Visualization of the MEDLINE Database for Prostate Cancer

Xingye C. Lei, Ph.D., Paul Whitney, Ph.D.,
Dennis McQuerry, B.S., Beth Hetzler, M.S.
Battelle Memorial Institute, Richland, WA 99352
and
Leroy J. Korb, M. D.
Northwest Hospital, Seattle, WA 98133

ABSTRACT

This paper describes the application of modern data analysis systems to prostate cancer research articles. The articles are studied as a group to address the questions:

1. What are the trends in diagnostic and treatment research?
2. Which organizations and individuals are engaged in prostate cancer research, and how has their participation varied over time?

In this paper, the text analysis software Spatial Paradigm for Information Retrieval and Exploration (SPIRE“) and general statistical software SAS¤ were used to address these questions.

INTRODUCTION

The study of cancer diagnoses and treatments can be facilitated and enhanced with the vast amount of information available on the Internet and in other databases such as MEDLINE. Because of its social and societal impact, much research has been conducted in the diagnosis, treatment, and management of prostate cancer[1,2]. >From 1992 to 1999, more than 11,600 citations were found in MEDLINE for prostate cancer alone. The following questions are important to researchers and to those considering the strategic direction of cancer research:

1. What are the trends in diagnostic and treatment research?
2. Which organizations and individuals are engaged in prostate cancer research, and how has their participation varied over time?

It is an impossible task for an individual, or a small group of individuals, to manually review (i.e., read and digest) all 11,600 articles about prostate cancer to answer these and other questions. However, recent advances in information retrieval and visualization have made it possible to consider the analyses of such collections.

In this paper, we will show how the SPIRE (version 3.3, 1998) text analysis software and the statistical analysis software SAS¤ facilitates answering the questions for the prostate cancer articles taken from the MEDLINE database.

DATA AND METHODOLOGY

The data for this analysis were obtained through downloading information about prostate cancer on MEDLINE from 1992 to 1999. 11,613 MEDLINE records and documents were obtained, including author and co-author names, titles, journal names, publication dates, keywords (medical subject headings or MESH), the first author s affiliation, and, when available, abstract.

To analyze text documents and records, text processing that eventually results in interpretable numerical encoding of the text is needed[1,2]. Strategies for obtaining numerical encodings of text are described in Lebart , Wise, and Deerwester[3,4,5]. For the analyses described below, we used SPIRE to extract text features and create the numerical coding for the abstracts. Further analyses with SPIRE and SAS were based on those extracted text features and the other available data describing the article.

RESULTS AND DISCUSSION

Selected analyses are shown below. Figure 1 shows an application of bi-plots (obtained using SAS) to the problem of understanding the trends in the topics in the collection of abstracts. It shows the relationship between the clusters of the documents and the time a document was published; relatedness is indicated by proximity. A cluster with topics (rr, workers, ci) and (rr, worker, black) are closer to 1995 and 1996, indicating that more epidemiological studies were performed in 1995 and 1996 than in other years.

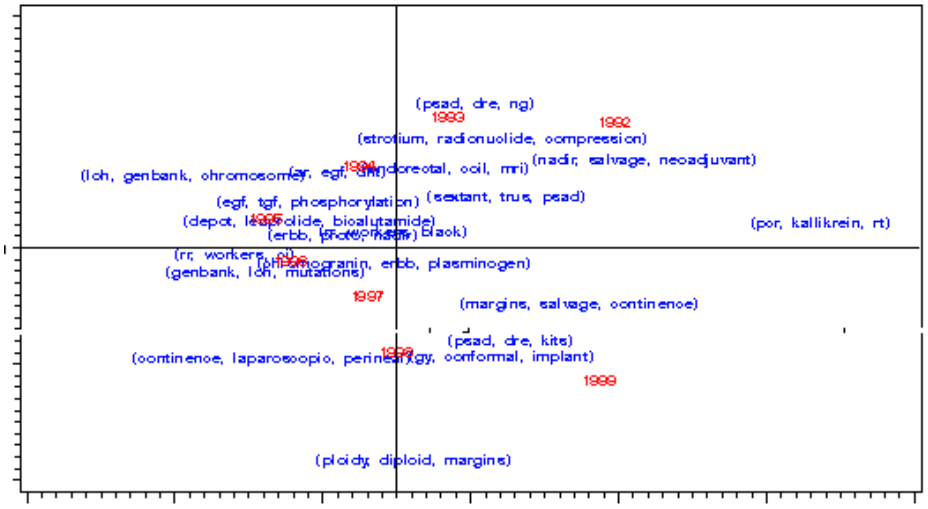

Figure 1: Correspondence analysis (SAS CORRESP procedure) showing the relationship between the clusters of the documents and the time the document was published; relatedness is indicated by proximity.

Figures 2 and 3 show some of the capabilities of the SPIRE system as applied to the analysis of how the prostate cancer research retrieved by the MEDLINE query is distributed topically and in time for the eight most-published research institutions. Comparing Figures 2 and 3 reveals that research interests over the two time periods changed for some institutions but not for others.

