IN-SPIRE InfoVis 2004 Contest Entry

Pak Chung Wong, Beth Hetzler, Christian Posse, Mark Whiting, Susan Havre, Nick Cramer, Anuj Shah, Mudita Singhal, Alan Turner, Jim Thomas
Pacific Northwest National Laboratory
{firstname.lastname}@pnl.gov

ABSTRACT

This is the first part (summary) of a three-part contest entry submitted to IEEE InfoVis 2004. The contest topic is visualizing InfoVis symposium papers from 1995 to 2002 and their references. The paper introduces the visualization tool IN-SPIRE, describes its strengths and weaknesses, summarizes the visualization process and results, and presents lessons learned.

1 INTRODUCTION

This year's InfoVis Contest invites participants to analyze InfoVis symposium papers from 1995 to 2002 and their references. Based mostly on the paper abstracts (i.e., full papers are not included), the participants are asked to use visualization tools to: 1) identify major research areas, 2) characterize these areas and their evolution, and 3) discover collaboration relationships among researchers. (Tasks 3 and 4 are combined into one throughout our submission as suggested by the contest html template.)

For over a decade, researchers at the Pacific Northwest National Laboratory (PNNL) have developed a number of text visualization tools customized for different applications. Many of them have been presented previously at this symposium. We select one of them—IN-SPIRE [1]—to answer the questions.

This paper introduces the background of IN-SPIRE, discusses its strengths and weaknesses, describes the exploration process, and summarizes the contest results. At the end, we share some of our lessons learned throughout the contest effort.

2 IN-SPIRE

IN-SPIRE [1] is a visual-analytic tool primarily designed to unveil common themes and reveal hidden relationships within a large corpus. While the underlying metaphors of its two signature visualizations—Galaxy and ThemeView—have remained stable since the debut of its predecessor (SPIRE) [2], the design algorithms behind the visualizations and the analytical tools surrounding them have gone through multiple generations of intensive research and development (R&D) effort. This includes over two dozen new analytical features (some of which have just been released in June 2004,) which required over 200 major design changes. Figure 1 illustrates a visualization process of a document corpus using IN-SPIRE.

Today, on a modest Windows desktop computer, IN-SPIRE can harvest large numbers of documents from multiple sources in different formats, ingest both static and dynamic corpora, identify major and minor themes, query topics and seek evidence, and conduct short-term analyses and long-term monitoring. As an ongoing R&D project at PNNL, IN-SPIRE continues to grow into a multi-purpose visual-analytic tool that has spawned two spin-off companies specializing in bioinformatics and text analysis. Interested readers can download a demo copy from [1].

3 STRENGTHS AND WEAKNESSES

The development of IN-SPIRE has been a long and fruitful R&D journey among the developers and end users. We have, in a sense, co-invented many practical solutions for many real-life problems. While we have enjoyed great success using IN-SPIRE in many applications, we still need to address open problems such as natural language processing and understanding. Here we discuss some of the strengths and weaknesses of IN-SPIRE.

The Galaxy and ThemeView visualizations help provide user insight into overview themes, relationships between themes,

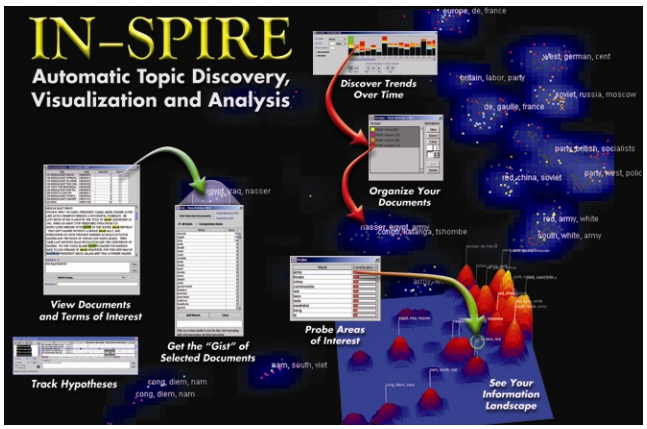

Figure 1: Visualizing a document corpus using IN-SPIRE.

document similarity, and relation of query results to overall thematic content. However, they require learning and practice for users to realize their benefits. While the visualization paradigms have remained stable for several years, much of our recent work has concentrated on visual interaction capabilities and high-value analysis features, based on studies such as [3]. This emphasis is reflected in the contest analysis as shown in the video.

