AD_____

MIPR NUMBER 96MM6709

TITLE: Evaluation of Spatial Paradigm for Information Retrieval and Exploration (SPIRE) Technology for Trauma Data Analysis

PRINCIPAL INVESTIGATOR: Sam N. Stevens, Elena Mendoza, Dennis McQuerry

CONTRACTING ORGANIZATION: Department of Energy, Richland
Richland, Washington 99352

REPORT DATE: October 1996

TYPE OF REPORT: Midterm

PREPARED FOR: Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1996	Midterm (19 Apr 96 - 31 Oct 96)

4. TITLE AND SUBTITLE
Evaluation of Spatial Paradigm for Information Retrieval and Exploration (SPIRE) Technology for Trauma Data Analysis

5. FUNDING NUMBERS
96MM6709

6. AUTHOR(S)
Sam N. Stevens, Elena Mendoza, Dennis McQuerry

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Department of Energy, Richland
Richland, Washington 99352

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, Maryland 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200*

The trauma data set was processed and analyzed by SPIRE. The data set was small and highly structured. Both of these facts were limiting factors in the SPIRE analysis. SPIRE accurately portrayed the information content; however, more research is needed for SPIRE to assess information from a small, structured data set.

14. SUBJECT TERMS
Information Retrieval, Themes, Textual Analysis

15. NUMBER OF PAGES
14

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Elena Mendoza 10-29-96

PI - Signature Date

Table of Contents

Introduction

The US Army MRMC is interested in applying Spatial Paradigm for Information Retrieval and Exploration (SPIRE) technology to data sets of trauma related information. SPIRE has generally been designed to work with larger data sets of a more unstructured nature (i.e. newswatch data, message trafficking). The purpose of this research was to determine if SPIRE would produce meaningful analysis of this trauma data. The trauma data provided by Dr. Howard Champion of the University of Maryland, has been processed and analyzed. The data consisted of 86 documents from two different sources pertaining to auto accidents. Both sources of documents discuss traumas; however, one set discussed more of the accidents, while the other source discussed more of the medical problems. SPIRE identified the two major differences in data sources, but, according to our analysis, could find little other meaningful differentiation among the documents. From our initial analysis, it appears as though the low number of documents and the structured nature of those documents impacted the performance of SPIRE on this data set.

Task Description and Study Results

SPIRE technology, as it exists today, assesses the patterns of word use in documents (e.g. frequency clusters of an individual word implies salience in theme) to determine important topics and relationships. Natural language communication utilizes definite strategies for conveying the content particularly when substantial knowledge is not assumed. SPIRE is dependent on this. Because of the inherent structure in the trauma data, the SPIRE system has the tendency to identify most documents in this data set as very similar. As a consequence, thematic differentiation of this data set is less pronounced than we would normally consider desirable.

Several different analyses were generated to assess the nature of this document set. First, we examined the correlation matrix (which establishes the correlations between all major terms and the key topics), and found consistently lower than normally acceptable values. A subset of the matrix can be found in **Appendix – Figure 1**. A random sample of terms is provided to convey the nature of the correlation matrix content. The term at the top of each column is the term to which other terms are related. The values are normalized and in a more strongly related set of documents may average 0.6 or higher. As the sample conveys, the average for the trauma data is about 0.15 for the ten most related terms. Additionally, the number of connected terms in a "normal" data set is typically much greater (terms with non-zero values), which again demonstrates the structured nature of this data. Generally, the information found in this matrix indicates that the documents don't tend to group into easily differentiated clusters and that relationships which would otherwise be small enough to ignore, can have a dominant influence on thematic distribution.

A second analysis we performed was to identify the number of statistically important topic terms in the data set relative to the unique word count. This is commonly called the noise to signal ratio, the "signal" being the important topics and the rest of the vocabulary being the "noise". There were only 13 major topics found in the data set with 2806 words in the vocabulary. This percentage (.005%) is extremely low for significant analysis. Significantly fewer high-value topical terms were found in this data set than were found in more normal expressions of natural language communication such as news articles or research papers or even WWW pages. For example, a data set run on CNN news data (635 small documents) gave approximately 450 major terms and 10,000 words in the vocabulary for a percentage of 4%. The trauma data set possesses a high noise to signal ratio, making it difficult to find meaningful information.