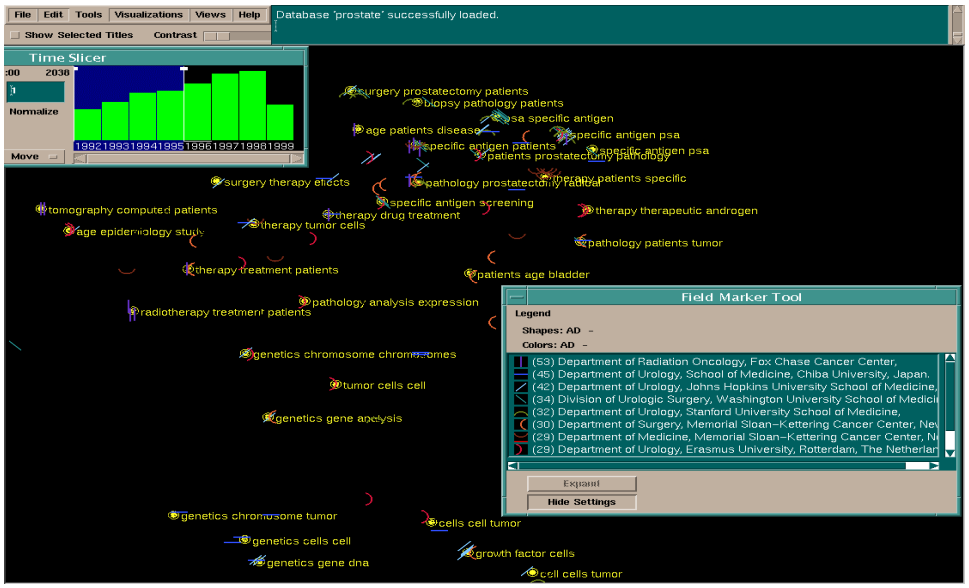

Figure 2: SPIRE Galaxies view for 1992 - 1995. Symbols and colors correspond to the top eight research institutions.

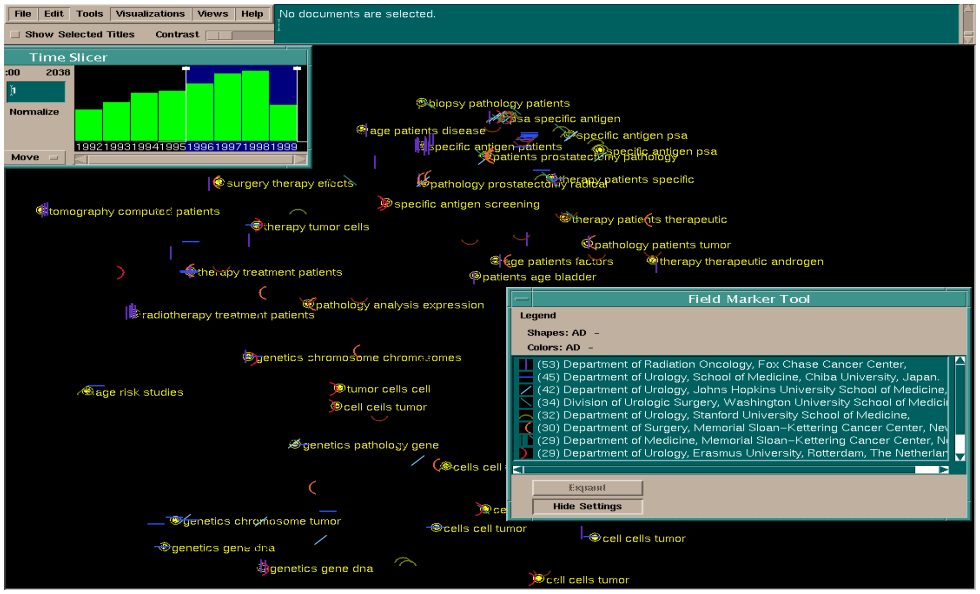

Figure 3: SPIRE Galaxies view for 1996 - 1999. Symbols and colors correspond to the top eight research institutes.

It is now possible to analyze document collections in the same way that collections of numerical or categorical data are analyzed. This analytic capability is a valuable help in increasing our understanding of the ever-growing research and related literature in medical science.

References

1. Cersosimo, RJ, Carr, D. Prostate Cancer: Current and Evolving Strategies. *Am J Health-Syst Pharm*, 1996; 53: 381-396.
2. Mettlin, C. Changes in Patterns of Prostate Cancer Care in the United States: Results of American College of Surgeons Commission on Cancer Studies, 1974-1993. *Prostate*, 1997; 32: 221-226.
3. Lebart, L. Visualizations of Textual Data. In *Visualization of Categorical Data* edited by Blasius, J, Greenacre, M. Academic Press. 1998.
4. Wise, JA, Thomas, JJ, Pennock, K, Lantrip, D, et al. Visualizing the Non-Visual: Spatial Analysis & Interaction with Information from Text Documents. In *Proceedings of IEEE '95 Information Visualization*, pp. 51-58, Atlanta GA, Oct. 1995. IEEE.
5. Deerwester, S, Dumais, ST., Furnas, GW, Landauer, TK., and Harshman, R. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information*, 1990; 41:391.