A recently added strength is the ability to change the Galaxy, e.g., to converge on a smaller portion of a dataset. Users can "move aside" uninteresting clumps to see more detail on the remaining ones that are re-clustered and re-projected as a result. Alternatively, users can select interesting documents, e.g., by querying, and easily see a re-clustering of those, with the remainder moved aside (see video). A weakness of our current implementation is that it provides no easy way to explicitly compare before and after groupings of documents.

Another way that the user can modify the information space is to select words, such as peak labels, and interactively remove their influence. This allows the user to begin to tune the analysis to their interests. A weakness is that there is currently no way for a user to interactively "nominate" words to promote as high influence words.

Until recently, a weakness was the lack of ability to add new documents to an existing dataset. Now users can update datasets, or merge two datasets, and the processing takes advantage of the previously known document analysis.

IN-SPIRE's text characterization and clustering capabilities are based on proven statistical word pattern technology to identify document themes of a corpus. It does not rely on any English-specific lexicons or grammars, except for the optional use of a stopword list. This makes it extremely flexible and allows us to apply the approach to foreign languages or specialty languages. A disadvantage of this approach is that IN-SPIRE has no ability to take advantage of existing domain knowledge, such as that encoded in an ontology. It does not apply the concept of natural language processing to try and "understand" the documents. It also is currently single-word based and doesn't include or consider word phrases in its analysis. The ability to leverage or combine some of these approaches is currently being researched. For example, the ability to use ontologies to link document themes to a larger knowledge context is a prime topic currently under investigation by researchers at PNNL.

In addition to flexibility, IN-SPIRE's approach and implementation give it speed. It can process tens of thousands of magazine-sized documents on a modest Windows PC in a few minutes. Once a corpus is processed, visual interactions respond very quickly.

IN-SPIRE excels at the analysis of unstructured text. It includes many practical analysis tools to query for relevant documents, see time trends, browse, and explore hypotheses. It is by no means a Swiss-Army knife. Certain prevailing visualization techniques, such as a traditional graph display with nodes and links, are missing in its toolbox. For tasks that require graph visualization, we usually use tools such as Starlight [4], which was also developed at PNNL.

4 Data Ingesting and Processing

IN-SPIRE can read free-formed ASCII text as well as most of the standard data formats, including XML and HTML. The XML tags in the contest data allow IN-SPIRE to query different combinations of fields to generate visualizations, query topics, and explore evidence. We used an in-house data cleanup tool to standardize the dates recorded in different formats and to mark as symposium papers the documents with source tags that matched any of the eight InfoVis proceedings titles.

IN-SPIRE uses the unstructured text content of documents (e.g., the "abstract" field) to identify an interesting set of words known as *topics* or *themes* that can be used to distinguish clumps of similar documents within the collection. This process is based on the particular word patterns in the collection at hand and does not transfer to other collections. The co-occurrence or lack of co-occurrence of these interesting words and other statistically associated words in documents is used to build a richer thematic meaning for these representative topic words. Commonly appearing words that do not directly contribute to the content—typically prepositions, pronouns, gerunds, etc.—are ignored.

The system uses these topic word and associative patterns to build n-dimensional signature vectors characterizing each document. The vectors are clustered and projected to 2-space to create two visualizations—ThemeView and Galaxy.