Appendix – Figure 2 is another illustration of the lack of discriminating dimensions in the data set. In this graph, there are three documents which appear in very different locations in the 2D scatter plot: two are proximal, the third is distant from the first two which is shown by **Appendix – Figure 3**. Thirteen topics along with their relative magnitudes are graphed. Typically, we would expect to be able to quickly identify proximal and distant document pairs due to the strong diversity in content represented by the magnitudes of each dimension--the relative magnitudes at each dimension of proximal documents would be close while the values for distant documents would be quite different. A couple of dimensions or topics in this data set, "pilon" and "travel" follow this pattern. However, the other topics show more random values for both proximal and distant pairs. This again shows that the data is not rich in discernible content, at least as measured by SPIRE.

There were, however, some positive results that are worthy of note. We were able to identify thematic clusters which could provide some insights to an analyst, depending upon their domain-specific requirements. For example, a query on the word, "tibia," shows that all tibial fractures tend to group together in the lower middle quadrant of the themescape as shown in **Appendix – Figures 4 and 5**. Further exploration with a larger data set might enable us to discover correlations with such elements as vehicle type or speed, age of "case," and so on.

Conclusions

In summary, what we've determined is that this particular data set has a degree of structure which makes it difficult for the SPIRE system to meaningfully process. Each document discusses the same general topics in the same general language; therefore, the language used doesn't convey the importance of words in the manner to which SPIRE is tuned. There is, however, information in the structure itself that has potential. Our domain expert, Dr. Howard Champion, agrees with this analysis and believes that SPIRE accurately portrayed the information in the trauma data.

At least two short term steps might be taken to improve the quality of the results. First, acquiring a larger data set might yield a better correlation matrix in terms of strength and number of word/word correlations, although we are skeptical that more data of exactly the same type would produce appreciably better results. Second, it might be possible to separate or eliminate some of the non-injury related vocabulary in the belief that the remaining data will map out more meaningfully. This effort would eliminate some of the "noise" in this signal.

Further research into this type of data and how to process and visualize it in a meaningful way is required for more substantive progress. The field of visual analysis of structured data is new and innovative. Research would include understanding the structure and gaining knowledge from the structure. It would also include new ways of visualizing structured data that might combine current data mining techniques with new visualization techniques such as SPIRE. SPIRE could be expanded to support this area in conjunction with it's current architecture for dealing with unstructured text.

Appendix – Figures and Charts

Figure 1. Correlation Matrix

The following matrices show a random sample of 12 terms and the normalized correlation value of 10 related terms. Strongly related document sets average 0.6. Trauma data average is approximately 0.15.

Term: 1st

Related Terms:		Normalized correlation value:
	victim	0.266667
	embankment	0.266667
	plateau	0.133333
	pilon	0.133333
	concrete	0.133333
	tibial	0.0666667
	pole	0.0666667
	metatarsal	0.0666667
	airbag	0
	travel	0

Term: 5th

Related Terms:		Normalized correlation value:
	metatarsal	0.266667
	cuboid	0.2
	embankment	0.2
	bimalleolar	0.133333
	airbag	0
	lisfranc	0
	victim	0
	travel	0
	tibial	0
	pilon	0

Term: 3rd

Related Terms:		Normalized correlation value:
	lisfranc	0.266667
	cuboid	0.266667
	victim	0.266667
	metatarsal	0.2
	airbag	0.0666667
	travel	0
	bimalleolar	0
	embankment	0
	pole	0
	pilon	0

Term: accelerator

Related Terms:		Normalized correlation value:
	travel	0.0666667
	tibial	0
	plateau	0
	pole	0
	concrete	0
	pilon	0
	victim	0
	metatarsal	0
	cuboid	0
	embankment	0

Term: 4th

Related Terms:		Normalized correlation value:
	metatarsal	0.333333
	cuboid	0.2
	embankment	0.2
	pilon	0
	pole	0
	concrete	0
	tibial	0
	plateau	0
	bimalleolar	0
	lisfranc	0

Term: access

Related Terms:		Normalized correlation value:
	bimalleolar	0.133333
	cuboid	0
	metatarsal	0
	embankment	0
	airbag	0
	lisfranc	0
	victim	0
	travel	0
	tibial	0
	plateau	0

Term: accident

Related Terms:		Normalized correlation value:
	airbag	0.0666667
	metatarsal	0.0666667
	plateau	0
	tibial	0
	pole	0
	concrete	0
	pilon	0
	victim	0
	bimalleolar	0
	cuboid	0

