Beth Hetzler, W. Michelle Harris, Susan Havre, Paul Whitney
Pacific Northwest National Laboratory

Visualizing the Full Spectrum of Document Relationships

Abstract: Documents embody a rich and potentially very useful set of complex interrelationships, both among the documents themselves and among the terms they contain. However, the very richness of these relationships and the variety of potential applications make it difficult to present them in a usable form. This paper describes an approach that enables the user to visualize a multitude of document or entity relationships. Two visual metaphors are presented that allow the user to gain new insights and understandings by interactively exploring these relationship patterns at multiple levels of detail.

Introduction

Traditional information retrieval systems focus on a particular relationship among documents: similarity of content. The goal in such systems is to find documents whose topical content matches the topical content of the user query. Likewise, information visualization systems often focus on a single or perhaps a small number of relationship types. Some focus on matching content; others focus on particular attribute values, such as author or publication date; still others focus on explicit links among documents, such as citations or hyperlinks. This paper describes an approach that not only enables the user to visualize a multitude of document relationships but also provides exploration capabilities that allow the user to gain new insights and understandings of the relationships that might not otherwise be apparent.

The Importance and Richness of Relationships

Why are document relationships important? Traditionally, library card catalogs list books and documents in multiple ways, such as by author and by subject. These simple relationships have proved very useful for many users for many years. Research by Fidel and Crandall showed that users often make decisions about which documents to read based on more complex attributes and relationships, such as whether the document included hard data, was written on a technical vs. non-technical level, or confirmed what the user already knew (Fidel, 1997). With the huge volume of information available today, tools that help users select which few documents to spend time reading can be invaluable. For example, a system that could help users understand the history and parentage relationships among documents, the citation relationships, and attribute similarities would help users select documents and possibly help them gain insights about the group of documents as a whole.

Documents, especially collections of documents, are rich and complex in the relationships they represent. At first glance, the fact that one document cites another might be considered a simple relationship between the two, easily conveyed and used. However,

document citation studies propose numerous possible relationships that may motivate such a citation (Liu, 1993). A study of bibliographic relationships represented in library cataloging systems led to a taxonomy with seven major categories (Tillett, 1991). In a workshop held at SIGIR '97, participants identified a large number of potential document-to-document relationships, which were grouped into eight different relationship categories (Hetzler, 1997). Examples included a number of non-traditional documents and relationships, such as documents supporting the same workflow process, images of adjacent objects, and Internet newsgroup question/answer messages.

Further, relationships among individual terms or phrases can be extremely useful as well as complex. The need for exploring and identifying relationships among terms in a thesaurus is described by Green (1996) and Molholt (1996). One difficulty is that although much work is being done on standard ontologies, the use of the words within the ontology still may vary from person to person and group to group. A system that detects relationships based on actual usage may help solve this difficulty. One approach to this task is described by Cooper and Byrd (1997). Rich relational structures have been developed based on analyses of English verbs (Green, 1996; Myaeng, 1994). Words and phrases are commonly used for document retrieval; however, the inadequacy of current methods that simply match topics between a query and the potential target documents has also been shown (Green and Bean, 1995). In the military intelligence community, relationships among terms (e.g., named entities and concepts) can be particularly valuable in identifying key information. In this domain, it may be very helpful to simply identify the fact that a relationship appears to exist between certain terms or entities within a collection of documents, even if the type of relationship is not known.

Visualization as a Tool for Understanding Relationships

Given that documents and terms can embody rich and potentially very useful relationships, one challenge is how to represent these relationships so that users can easily understand and use them. The primary goal of information visualization is to help users glean insight from large collections of information. Several tools have been developed that graphically represent document relationships. A common representation is some form of link-node diagram that depicts one or at most two types of explicit relationships. Examples include systems that show call dependencies in computer code (Storey, 1997) and systems that show visualizations of World Wide Web link structures (Card, 1996; Munzner, 1997). A two-dimensional matrix approach to showing relationships has also been applied by Becker (1995) to portray telephone network overload among major cities and by Gershon (1995) to portray how words appear near each other in documents.

