Eurographics Italian Chapter Conference (2011)
A. F. Abate, M. Nappi, and G. Tortora (Editors)

Information visualization on organized crime trials

L. Di Silvestro[1], G. Gallo[1], G.Giuffrida[2], C.Zarba[3]

[1]Department of Mathematics and Computer Science, University of Catania, Italy
[2]Department of Social Science, University of Catania, Italy
[3]Neodata Group, Catania, Italy

Abstract

Today everyone on the Internet becomes an active writer. It is simpler than ever to produce and to share new knowledge on the net. Users risk to be overwhelmed by too much information. To handle unorganized data may involve a huge effort by people and this could lead to information overload. A promising way to face these problems is Information Visualization. This contribution is a report of using visuo-spatial reasoning to help sociology scholars to understand and to manage large amounts of textual data. We use a very interesting dataset, especially for sociologists and jurists, extracted from a collection of sentences of trials on organized crime activities in Sicily. We report the results obtained in so far clustering and visualizing data.

Categories and Subject Descriptors (according to ACM CCS): I.3.m [Computer Graphics]: Miscellaneous—

1. Introduction

Since 1960 thanks to Miller's psychological studies [Mil60] the term *information overload* has been used to refer to the difficulty a person faces understanding an issue or taking a decision because of the availability of too much information. In *digital age* an increasing number of people are connected to the Internet, they can use data and create news as well. They become active writers and produce more data for other viewers. Thousands of pages of text data are produced daily: now more than ever information overload is becoming a serious problem. It is easy to access information but too much of it is hard to manage and to understand and this can lead to misinformation. We need to find out efficient systems to manage large amount of information, exploring and analyzing the huge flow of new data gathered so far. Managing, exploring, and analyzing the flow of data are among the most important tasks for scholars of various disciplines.

One of the most efficient method to handle large amount of data and make them simpler to understand for people is using the visuo-spatial reasoning abilities of humans. [Tve05] It is clear that visualization is the key for content analysis. In this scenario a new field of research has been developed, to design and study interactive visual representation of abstract data. Information Visualization (InfoVis) is a rapidly growing field that is emerging from research in human-computer interaction, computer science, graphics, visual design and psychology. [SB03] Text visualization is considered one of the big challenges of the newly defined field of visual analytics. [Hea95] [Pal02] [HSH*02]

In this contribution we report of our initial experiments and efforts to use InfoVis to make a specialized corpus of textual information more accessible and useful. In particular we apply graph visualization to a collection of organized crime sentences. The final objective of this research is to provide to the crime analysts a tool to pool existing information into an organized database in order to gain a better understanding and forecasting of crimes. However, since we are still at the begin of this project our aim for the present is to gain some know-how about the major issues related to this specialized field.

This paper is organized as follows: Section 2 reports details of a case study. Section 3 describes some of our results and future works. In Section 4 conclusions are drawn.

2. Case study

Our data come from the legal domain. Legal scholars and social scientists are often interested in information extraction from a large number of texts. They need to analyze a very large amount of data to find out useful information to formulate and to verify social theories.

A valuable source of information about organized crime are the official trials' documentation. Empirical studies on

© The Eurographics Association 2011.

DOI: 10.2312/LocalChapterEvents/ItalChap/ItalianChapConf2011/035-039

delivered by EUROGRAPHICS DIGITAL LIBRARY www.eg.org diglib.eg.org

whole trials are not, up today, practical, due to the huge size of the complete trials' documentation. It is hence wise to restrict the analysis only to the final sentences. This is reasonable because a sentence contains all relevant elements that allow judges to take a decision. Indeed reading it, we can reconstruct the decision process. From sentences moreover, we can extract data to find out statistical results on age, genders, locations, etc. on criminal activities.

2.1. Collecting data

Sociologists are interested in studying organized crime. In Sicily organized crime is largely connected to mafia affairs. Italy has yet no central digital database of past sentences. This makes expensive and difficult to gather this kind of data. Since sociologists' interest in this topic is high, a research group in Catania decided to invest into this data gathering activity.

It took about 30 months/man to get together all the information to create the dataset used in this study. The gathering of sentences have been done in the archives of all the major appellate justice courts in Sicily where the trials have been conducted.

Every paper sheet of the sentences has been xeroxed. The entire set is made of about 55000 pages (sentences length goes from 2 to 3268 pages). Every page has been scanned producing PDFs files; after that, an OCR system has been used to extract textual information. A set of (unchecked) text files has been produced. [DGGZ10]

This expensive work makes the dataset very interesting and important to use for social studies, because it represent the only example in Italy of digitalized crime sentences corpus on mafia topics.

