The Knowledge Space Visualizer: A Tool for Visualizing Online Discourse

Chris Teplovs
Institute for Knowledge Innovation and Technology
University of Toronto
chris.teplovs@utoronto.ca

Abstract
In this paper I describe a new visualization tool for data from CSCL systems. The Knowledge Space Visualizer (KSV) projects writings ("notes") from participants into a two-dimensional space and shows relationships and interactions between notes as a graph. These relationships can be structural (e.g. relationships such as "reply-to", "build-on", "reference", "annotation", "contains"), authorial, semantic, or derived from codings applied by researchers. In the case of semantic links, similarity is determined by using Latent Semantic Analysis. The KSV employs a force directed layout algorithm to minimize the distortion of the relationships as viewed in two dimensions. In its current form the KSV is tuned to import data from Knowledge Forum, although any data set comprising information about content, creation time, and authorship could be visualized. Results of visualizations can be fed back into Knowledge Forum to promote further building on of analyses by other researchers. This affordance leads to community-based "rise-above" analyses that foster a cycle of continual analytic improvement.

Introduction
The amount of discourse generated online is growing at a staggering rate. Wikis, blogs, bulletin board systems and knowledge building environments provide exciting opportunities for participants and researchers in the Learning Sciences alike. But how can participants and researchers make sense of the discourse and the data? Traditional qualitative, quantitative and statistical analyses are often either impracticable or fail to capture the richness of the data.

Visualization of data from CSCL settings can provide insights that would be lost using summarization statistics while at the same time providing sufficiently rich details that users can come to understand the nature of the discourse and the interactions that underpin its production.

The theme of this workshop is interaction analysis, which I take to mean interaction between people. My focus is on the interaction between ideas, with interactions between authors being of secondary import. Of course, ideas are expressed by people so the shift in focus is a subtle one. To this end I use Latent Semantic Analysis (LSA) to determine semantic similarity between pieces of writing. LSA is a statistical technique used to extract the deep meaning of patterns of words in specific contexts of use. The technique is performed by applying methods from
linear algebra (matrix decomposition and dimension reduction) to matrices that represent usage patterns of terms in documents (Deerwester et al. 1990, Landauer et al. 1998). It is also a theory about knowledge acquisition in human beings (Kintsch, 1998; Landauer & Dumais, 1997).

Landauer et al. (2004) suggest visualization of LSA-derived similarities is problematic due to unacceptable reduction of dimensionality to two or three dimensions suitable for visualization from that which is optimal for LSA (typically around 300). The technique used here differs from the sort of dimension reduction or selection examined by Landauer et al. (2004) in that document-to-document similarities are used to construct an undirected graph. Edge lengths are scaled according to semantic similarity. To simplify the display edges that represent document similarities that fall below a flexible threshold are not rendered. The resultant semantic clusters seem to map onto intuitive notions of thematic clusters, although face validity of these clusters has not been rigorously studied. The focus of this paper is the visualization tool, which takes into consideration semantic relationships, and not the semantic analysis itself.

**Escaping Flatland: Visualizing Multivariate Data**

CSCL systems provide us with lots of data but we tend to examine only select parts of it, focusing only on those features that we pre-determine to be of interest. I contend that it is better to try to consider more dimensions. Tufte's (1990) chapter titled "Escaping Flatland" calls for increasing the data density in visual displays.

Knowledge Forum is a knowledge building environment with a 25-year development history (Scardamalia & Bereiter, 2003). It possesses an inherent two-dimensional visualization: contributions (notes) are positioned in two-dimensional spaces (views), often according to collaboratively negotiated conventions. In some cases the horizontal and vertical axes have interpretable meaning. In other cases the 2D space is used is a less methodical way to group/structure notes (Figure 1). Early attempts to make Knowledge Forum web-accessible saw a concomitant loss of data and functionality. Even in its current form the nature of the data that is encoded onto the 2D view is limited. Note status (read, unread, modified-since-read) is easily gleaned, but other information (e.g. authorship) is present but less obvious.
Figure 1: Typical view from Knowledge Forum

To address these shortcomings a more powerful visualization system has been created. The design of the Knowledge Space Visualizer (KSV) follows Shneiderman's threefold mantra for information visualization: (1) provide an overview, (2) provide the ability for the user to zoom and filter, and (3) provide details on demand. A screen capture of the KSV in action is shown in Figure 2. Of note are the major components: (a) the main display area, (b) a display area for showing semantic fields (c) an overview display, which is always zoomed out to include all visible entities, (d) controls for filtering according to creation dates, title, author, or note content, (e) controls for showing and hiding various object types and the links between them, (f) controls for scaling and shading objects, (g) controls for general display options and database functionality, (h) a panel for displaying details about individual objects or analyses, and (i) information about data sources and extraction details.
The main display area supports zoom and pan operations, both manual and computer-supported (e.g. "zoom to fit"). Any number of objects can be selected with a marquee-style selector for subsequent manipulation or analysis. "Tool-tip" style windows that contain information about authorship and a summary of content appear when a user hovers over an individual node. Links between nodes are shown using cubic spline curves to minimize problems with occlusion when nodes are multiply linked. Directionality of links is indicated by arrowheads.

Notes can be selected and then moved in the space. Alternatively, the user can request the generation of a semantic field (described below) and/or the execution of a simple analysis that shows some surface features of the writing (e.g. word count).

The selection of objects can be constrained to include only those created or modified within a specified date range. The date range specification is controlled by a slider that operates in such a way that the patterns of contribution over time can be seen. By manipulating both the minimum and maximum date values users can create virtual time slices of the discourse space. Objects matching search criteria can similarly be highlighted using text-based text searching facilities. The display of objects and links between them is configurable by the user.