Term: airborne

Related Terms:		Normalized correlation value:
	victim	0.266667
	embankment	0.266667
	concrete	0.266667
	metatarsal	0.2
	plateau	0.133333
	airbag	0.0666667
	tibial	0.0666667
	cuboid	0
	bimalleolar	0
	pilon	0

Term: acetabulum

Related Terms:		Normalized correlation value:
	cuboid	0.133333
	victim	0.133333
	pilon	0.133333
	metatarsal	0.133333
	plateau	0.133333
	pole	0.0666667
	tibial	0.0666667
	travel	0
	lisfranc	0
	embankment	0

Term: apparently

Related Terms:		Normalized correlation value:
	embankment	0.4
	cuboid	0.4
	pilon	0.2
	metatarsal	0.133333
	tibial	0.0666667
	concrete	0.0666667
	victim	0.0666667
	lisfranc	0.0666667
	pole	0
	airbag	0

Term: airbag

Related Terms:		Normalized correlation value:
	airbag	1
	victim	0.2
	cuboid	0.2
	metatarsal	0.133333
	tibial	0.0666667
	pole	0
	pilon	0
	plateau	0
	travel	0
	lisfranc	0

Term: arm

Related Terms:		Normalized correlation value:
	airbag	0.2
	cuboid	0.2
	victim	0.2
	tibial	0.0666667
	bimalleolar	0.0666667
	pole	0.0666667
	travel	0
	plateau	0
	lisfranc	0
	concrete	0