Some visualization systems calculate a measure of overall relatedness among documents, based on word or concept usage, attribute similarity, explicit relationships, or some combination thereof. They then create a visualization where proximity indicates relatedness (Wise, 1995; Ilgen, 1997; Brodbeck, 1997). For example, Figure 1 shows a visualization of Shakespearean scenes using the SPIRE[1] Galaxies visualization approach. Each dot represents a single document. Documents that are more closely related are shown closer together; documents farther apart are less related.

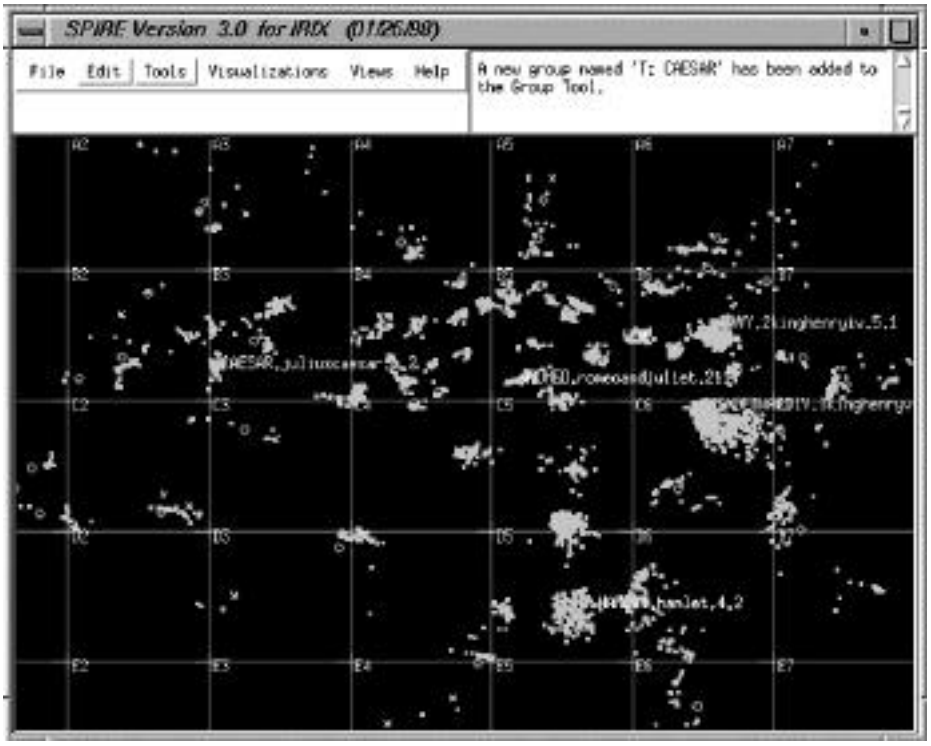

Figure 1: Galaxies Visualization Using Proximity to Show Semantic Similarity

Some systems represent document attributes more explicitly and help the user to perceive the relationships themselves (Nowell, 1996). For example, the type of document may be shown as a icon, the date of publication may be represented by shade of color, and the relevance to a particular query may be shown by vertical or horizontal position.

A Proof of Concept System for Visualizing Complex Relationships

Two prototype visualizations being developed at the Pacific Northwest National Laboratory are designed to help people understand and explore multiple relationships among documents. In our visualizations, we allow the user to control whether to view a combination of multiple relationships or to view them individually. We rely on the concept that relationships can be categorized into higher-order groupings; we also note that the specific assignment of relationships to categories is user- and context-dependent. The visualizations are 3-dimensional. Each has specific interactions as well as general zoom, pan, and rotate capabilities. The goal is to create a flexible tool that enables users to explore various relationship patterns among documents, according to their individual needs and interests.