2.2. Dataset description

Our dataset collects all criminal sentences of trials on crime activities in Sicily pronounced from 2000 to 2006. In this set are included only crime sentences that became definitive in that years, about mafia and drug dealing cases. According to these principles, 721 sentences have been included in our study.

These text are obtained by using OCR system on a PDF copy of the original papers of sentences. Those papers are often written using typewriters, so some characters are difficult to read and to recognize. Sometimes there are handwritten notes on the sheets that are obviously not recognized. For these reasons there are a lot of characters with no meaning in our digital text. This makes our work for automatic information extraction harder.

2.3. Information extraction

Information extraction is a method to obtain structured data from unstructured natural language texts. With these techniques we can extract four type of information: entity, attributes, associations, and events. Entities can be persons, objects, dates and measures. Attributes are characteristics of entities, like birthday and birthplace of people, or their job. Associations are relationships between two entities that link them to each other. Events are associations for which temporal dimension is important.

For this preliminary work about visualization data to make easier the work of sociologists and jurists we decide to use mainly entities. Among entities we decide to extract only those that represent people, leaving out places and other entity types. During extraction we use context to understand the role of the person we have just found in the text. Four kinds of roles are recognized: "prosecutor", "judge", "lawyer" and "defendant".

Several finite state transducers (FST) are used to scan every document in sentences corpus. An FST is an automaton able to recognize specific patterns in an input string. [BB79] An automaton is build for each role we wish to identify in the text. For example layers are simple to recognize because their name come after an exact string that in Italian language denote their qualification (i.e. "`avv.`", "`avv.to`", "`avv.ti`", "`avvocato`", etc.).

During data extraction some important information about name's position is saved. For every occurrence of a name we know the unique id of the paragraph in which the name is found and the id of the sentence.

In the whole dataset there are 2475 entities referring to people.

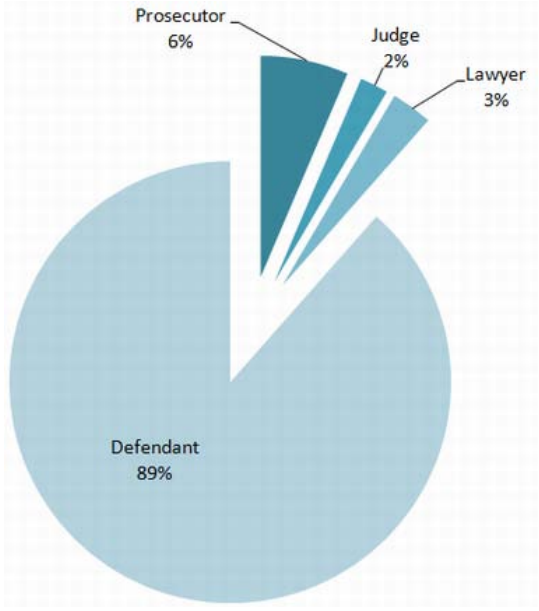

Figure 1: *Pie diagram for people divided by their roles.*

© The Eurographics Association 2011.

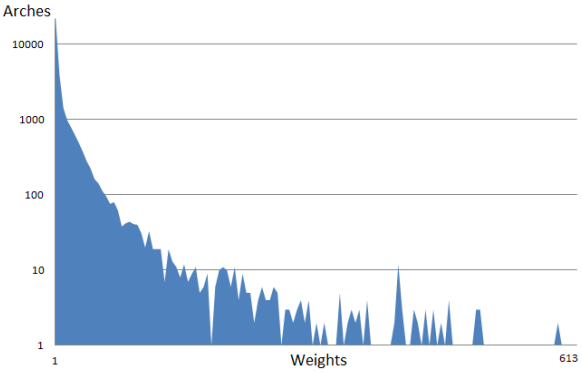

Figure 2: *Histogram showing number of arches for weight's value.*

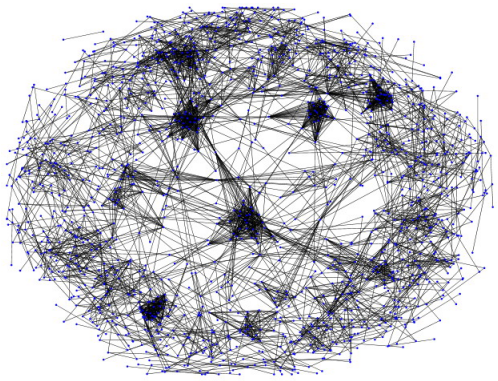

Figure 3: *Complete social network from data of sentences.*

A pie chart of the four roles for persons is shown in Figure 1.