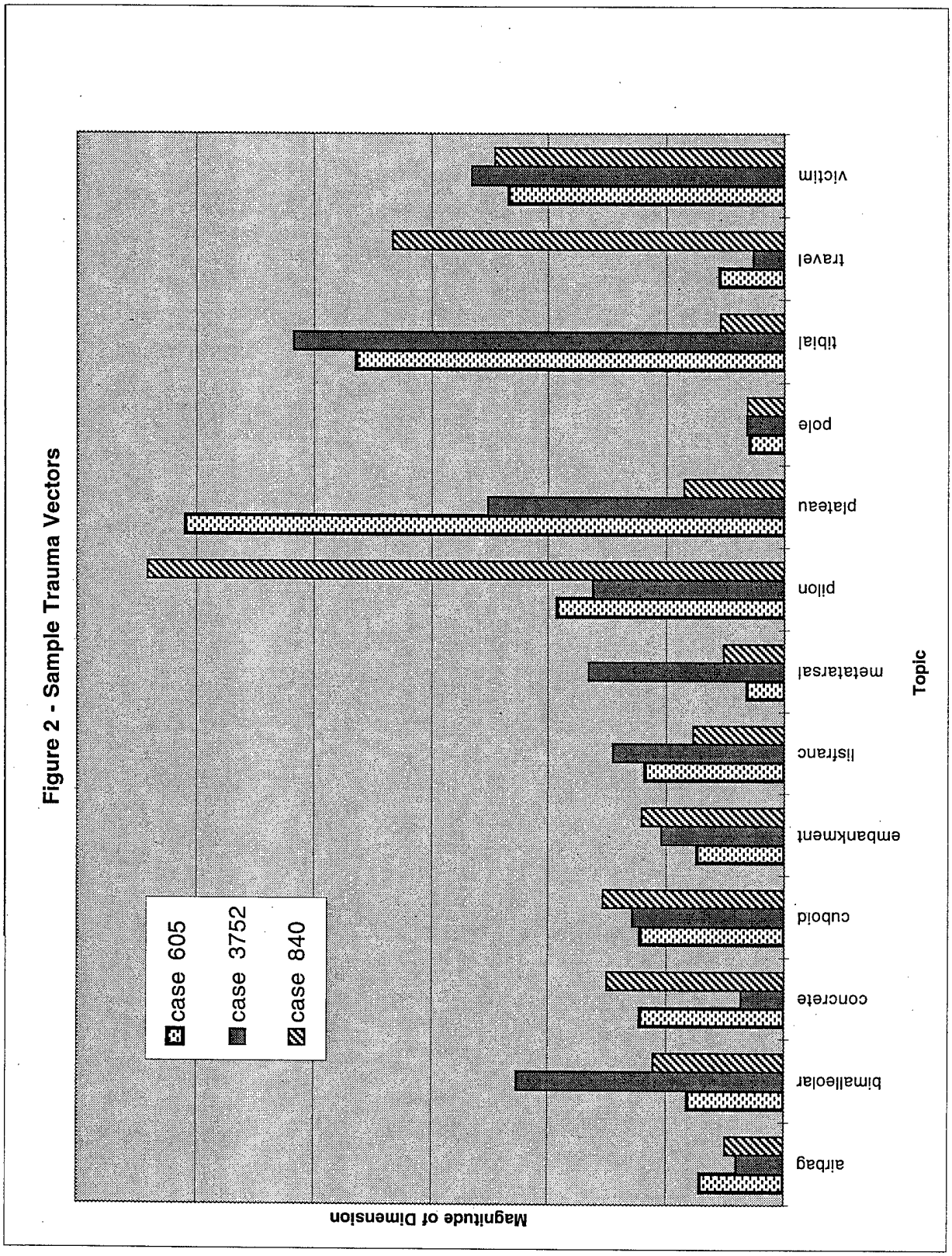
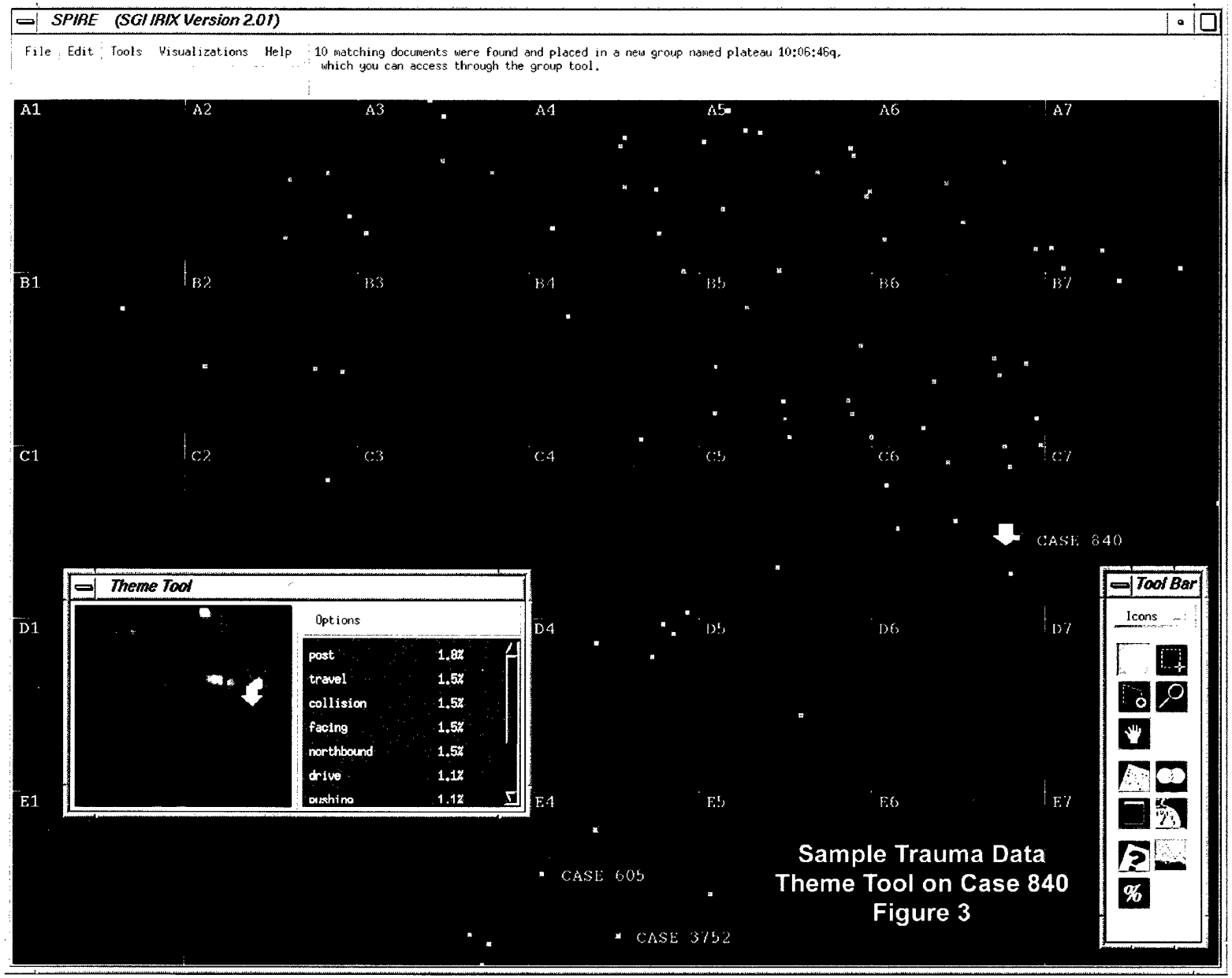