As an example, we analyzed a diverse collection of documents related to Shakespeare. All were found on the World Wide Web from a variety of sites. One site publishes the scenes from all of Shakespeare's plays. Another site at Emory University publishes descriptions of paintings illustrating some of those Shakespearean scenes and histories of the artists who created them. The Richard III Society web site is replete with essays and debates disputing the English king's murderous portrayal by Shakespeare in his historical

plays, plus alternate accounts of King Richard's life from more sympathetic perspectives. Several sites publish argument essays and answering rebuttal essays debating whether plays currently attributed to Shakespeare were actually authored by others. Even more sites publish treatises by literary scholars discussing esoteric points of Shakespearean characters and themes. Altogether, we collected 900 Shakespeare-affiliated documents, representing a myriad of explicit and obscure relationships.

Our first visualization experiments used the explicit attributes and document relationships among this collection. The visualization shown in Figure 2 is based on a 3-dimensional extension to the matrix approach used by Becker (1995). In our visualization, the x and y axes both correspond to an ordered list of documents in a subset of the collection. The z axis corresponds to the various types of relationships that occur in this subset. Thus, in this example, a z-position corresponds to the relationship "precedes." If document "i" precedes document "j," then a small bead shape in that z-plane will correspond to the "i,j" spot. If document "i" has multiple relationships to document "j," there will be several such beads — perhaps a stack of beads — at various z-positions.

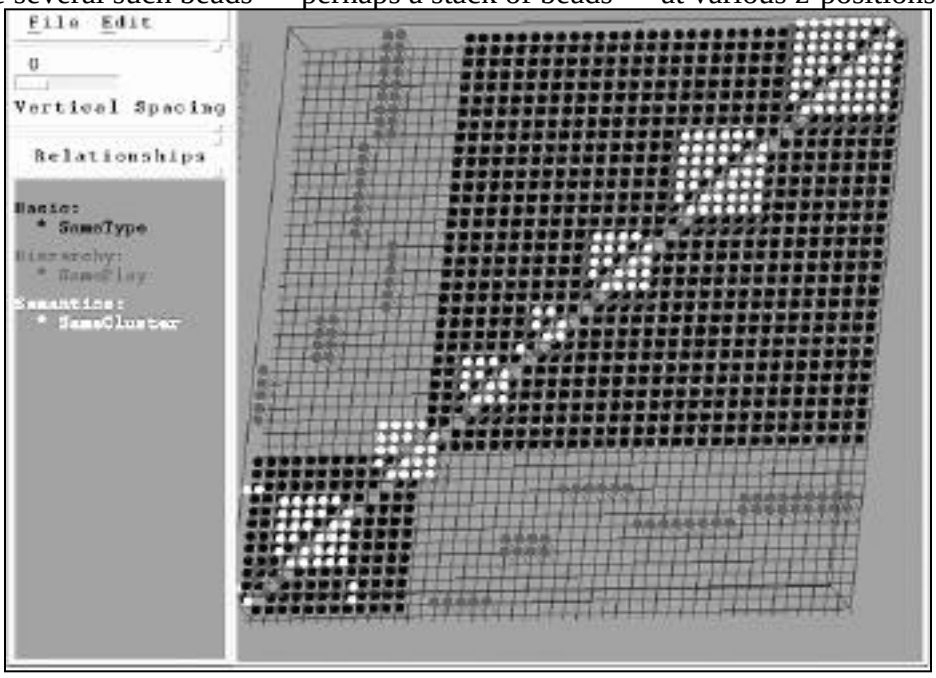

Figure 2: Connex View of Three Relationships

This visualization is called "Connex" because it visualizes relationships that connect the documents. The Connex visualization tool allows the user to categorize the various relationships, hide or show particular relationships at will, highlight asymmetric relationships, and reorder the document display. Colors can be assigned to categories to help group or distinguish different relationships. For example, in one situation, a user might find it helpful to group together relationships based on attributes, such as documents with the same author, same year of publication, same language, or same level of detail. A second relationship category might be thematic similarity; a third category might be sequential relationships. The visualization allows the user to examine such relationships individually, or in combination, seeing where documents contain the same characters or themes, yet do not come from the same play. Statistical methods can be

used to reorder the documents, making it easy to identify high concentrations of inter-document relationships.

