Understanding Large Legal Datasets through Visual Analytics

Erick Gomez-Nieto*, Wallace Casaca*, Ivar Hartmann†, Luis Gustavo Nonato*
* ICMC, University of São Paulo, São Carlos, Brazil
† Law School, Getúlio Vargas Foundation, Rio de Janeiro, Brazil

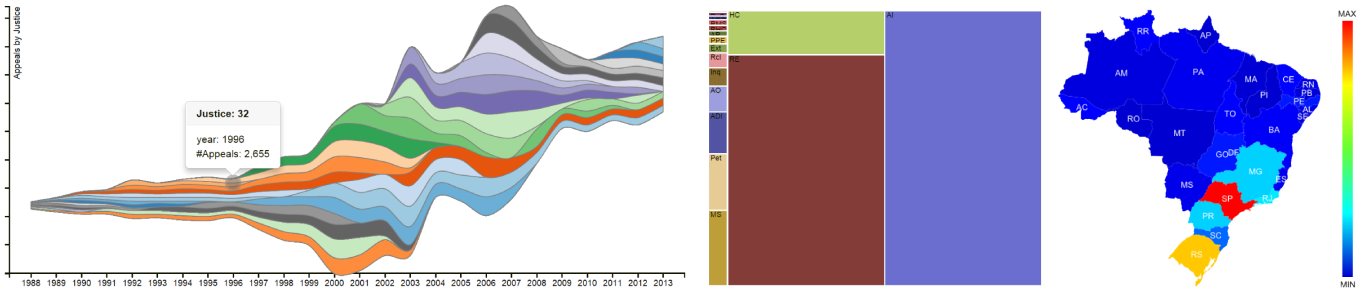

Fig. 1. Our approach showing in linked-views all the appeals from Brazilian Supreme Court database grouped by justice during 1988-2013: (left) we use a streamgraph to display a global view enabling users to select an arbitrary point of time for a detailed analysis, (middle) a treemap summarizes the type of appeals reviewed in the year and by the justice selected, and (right) a geographical heatmap illustrating how the frequency of a specific appeal type from the treemap, in this case "AI" type, is distributed across the Brazilian nation.

***Abstract*—Databases containing more than one million of legal documents, each with dozens of variables, pose special issues as to the detection of patterns of interest for judges and prosecutors. Most state-of-the-art methods rely on automatic schemes to classify/group data according to their similarity in the hope of uncover useful information, neglecting the knowledge and skill of specialists in the information extraction process. In this work we propose an visual analytics tool made up of a set of linked-views to explore 25 years of data from the Brazilian Supreme Court, all with the purpose of extracting information that traditional methods could not reveal.**

***Keywords*-Legal Data; Brazilian Supreme Court; Visual Analytics; Information Visualization;**

I. Introduction

One of the novelties that the advances in information technology have brought to law in the last decade is the quick and facilitated access to extremely large databases of legal documents. Metadata concerning such documents is likewise easily available. By "extremely large databases" we mean those in excess of one million documents, preventing individual analysis by humans. Furthermore, it is not uncommon for such collections of contracts, petitions or court rulings to contain upwards of twenty variables. Combined with the infinite possibilities of associated documents, this raises a fundamental problem that precedes any attempt at testing a hypothesis, running a regression or designing a model: how to know where to look for a pattern, trend or relationship?

The effectiveness of visual mechanism for exploring and analyzing data in different fields of study has been demonstrated. Although superficially addressed, visualization on laws is no stranger to this development. Some examples are judicial infographics [1], legal documents and opinion inspection tools [2], and browsers for enacted laws [3], which evidence the potential of using visual analytics for understanding concisely legal data.

One of the key problems that the eleven Brazilian Supreme Court Justices have to solve is precisely combing through dozens of thousands of appeals each year in order to decide which ones to admit for constitutional review. The number of appeals reached nearly 130.000 in 2006, but has since decreased to approximately 70.000 [4]. Empirical evidence shows that the filtering process started in the last few years has been unsuccessful, allowing thousands of repetitive appeals by a small group of parties to be granted certiorari [5].

Contributions: This paper proposes a new approach to support interactive analysis and visualization of 25 years worth of rulings from the Brazilian Supreme Court and respective metadata. The main purpose of finding patterns and discovering hidden information is to enable the Supreme Court Justices to visually inspect the case trends and comb faster through the new appeals as they reach to it.