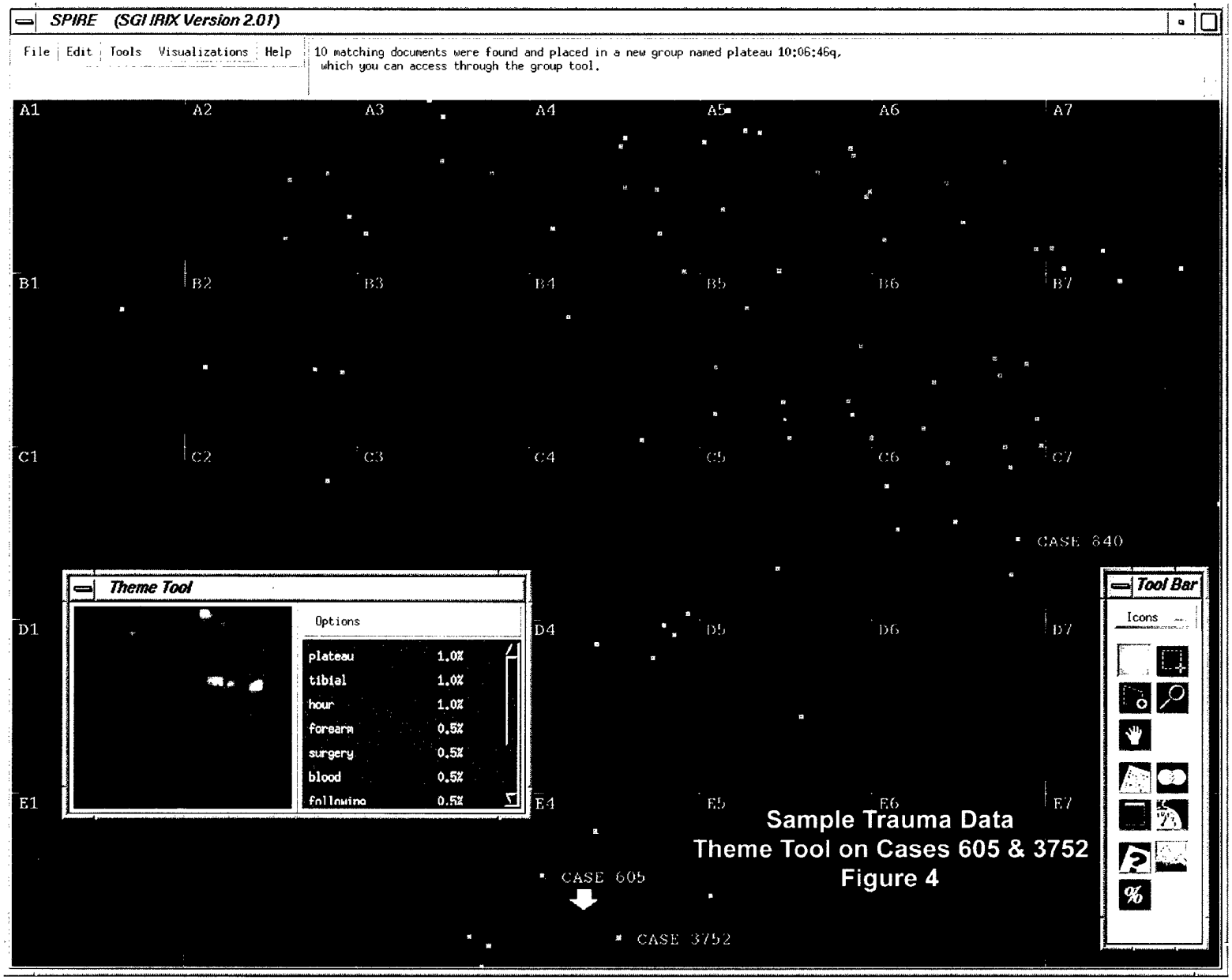