Figure 2 illustrates the interplay of three different relationships. The black beads indicate items that are the same type. The order of the rows and columns has been assigned by mathematical correlation; the result shows the two document types represented here. The smaller group of black beads corresponds with descriptions of art illustrating Shakespeare's works; the larger group corresponds with selected scenes from several plays. The gray beads show items that relate to the same play. The white beads show items that have similar semantic content, as determined by SPIRE.

For users who are interested in particular relationships, the colors and layers help to highlight individual ones. For users who are interested in relationship combinations, the stacking of multiple colored beads may help to identify unanticipated combinations. For example, in Figure 2, the semantic content relationship generally matches the type relationship (scenes from plays tend to group together and art descriptions tend to group together). However, there is one exception. One set of documents that is semantically very similar contains both scenes and art descriptions; this is shown by the group of white beads that overlaps both groups of black beads near the lower left corner.

In our second experiment, we wanted to identify relationships among characters and concepts *within* the documents. One key goal of this experiment was to identify methods for detecting both evidence of association and evidence of disassociation. The fundamental data used for measurements of association and dissociation of a character or concept were based on counting the number of times the word(s) indicating the character/concept appeared in a document chunk (e.g., an essay, critique, or portion of a play). The table below shows part of these data for some characters occurring or mentioned in Shakespeare's plays.

Table 1: Occurrence of selected characters in a subset of documents.

Character	Romeo & Juliet 1.0	Romeo & Juliet 1.1	Romeo & Juliet 1.2	Romeo & Juliet 1.3	Romeo & Juliet 1.4	York Debate 1	York Debate 2	York Debate 3
Antony	0	0	0	0	0	0	0	0
Caesar	0	0	0	0	0	3	1	0
Romeo	0	20	13	0	15	1	0	0
Tybalt	0	4	1	0	0	0	0	0

These data can be viewed as providing a *vector* for each character or concept. Such data can be evaluated for associations via a number of mathematical or statistical approaches. After some experimentation and thought, we are currently using *cosine distance* as our association measure. The cosine distance takes two of the vectors and returns a number between 0 and 1. Values close to 0 indicate that the two characters or concepts rarely (if ever) occurred in the same text. Values close to 1 indicate that the two characters occurred together in the text. Thus, potential disassociations were indicated for pairs of concepts/characters with a distance near 0, and potential associations were indicated for pairs of concepts/characters with a distance near 1. We found it particularly interesting to compare such association evidence within various subgroups of the collection. For example, some characters were related within the play critiques but not within the plays themselves. In addition to such co-occurrence relationships, we also identified instances where one character mentioned terms of interest, such as other names or chosen concepts.

Note that the data in Table 1 are indicative of a disassociation between Caesar and Tybalt and positive association between Romeo and Tybalt.

A prototype visualization, called "Rainbows," provided a good method for displaying the results (see Figure 3). In Rainbows, the entities (in this case, mainly Shakespearean character names) are displayed as green dots on a plane. The location of the dots was determined by how the names appeared within the lines of the Shakespearean plays; names that tend to appear together are shown closer together on the plane. Mathematical clustering techniques can be used to determine groups of characters.

