth Hetzler, Paul Whitney, Lou Martucci, Jim Thomas. Multi-faceted Insight through Interoperable Visual Information Analysis Paradigms. In Proceedings of IEEE Symposium on Information Visualization, InfoVis '98. 1998. IEEE CS

[5] Michael McQuaid, Thian Huat Ong, HsinChun Chen, Jay F. Nunamaker, Jr. Multidimensional Scaling for Group Memory Visualization. 1998

[6] Russell Rose. P1000 Planning Committee. Science and Technology Strategy for Information Visualization. Sept. 1996

[7] David Rushall, Marc Ilgen. DEPICT: Documents Evaluated as Pictures, Visualizing Information Using Context Vectors and Self-Organizing Maps. In Proceedings of IEEE Symposium on Information Visualization, InfoVis '96. 1996 IEEE CS

[8] G.A.F. Seber. Multivariate Observations. John Wiley & Sons, 1984

[9] Min Song. BiblioMapper: A Cluster-based Information Visualization Technique. In Proceedings of IEEE Symposium on Information Visualization, InfoVis '98. 1998. IEEE CS

[10] Edward Tufte. Envisioning Information. Graphics Press, 1990

[11] Edward Tufte. The Visual Display of Quantitative Information. Graphics Press, 1983