Information Visualization Interfaces for Retrieval and Analysis (IVIRA) Workshop Summary

Contributed by:
Javed Mostafa
Laboratory for Applied Informatics Research
<http://lair.indiana.edu>
Indiana University, Bloomington
Bloomington, Indiana, USA
<jm@indiana.edu>

The IVIRA workshop was held on Saturday May 31st in conjunction with the Joint Conference on Digital Libraries 2003 at the Rice University Campus in Houston, Texas. Katy Börner and Javed Mostafa jointly organized the workshop.

Javed Mostafa opened the half-day with a brief introduction in which he summarized the significance of three key areas in information retrieval visualization: functions, infrastructure, and evaluation.

- Functions of IV/IR systems include "prototypical" ones such as browse, search, refinement, and presentation. However, increasingly users demand more — features and tools to help them interactively mine data, generate patterns, and conduct analysis on data.

- In IV/IR systems infrastructure can play a critical role by providing access to different standard corpora for training and testing systems, pre-established results for validation, and well-tested modules to reduce development time.

- Evaluation is a related area requiring attention especially in an emerging field such as information visualization as evaluation can guide development and ensure that new techniques are based on sound empirical studies.

Javed concluded his introduction by observing that IV/IR systems are now being developed and used in a wide variety of application domains, ranging from music to proteomics; hence, it is imperative that researchers and practitioners reflect on the three general areas (i.e., function, infrastructure, and evaluation) in terms of domain-level implications.

Seven papers were presented at the workshop, each lasting approximately 20 minutes. The visualization approaches discussed can be grouped into three broad categories: user-centered, metadata-driven, and data mining applied on raw data.

The first paper entitled "Collection understanding through streaming collage" was presented by Michelle Chang. She discussed a technique that is unusual in that its main aim is to support comprehensive understanding of an image collection instead of facilitating retrieval of a narrow subset from the collection. The presentation method

takes advantage of Andruid Kerne's "Collage Machine" to generate streaming collages from the collection. Users can review randomly produced collages from the collection, manipulate collages by moving around individual images displayed, or filter collages by specifying metadata values associated with images. This is an example of a user-centered approach with significant control given to the user for manipulating the presentation of information.

The second presentation was by Yueye Fu and focused on visualization of email-boxes with the main goal of identifying patterns in communication behavior. A set of emails related to a collaborative software development project was used as the "collection" to illustrate the application of the Treemap visualization algorithm. Two Treemap visualizations were produced: one that coded individual senders differently based on the amount of emails each contributed and the other grouped emails based on specific projects. The Treemap approach presented is an example of a basic data mining technique applied on raw data to produce visualizations.

Eric Isaacson was the third presenter and his paper dealt with content visualization of a music digital library. Eric discussed several techniques that aid in visualizing musical scores and a technique that permits display of different homogeneous sections of music over time. An important aspect of the visualization techniques discussed by Eric was the fact that they permit the user to annotate or "edit" the presentation and thus provide a significant degree of control to the user to manipulate the visualization.

Following a brief break after the third presentation, the workshop continued with a presentation by Carlos Monroy. He described a system to create, visualize, and analyze series of art objects. For a given collection of art object images, the system permits the user to browse and select a subset, group them as a series, and save the series. Thus the user can build a catalog of self-created series. Additional functions offered include analysis of series content by creation time-line, animation of series content, and comparison of up to four series in parallel. This system is also a good example of a user-centered approach.

Next, Rao Shen described the GetSmart project — a collaborative effort between Virginia Tech and University of Arizona to integrate digital library tools in education. For this project, a knowledge visualization method has been implemented called "concept maps": the method involves displaying prominent concepts and association as a network graph with connection among nodes representing associations among concepts. Rao mentioned that such maps have been helpful in summarizing the main themes of a domain, allowing students to check their own understanding, and promoting collaborations.

The next paper presented was on development of a search system named Oncosifter that allows users to locate information on cancer using multiple means. One particular function supported by this system is search based on visualization of key concepts. The system maintains a database of "consumer health vocabularies" that are presented to the user as hyperbolic tree structure – the user can begin at any of the dozen or so top level

concept nodes and gradually refine the information needed by selecting lower level concepts that are also presented as clustered branches in the visualization. Based on the concepts selected the MedlinePlus and CancerGov databases are queried to retrieve related documents. Oncosifter is an example of a system that generates visualization from metadata that are then used to support retrieval functions.

The final paper was presented by James Cooper of IBM who discussed development of a concept relation discovery system in the biomedical domain. It uses a noun and technical term discovery library called JTalent, the output of which is filtered using the Medical Subject Headings (MeSH) dictionary. The MeSH dictionary is used to identify biomedical terms and a mutual information formula is employed to calculate association weights among terms. Finally, using a spring graph algorithm, the terms and term associations are displayed for the user to browse. This research is clearly in the category of data mining applied on raw text.

The workshop concluded with a brief discussion of the highlights of the presented techniques and how they relate to the basic taxonomy of user-centered, metadata-driven, and data mining approaches. It was pointed out by Javed that the workshop was being held in conjunction with one of the major digital libraries forums, and it is not unrealistic to expect that a large digital library may cover the whole range of basic data formats, i.e., text, image, audio, and video. Hence, a challenge for researchers and practitioners is to develop scalable information visualization techniques that can also be generalized across heterogeneous data formats.

Papers as well as presentation slides of all talks are online at <http://vw.indiana.edu/ivira03>.

Acknowledgments

Related Sites:

SPIE Conference on Visualization and Data Analysis, San Jose, CA, USA, January 18-22, 2004. <http://vw.indiana.edu/vda2004>.

Special Issue on Collaborative Information Visualization Environments in PRESENCE: Teleoperators and Virtual Environments. Submission Deadline: September 1st, 2003. <http://vw.indiana.edu/cive03/03-cive-presence.pdf>.

Symposium on Collaborative Information Visualization Environments, IV 2003, London, UK, July 16-18, 2003. <http://vw.indiana.edu/cive03>.