2.4. Finding associations

We use data on entities and their location to find out a correlation among them. For our experiments a co-occurrence relationship is defined. Two entities are related if they appear in the same paragraph or in the same sentence. A Python script is used to identify associations. If a couple of entities co-occurs more than one time, the association between them is weighted accordingly. There are 32537 association. The weight for an association go from 1 to 613. The simplest way to show entities and their relationships is to create a graph in which vertices represent entities and arches between vertices represent the relationship of co-occurrences.

It is very tricky to build and to show a readable and easy to handle graph with more than ten thousand arch, so we choose to prune arches with lesser relevance. As diagram in Figure 2 shows, the arches with weight smaller than 5 are about 87,12%. Those arches represent a weak relation between entities, representing a very rare co-occurrence of names in the same sentence; if two person are not together in a paragraph more than few times, we can assume that the co-occurrence is not significative for our goals. If we maintain only arches with a weight grater or equal to 5, we have 4191 arches and 1436 entities connected in our graph.

2.5. Visualizing data

To build a graph with entities extracted from sentences in this initial study, we used a powerful free and open-source tool developed by the Social Media Research Foundation. NodeXL was created by Marc Smith's team while he was at Microsoft Research. [SSMF*09] It is a template for Excel that allows to easily build a graph entering a network edge list. With this tool it is not difficult to filter vertices and edges or calculate some graph metrics. We use this tool to obtain a clustering on our data. To identify vertices that are clustered together into subgroups of interest is indeed of great help. In this case clusters could identify important aggregates among offenders. We report anecdotically a correlation between network's clusters and mafia families, and considering the nature of connections among clusters sociologists have been able to infer some new knowledge.

2.6. Clusters

Clustering is among the main tasks of explorative data mining, and a common technique for statistical data analysis. NodeXL implements three clustering algorithms One of those is generally used to find community structure in very large networks: Clauset-Newman-Moore algorithm [CNM04]. Using this algorithm we have assigned a different color to each cluster and bound every cluster in a box.
As is shown in Figure 4, each box has an area proportional to the number of vertices that are contained in the cluster. We use different shapes to specify the role of person: circles for defendant, squares for lawyers, triangles for judges and diamonds for prosecutors.
The size of vertices depends on the *betweenness centrality* of the node [Fre77]. Betweenness centrality is defined as follows:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{1}$$

where σ_{st} is total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass trough v.

A vertex with high betweenness centrality often acts like a bridge between two clusters, perhaps indicating a key role of the person in the small society depicted by our sentences data.

© The Eurographics Association 2011.

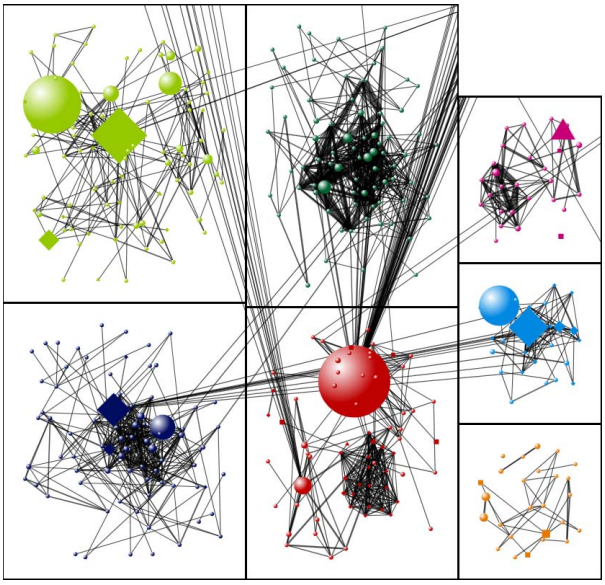

Figure 4: *The diagram shows a portion of the whole graph. The seven rectangles bound seven clusters of persons. Bigger icons denote a higher betweenness centrality. The shape of a vertex represents the role of the corresponding person. See text for details.*

The thickness of edges in the diagram is proportional to their weight.

It is possible to zoom in and visualize a cluster. In the example in Figure 5 we can see a cluster of defendants only labeled with their names. They are connected to a lawyer with very high betweenness centrality. This shows that almost all people in this small group use to be defended in court by the same lawyer. This lawyer is, moreover, representative of another cluster of people, he acts like a link between two group of offenders.