5 Exploration Results

In answering the contest tasks, we used only the abstract and title fields to characterize the articles. We chose not to use the keyword fields for three reasons: 1) Many documents do not contain keywords, 2) many prevailing visualization keywords are missing from the ACM Computing Classification System [5], and 3) we wanted the text content to determine the similarity, not the human keywording. Due to limited space, we present only the most important result of each task. Readers are referred to the video and HTML submissions for demonstrations and detailed discussions.

5.1 Static Overview of 10 Years of InfoVis

We use two images to reflect and contrast the magnitude of contributions of the symposium papers to that of the reference papers. Our visualizations indicate that every major area identified by IN-SPIRE contains both symposium and reference papers. This suggests that the growth of the symposium papers through the years is consistent with the reference papers that represent a much larger community effort.

5.2 InfoVis Research Areas and their Evolution

We demonstrate the IN-SPIRE time slicer tool to animate the evolution of major research areas and the IN-SPIRE outlier tool to examine the details of the document clusters in a multi-scale fashion. Our work clearly reveals the evolution of the community since 1974: from the humble beginning when there were no focal topics in the community, to the domination of user design themes prior to InfoVis '95, to the diverse and topic-rich period throughout the rest of the 90s, and the dominance of the trees/hierarchies/graphs themes from 2001 to 2002.

5.3 The People in InfoVis

We apply the IN-SPIRE word usage statistics and probe tools to look into the contributions of two information visualization research pioneers—George Robertson and Stuart Card, their professional relationships, their technical contributions to the community, and their influence to the research community overall. We also conduct a similar investigation on researchers at PNNL. In all cases, the results generated automatically by IN-SPIRE are consistent with reality.

6 Observations and Lessons Learned

This has been a great learning experience for us at PNNL who participated in the contest effort. We are compelled to share some of our observations and lessons learned with our fellow readers.

We all agree that we could have done a more comprehensive analysis if we could work with the full papers instead of just abstracts and titles. While most of the abstracts are descriptive enough to identify the main themes of the papers, IN-SPIRE is capable of further distinguishing the characteristics of the contents and giving a finer degree of separation in its multi-scale visualization.

The contest organizers get an enthusiastic thumbs-up for providing the contest data in XML format. The hard work of the volunteers has paid off. We found the pre-defined field tags extremely useful to query the data fields in IN-SPIRE and later describe the process in the contest report.

Unfortunately, we did find a few data entry errors in the XML file. For example, one of the Vis95 papers has an incomplete "datefrom" entry, which caused IN-SPIRE to miss the paper in our time evolution task until we found the mistake.

We all agree that the contest tasks are challenging and stimulating. It is a great opportunity to look back at the history of the symposium and review our past. As long-time practitioners and technical contributors of the information visualization community, we can't help but ask, "What is the Next Big Thing in Information Visualization?"

7 Conclusions

We use a locally developed visual-analytics tool—IN-SPIRE—to take on the challenges of analyzing eight years of InfoVis symposium papers and their references. Using only IN-SPIRE's built-in tools, we are able to answer all the contest questions and provide quality insights into the corpus. The results presented in this summary, the video, and the supplemental file are consistent with the reality of the community.

Acknowledgments

The Pacific Northwest National Laboratory is managed for the U.S. Department of Energy by Battelle Memorial Institute under Contract DE-AC06-76RL1830.

References

[1] http://in-spire.pnl.gov/

[2] J.A. Wise, J.J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, "Visualizing The Non-Visual: Spatial Analysis And Interaction with Information From Text Documents,' *Proceedings IEEE Symposium on Information Visualization '95*, pages 51-58, IEEE CS Press, October 1995.

[3] Elizabeth G. Hetzler, Alan E. Turner. "Analysis Experiences Using Information Visualization," IEEE *Computer Graphics & Applications*, Fall 2004. To appear.

[4] http://starlight.pnl.gov/

[5] http://www.acm.org/class/