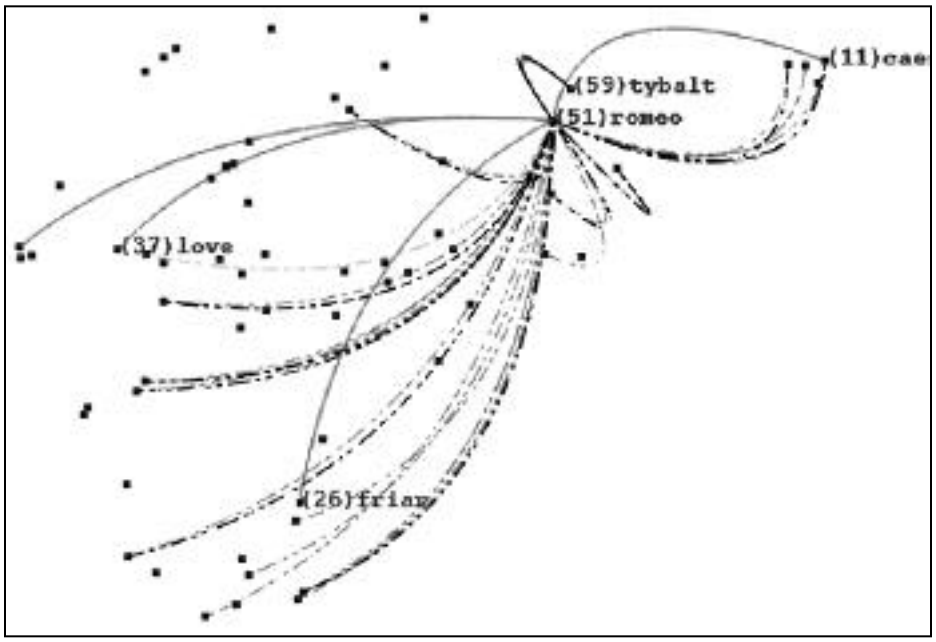

Figure 3: Rainbows View of Relationships

The user can interact with the visualization in various ways. For example, a user may wish to see what relationships exist between two clusters of entities. White arcs between clusters indicate the presence of such evidence. Arcs that go above the plane indicate evidence of an association (values within a threshold of 1); evidence of a disassociation (values within a threshold of 0) are shown by dotted arcs drawn below the plane. Like separating white light into its component colors, the user can expand these white arcs into multi-colored ones, each representing a different kind of relationship evidence, detected using the statistical analysis described above. The user may also see which entities in the cluster contribute to these relationships, or may expand the detail to show all entity-entity relationships between the two clusters. In addition, he or she may explore the relationships by type rather than by cluster.

Figure 3 shows relationships between selected characters and concepts. In this view, we have displayed relationships pertaining to Romeo. Note the expected (positive) associations to "friar," to Tybalt, and to love, and the unexpected one to Caesar. Some of the debate documents in our collection mention both Romeo and Caesar.

Discussion

The Connex and Rainbows visualization methods are independent of the particular data and relationships shown here. They could be used to visualize citations or hyperlinks, attribute similarities, indications of relationships between individuals or countries, etc.

Part of their strength comes from the ability to interact flexibly with the information. The abilities to see overview information, get more details about particular categories of relationships, move from cluster relationships to individual entity relationships, and hide or show relationships are all important capabilities in helping the user understand and explore such complex information.

Our current informal feedback indicates that the Rainbows visualization is particularly appealing to many people. The use of positive and negative arcs to represent and compare evidence of association and disassociation fits the way they think about the information. We would like to add the capability for users to interactively modify the thresholds. Being able to move a slider and see new arcs appear at different thresholds may provide a powerful addition.

We believe the strength of the Connex visualization is in the ability to re-order the entities according to many methods and to emphasize asymmetric relationships. We would like to add the ability to search for particular combinations of relationships.

Conclusion

Documents embody a rich and potentially very useful set of complex interrelationships. Information visualization is a powerful tool to help users understand and explore such relationships. This paper presented two novel visualization approaches for representing and interactively exploring these relationships among documents or entities. Research is continuing into alternative statistical methods for detecting evidence of relationships and in making the system more interactive and flexible to use.

Acknowledgements

We wish to thank the CIA's Office of Research and Development for their support in this research. We also wish to thank the rest of the Pacific Northwest National Laboratory Information Visualization Team, especially Vern Crow, Sharon Eaton, Tonya Martin, Grant Nakamura, Wendy Cowley, and Dan Donohoo, for their assistance in this work. Finally, we acknowledge the guidance and support of Renie McVeety and Jim Thomas, without whom this work would not be possible.