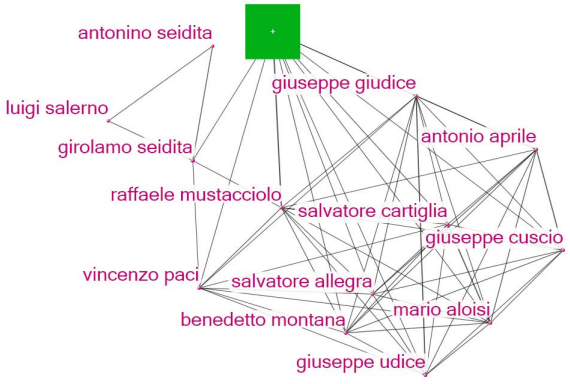

Figure 5: *A cluster in detail.*

3. Results

Working on this case study, we have been able to experiment on classic principles of information visualization. We have worked on data to achieve the five advantages defined in 2004 by Ware [War04]:

1. Comprehension: Visualization provides an ability to comprehend huge amounts of data.
2. Perception: Visualization reveals properties of the data that were not anticipated.
3. Quality Control: Visualization makes problems in the data (or in the data collection methods) immediately apparent.
4. Focus + Context: Visualization facilitates understanding of small-scale features in the context of the large-scale picture of the data.
5. Interpretation: Visualization supports hypothesis formation, leading to further investigation.

Simple information extraction from sentences produces a rough network from the data (Figure 3). This is nearly useless. It is a way to represent data, but it is not readable: we can't understand data and it does provide very little insight in the structure that it wish to represent. We have tried several combinations of graphical accessory elements like color, size, shape, location, thickness, to code as much information as possible in a single image. By trial and errors we believe that we obtained a graphical representation that at a quick glance may help scholars to roughly grab many important information otherwise very hidden in our data.

Although a rigorous usability test of the proposed graphical layouts are still in progress, we may safely claim that these elements are of great help to navigate and understand network data. Definitive data about the testing will be produced soon.

4. Conclusion

This work has to be considered an attempt to realize a more powerful tool to handle large collections of data extracted from texts. Clustering is our first approach to data visualization because it appears to be the natural choice for our particular dataset. Working with person type entities is desirable to find out classes and groups reflecting those in the real world. With simple tools we have a way to visualize data and help scholars to manage thousands of entities and relationships, to identify cluster of people and highlight who of them is more important for his role in the small mafia sentences society. We intend to continue our study on information visualization adopting methods and knowledge we have learned during this work.

Future work will focus on: testing and improving the suggested graph layout; testing the proposed technique on other text collections; compare the proposed technique with other published methods.

© The Eurographics Association 2011.

References

[BB79] BERSTEL J., BOASSON L.: Transductions and context-free languages. *Ed. Teubner* (1979), 1–278. 2

[CNM04] CLAUSET A., NEWMAN M. E. J., MOORE C.: Finding community structure in very large networks. *Phys. Rev. E 70* (Dec 2004), 066111. 3

[DGGZ10] DEFELICE D., GIUFFRIDA G., GIURA G., ZARBA C.: La descrizione dei reati di mafia nel testo delle sentenze. *Quaderni di sociologia LIV*, 3 (2010), 57–80. 2

[Fre77] FREEMAN L. C.: A Set of Measures of Centrality Based on Betweenness. *Sociometry 40*, 1 (Mar. 1977), 35–41. 3

[Hea95] HEARST M. A.: Tilebars: Visualization of term distribution information in full text information access, 1995. 1

[HSH*02] HAVRE S., SOCIETY I. C., HETZLER E., WHITNEY P., NOWELL L.: Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics 8* (2002), 9–20. 1

[Mil60] MILLER J.: Information input overload and psychopatology. *Am J Psychiatry 116*, 8 (1960), 695–704. 1

[Pal02] PALEY W. B.: Textarc: Showing word frequency and distribution in text. *Poster presented at IEEE Symposium on Information Visualization 2002* (2002). 1

[SB03] SHNEIDERMAN B., BEDERSON B. B.: *The Craft of Information Visualization: Readings and Reflections*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003. 1

[SSMF*09] SMITH M. A., SHNEIDERMAN B., MILIC-FRAYLING N., MENDES RODRIGUES E., BARASH V., DUNNE C., CAPONE T., PERER A., GLEAVE E.: Analyzing (social media) networks with NodeXL. In *C&T '09: Proceedings of the fourth international conference on Communities and technologies* (New York, NY, USA, June 2009), C&T '09, ACM Press, pp. 255–264. 3

[Tve05] TVERSKY B.: Visuospatial reasoning. In *The Cambridge Handbook of Thinking and Reasoning*, Holyoak K., Morrison R., (Eds.). Cambridge University Press, 2005, pp. 209–240. 1

[War04] WARE C.: *Information visualization: perception for design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2004. 4

© The Eurographics Association 2011.