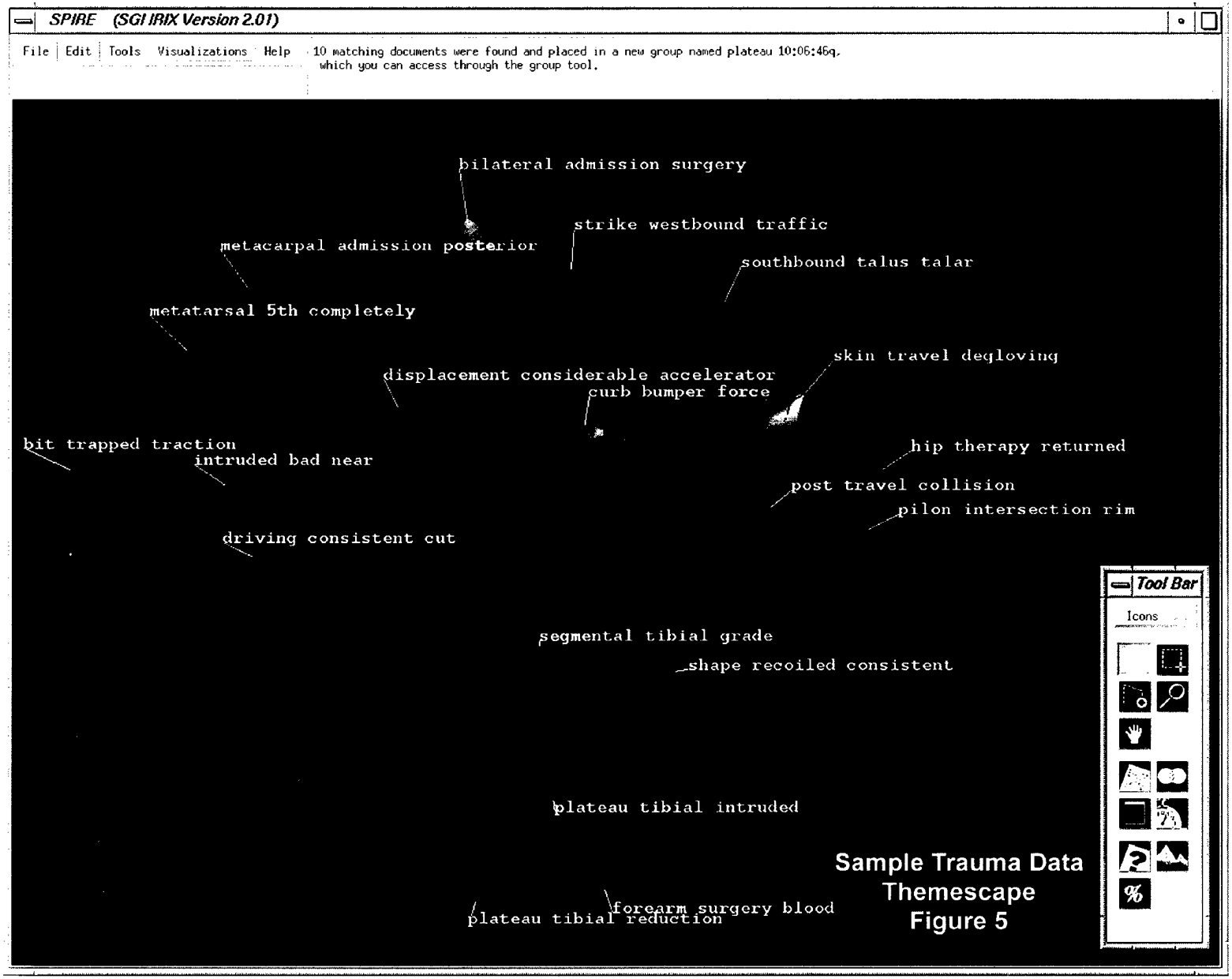
Figure 2 - Sample Trauma Vectors
case 605
case 3752
case 840
Magnitude of Dimension
Topic
airbag
bimalleolar
concrete
cuboid
embankment
listranc
metatarsal
pilon
plateau
pole
tibial
travel
victim



Sample Trauma Data
Theme Tool on Case 840
Figure 3



Sample Trauma Data
Theme Tool on Cases 605 & 3752
Figure 4



Sample Trauma Data
Themescape
Figure 5