Using the Knowledge Space Visualizer

The KSV creates internal tabular representations of the data using the prefuse toolkit (Heer et al., 2005). The first step in using the KSV is to import the data from whatever external system one is interested in. Any well-formed XML can be translated into a form suitable for importation into the KSV. The goal in importing data is to reproduce with as much fidelity of the original information as possible.
Figure 3 shows representations of Knowledge Forum data in the client (a), visualizer (b) and HTML (WebKF) representations (c). Note the loss of data in the translation to the threaded-discourse form.
To start with, structural links (edges) between notes (nodes) are shown. Build-ons ("replies" in other systems), references (track-backs or ping-backs in other systems), and annotations are distinguished by colour.

In keeping with the idea of adding (rather than subtracting) data from visual displays, the visualization tool allows the user to add information about semantic similarities between notes. (Figure 4). The user can vary the threshold value at which two notes are considered similar enough to justify the presence of a visual edge. By adding information about semantic similarity a richer description of the nature of the note content emerges. Note that semantic linkages are bidirectionally equivalent and therefore lack arrowheads.

![Figure 4: Semantic links (red) between notes](image)

With the addition of semantic linkages the display becomes considerably more complex, especially when the similarity threshold is lowered to create more inclusive semantic clusters. The user can manipulate the positions of nodes to create more clarity but this strategy becomes impractical when dealing with large numbers of nodes or in dense graphs. In these situations it is helpful to employ computer-assisted layout algorithms such as force directed layout (FDL). When activated, the FDL brings together those items that are linked by visible edges, and repels those that lack such connections. An example is given in Figure 5.
Figure 5: Semantic clusters after applying force directed layout

Force directed layout can also be used in the absence of information about semantic similarities to reveal interesting observations based on structural data. For example, invoking FDL on a view showing notes and the build-on and referencing relationships can yield easily interpretable information about the nature of build-on structures. Consider the differences between the two images in Figure 6. In Figure 6a, we see circular build-on structures; in Figure 6b we see mostly linear ones. In the former case there tends to be one note with many build-ons whereas in the latter there tends to be build-ons to build-ons.

Figure 6: Many build-ons to one note vs. building on to build-ons
Objects other than notes can be visualized as nodes as well. Authors, views, supports, keywords and problem statements can all be used to generate graphs using FDL. Knowledge Forum allows researchers to add codes to notes according to any arbitrary classification scheme. These can be for schemes as simple as author gender or as complex as assessment of explanatory power of theory statements. These objects can then be used to cluster notes in the visualizer (Figure 7).

![Figure 7: Clustering by author. Authors are represented by magenta triangles and appear as centroids of note clusters.](image)

In general, objects in the visualizer can be clustered by some criterion: semantic relatedness, build-on or reference structure, authorship, or any other coding scheme. Objects can also be shaded or coloured according to other criteria: authorship and researcher coding are currently implemented. The process of visual analysis in the KSV involves layering on new dimensions of data by employing linking, clustering, positioning, and shading. As a specific example, one can colour by author and then cluster by semantic relationships to see the degree to which various authors share semantic content.

It's also possible to use various variables to determine layouts. For example, one can order notes chronologically across the X-axis, or by author vertically, or both simultaneously.
Multiple Data Sources

The KSV can be used as a tool to investigate curricular coverage by examining the semantic relationships between curricular guidelines and contributions from participants. This has been done by injecting government-mandated curricular guidelines into the same database as the students work in, and then using the visualizer to show relationships between the notes. The progression of semantic overlap with curricular guidelines in a class of students ranging in age from 10 to 12 is shown in Figure 8. It is possible to see, at a glance, increasing coverage of curricular concepts (shown as nodes near the top of each display). Red lines connecting student contributions to curricular guideline are indicative of coverage.
Figure 8: Semantic overlap between curricular guidelines (dark icons across top of each image) and student contributions in early, middle, and late parts of unit

Semantic Field Visualization
The semantic fields for any collection of notes can be generated. This is currently done by one of three algorithms: author-chosen keywords, LSA-derived keywords, or keywords derived through a term extraction service such as that provided by Yahoo!. The user of the KSV can select one or more notes and request the creation of a semantic field. The process can be repeated with other selections and the resultant overlap can be examined for interesting features. Figure 9 shows an example of two overlapping semantic fields.

Figure 9: Semantic overlap of two sets of notes. The terms "organs", "genes", "cells", "plants", "organisms" and "dna" are shared by the two sets

Semi-Automated Visualization of Inquiry Threads
Finally, inquiry threads analysis (Zhang et al., 2007) can be done more easily by using the Knowledge Space Visualizer according to the following steps.

1. Use semantic similarity to show note linked by content
2. Use force directed layout to create semantic clusters
3. Flatten each cluster horizontally
4. Arrange each flattened cluster chronologically
5. Colour by author if desired
6. Hide semantic links and show structural links
Cybernetic Analysis

Analyses of CSCL systems tend to be done once and not generally repeated but for meta-analyses from published summary statistics. One of the shortcomings of visual analyses is that they tend to be difficult to reproduce. Screen captures are an effective way to convey the resultant images, but the state of the system is not typically saved. The KSV works in concert with Knowledge Forum (assuming the data originated from Knowledge Forum) by allowing the state of a visualization to be saved to the database. This process allows researchers to share their visual analyses with each other, which in turn should promote collaborative analysis of data and more powerful interpretations of the data.

References