Notes

SPIRE stands for Spatial Paradigm for Information Retrieval and Exploration

References

Becker, R., Eick, S., and Wilks, A. (1995). Visualizing Network Data. In: IEEE Transactions on Visualization and Computer Graphics. Vol. 1, No. 1, March 1995, p.16-28.

Brodbeck, D. et al. (1997). Domesticating Bead: Adapting an Information Visualization System to a Financial Institution. In: Information Visualization '97. Proc. Phoenix, October 1997. IEEE Computer Society, p.73-80.

Card, S., Robertson, G., and York, W. (1996). The Webbook and the Web Forager: An Information Workspace for the World-Wide Web. In: ACM SIGCHI '96. Proc. Vancouver, Canada, April 1996.

Cooper, James and Byrd, Roy. (1997). Lexical Navigation: Visually Prompted Query Expansion and Refinement. In: Digital Libraries '97. Proc. Philadelphia, PA, July 1997. ACM Press, p.237-246.

Fidel, R. and Crandall, M. (1997). Users' Perception of the Performance of a Filtering System. In: ACM SIGIR '97. Proc. ACM Press, New York, New York. 1997, p.198-205.

Gershon, Nahum, LeVasseur, Joshua, Winstead, Joel, Croall, James, Pernick, Ari, Ruh, William. (1995). Case Study of Visualizing Internet Resources. In: Information Visualization '95. Proc. IEEE Computer Society, p.122-128.

Green, R. (1996). Development of a Relational Thesaurus. In: Knowledge Organization and Change. Proc. 4th Int. ISKO Conf., Frankfurt/Main: Indeks Verlag. p.72-79.

Green, Rebecca and Bean, Carol A. (1995). Topical Relevance Relationships. II. An Exploratory Study and Preliminary Typology. Journal of the American Society for Information Science, 46/9, p.654-662.

Hetzler, E. (1997). Beyond Word Relations. SIGIR Forum, Fall 1997. Vol 31, No. 2. ACM Press, p. 28-32.

Ilgen, M., Rushall, D. (1997). Visualization of Multi-lingual Free Text Using Self-Organizing Neural Network Techniques. In: Advanced Information Processing and Analysis. Symp. Tysons Corner, VA. March 1997. p.30-31.

Liu, M. (1993). Progress in Documentation — The Complexities of Citation Practice: A Review of Citation Studies. Journal of Documentation, 49, p.370-408.

Molholt, Pat. (1996). Standardization of Inter-Concept Links and Their Usage. In: Knowledge Organization and Change. Proc. 4th Int. ISKO Conf., Frankfurt/Main: Indeks Verlag. p.65-71.

Munzner, T. (1997). H3: Laying Out Large Directed Graphs in 3D Hyperbolic Space. In: Information Visualization '97. Proc. Oct. 1997, Phoenix: IEEE Computer Society, p.2-10.

Myaeng, S. H., Khoo, Chris Khoo, Li, Ming (1994). Linguistic Processing of Text for a Large-Scale Conceptual Information Retrieval System. In: Lecture Notes in Artificial Intelligence. Proc. 2nd Int. Conf. on Conceptual Structures, 1994. 835, Springer-Verlag, p.69-84.

Nowell, L. T., et al. (1996). Visualizing Search Results: Some Alternatives to Query-Document Similarity. ACM SIGIR '96. Proc. Aug. 1996, Zurich: ACM Press, p.67-75.

Storey, M., et al. (1997). On Integrating Visualization Techniques for Effective Software Exploration. In: Information Visualization '97. Proc. Oct. 1997, Phoenix: IEEE Computer Society, p.38-45.

Tillett, Barbara. (1991). A Summary of the Treatment of Bibliographic Relationships in Cataloguing Rules. In: Library resources and technical services. 35, no. 4, p.393-405.

Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., Crow, V.

(1995). Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents. In: IEEE ’95 Information Visualization. Proc. Oct. 1995, Atlanta: IEEE Service Center, p.51-58.