Visualising Semantic Pathways in Document Collections

C. Rowland[1] and J. Anderson[1]

[1]Duncan of Jordanstone College of Art & Design, University of Dundee

Abstract

In this paper we discuss our investigation of the potential uses of animation and visualisation techniques for exploring unstructured text-based data. We reflect upon a design methodology that is rooted in art and design practice and introduce a conceptual model for following "semantic pathways" through document collections. We also describe an implementation of a working model of an interactive visualisation. We see potential applications for our conceptual model in domains where there is a requirement to infer narrative from multiple sources of evidence such as in counter-terrorism and criminal investigations.

Categories and Subject Descriptors (according to ACM CCS): I.3.8 [Computer Graphics]: Applications—

1. Introduction

In this paper we discuss our initial investigation into the potential use of animation and visualisation techniques within the field of visual analytics.

Our research, which is work in progress and currently at an early stage, is concerned with real time visualisation and exploration of unstructured text-based document collections using a conceptual model which makes use of "semantic pathways". Although our conceptual model and visualisation techniques are currently only applied to a relatively small collection of plain text documents, we are working towards investigating their potential in much larger multimodal document collections.

2. Design Methodology

We are aware of visual analytics systems such as JIGSAW [SGL08] and IN-SPIRE [WJ04]. Our novel approach, grounded in design research methodologies, may offer a different perspective. Art and design is by nature concerned with hard to quantify factors such as aesthetics and intuitiveness, but may bring tangible benefits with regard to the usability of visual analytics software.

Our design language comprises minimal graphics, subtle animation/compositing and typography. Our design philosophy is about working within self imposed constraints and reducing the amount of interface between the user and the data.

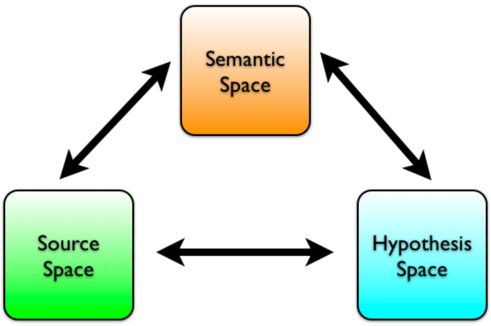

Figure 1: *Conceptual model for a data space.*

3. Conceptual Model

The conceptual model for our design is based around a flexible "data space" (Figure 1). This space functions both as a canvas for displaying the visualisation and as a workspace for interacting with data. Graphically this is similar to examples of other two dimensional representations of semantic spaces such as can be found in Widdows et al. [WCD02]

Within the data space there are three conceptual sub-spaces referred to as "semantic space", "source space" and "hypothesis space". Semantic space is concerned with query formulation and execution, source space with providing mediated access to the source data, and hypothesis space with inferring narrative and sequencing evidence. Our current focus is on the interplay between the semantic and source spaces.

4. Data Preprocessing

We selected the VAST 2010 Challenge (Mini Challenge 1) data set which we preprocessed in two stages. First we separated the composite Word documents into individual plain text documents and classified these by document type (e.g. email intercept, intelligence report, etc.). This yielded 103 documents in 6 classifications. Secondly we devised a simple extraction algorithm to produce an XML metadata file corresponding to each text file. The XML metadata encapsulated the ten most frequently occurring words in each plain text file (excluding English stop words) as well as for the entire collection.

5. Implementing an Interactive Visualisation

In order to implement our interactive visualisation we required a prototyping environment which would allow us to utilise the data in the context of real time interaction and blend our design methodology with our conceptual model. We considered several candidate tools including Quartz Composer, Processing, Field and vvvv. We felt that Apple's Quartz Composer environment offered the most appropriate combination of design, animation, compositing and programming capabilities for our immediate needs.
Our working model (WM) currently implements only the semantic and source spaces from our conceptual model.

In its default state the WM presents in semantic space a menu of search terms, corresponding to the ten most frequently occurring words in the data set. The "default query" is simply the Collection itself. In the source space the entire document collection is presented as a row of icons grouped and labelled by document classification.
The user initiates a new query by selecting a search term from the display. Using animation, the previous query recedes into a background layer and the new current query appears in the foreground. Query results, in the form of a document subset, are returned in the source space.

The new current query which appears in the foreground of the semantic space is presented along with a new display of search terms harvested from the document subset that was returned in the source space. These harvested search terms comprise the most frequently occurring words taken from each returned document in turn (one from the first, one from the second and so on). The user may then initiate a new query by selecting one of these new search terms. The user may also expand the available search terms by adding, from each document in turn, additional frequently occurring words from the second most frequent to the tenth most frequent. We call this expansion of the search terms "fanning" (Figure 2) as they take on the appearance of a fan of keywords emanating from the current query term.

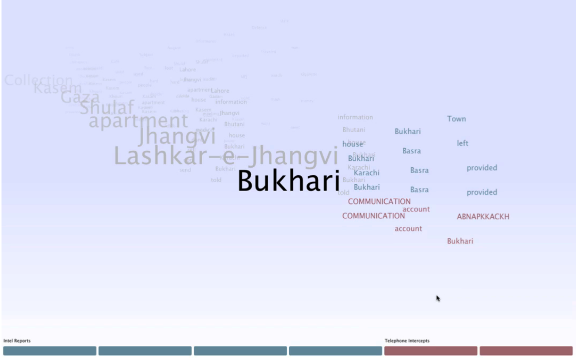

Figure 2: *A semantic pathway with "fanning" of current query.*

The pattern of querying with the returned search terms proceeds recursively as the user follows the chosen semantic pathway (Figure 2). The subtle animation and layer based compositing is used to preserve the semantic pathway in a visual history.

6. Conclusions and Further Investigation

Our research is at an early stage and clearly further investigation will yield more concrete results. Our work in progress demonstrates an interesting conceptual model and a design methodology which we have tested in a working model implementation.
In the immediate future we wish to further investigate the role and nature of the hypothesis space. We see in the longer term an obvious need to devise qualitative and quantitative studies to test the efficacy of our conceptual model. We also hope to investigate whether the model may be applied to established formal data structures, ontologies and knowledge bases. Ultimately we see potential applications in domains such as counter terrorism and criminal investigation.

7. References

References

[SGL08] Stasko J., Görg C., Liu Z.: Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization 7*, 2 (2008), 118–132. 1

[WCD02] Widdows D., Cederberg S., Dorow B.: Visualisation techniques for analysing meaning. In *Fifth International Conference on Text, Speech and Dialogue (TSD 5)* (September 2002), pp. 107–115. 1

[WJ04] Wong P. C., J.Thomas: Visual analytics. *IEEE Computer Graphics and Applications 24*, 5 (Sept-Oct 2004), 20–21. 1