II. Related Works

The task of processing legal data in the search for patterns [6], [7] – which might include clustering techniques [8], [9] – has been previously taken up for different purposes and contexts. Furquim et al. [10] use sophisticated machine learning techniques to categorize a set of three-years of Brazilian legal documents taken from the Federal Tribunal of one of five Brazilian Regions. However, as most automatic mechanisms

for pattern discovery in this context [11], [12], user knowledge is not consider during the exploratory process.

Another approach widely used for exploring legal data is network analysis. For instance, Bommarito et al. [13] use an acyclic digraph scheme to construct a case-to-case network on two-dimensional space. The data comprises the first quarter century decisions from the United States Supreme Court. Recently, Bommarito et al. [14] have extended the graph representation to 3D space, including an incremental mechanism to update dynamically time-varying data. Boulet et al. [15] rely on a clustered graph to visualize the ratification of Multilateral Environmental Agreements (MEA) between different countries during the last 35 years. Lawvis [16] is an exploration tool for visualizing heterogeneous relevant legal documents connected to a specific piece of Italian legislation, allowing to perform queries inside different data sets. However, even if networks are an effective mechanism to analyze legal information, there are some problems when a large number of edges connecting vertices are displayed, rendering the visualization as a dense and unreadable map.

III. The Brazilian Supreme Court Data Set

The Brazilian Supreme Court acts as a Constitutional Court, exercising both concentrated and diffuse review of constitutional aspects involved in appeals originating from lower courts in the country. However, because the Brazilian Constitution is very comprehensive, almost any lawsuit can be made to contain a constitutional issue. This, coupled with very generous criteria for appealing to the Supreme Court, has caused a monstrous number of new cases to enter the dockets of the eleven Justices each year.

We have access to a database of Supreme Court cases produced by the Law School of the Fundação Getúlio Vargas in Rio de Janeiro. The data is offered as part of an academic cooperation agreement with the Court and it includes information on the parties of the case, where the appeal originated, the legal subject of the case, all the docket entries, dates such as when the case entered the Court's docket or when it was tried, number of pages, textual content of the rulings and other similar variables. All in all, the database contains information of 1.524.060 cases (as of December 31st, 2013), which includes 14.985.497 docket entries with free text. Text files with preliminary, intermediary and merit rulings are also included, totaling beyond 2 million files in txt, rtf and pdf formats.

The rulings can be compared by matching key expressions, the citation of the precedents and laws, etc. We use those metadata as well as free texts and associate attributes as basis for comparing lawsuits.

IV. Linked-views and visual resources

In this section, we describe the proposed visual resources employed by our visualization system to explore the Brazilian Supreme Court data.

A. Stacked Graphs

Stacked graphs are visual mechanisms typically used to simultaneously visualize multiple levels of data categories over time. In this type of visualization, data representatives are ordered and heaped in a layered graph from which on can visualize the distribution of accumulated values over time.

In [17] the authors designed a stacked-graph based visual metaphor called *ThemeRiver*, which creates a set of wave-like layers corresponding to time series of the exploited data. *Streamgraph* [18] introduces significant improvements by optimizing the resulting layout, resulting in smother and more compact graphs. We make use of Streamgraph to depicted a very large collect of legal documents. These documents are organized in a timeline graph where colors reflect the Brazilian states/justices and the vertical-axis illustrates the number of appeals from each state/justice.

B. Treemaps

Shneiderman [19] introduces the notion of treemap-based visualization, in which hierarchical data represented by tree structures are progressively subdivided into geometric primitives and then placed in a given display area. Techniques devoted to build treemaps differ as to the region subdivision criterion as well as the nature of the graphical representation [20], [21].

In our exploratory analysis, we use the squarified treemap mechanism [22] implemented in D3.js library[1]. The treemap is used to narrow down the visualization of particular streamgraph points of time, creating a linked mechanism of visual data exploration. The linked view allows users to read and interpret state/justice streamgraphs, allowing the selection of particular data points for further exploration through the treemap layout, as shown in Fig. 2.

C. Heatmaps

A heatmap [23] is a visual abstraction where colors are used to convey data variation. A readable heatmap provides a global visual summary of information, allowing users to identify salient spots of color that correspond to particular data values. A survey that addresses intrinsic aspects of heat maps such as mathematical issues and their context of application can be found in [23].

In our implementation, we generate a geographical heatmap of Brazil states with the purpose of analyzing the distribution of specific type of appeals across the country. A particular appeal is picked out by the user and its occurrence around the Brazilian states are color mapped in the Brazil map. The advantage of using color map is that users can detect patterns by simply inspecting the variation in color levels, as shown in Fig. 3.

D. Similarity grid

Recently, the use of multidimensional projection to visual data analysis has received strong interest. Some important

[1] Available at http://www.d3js.org/

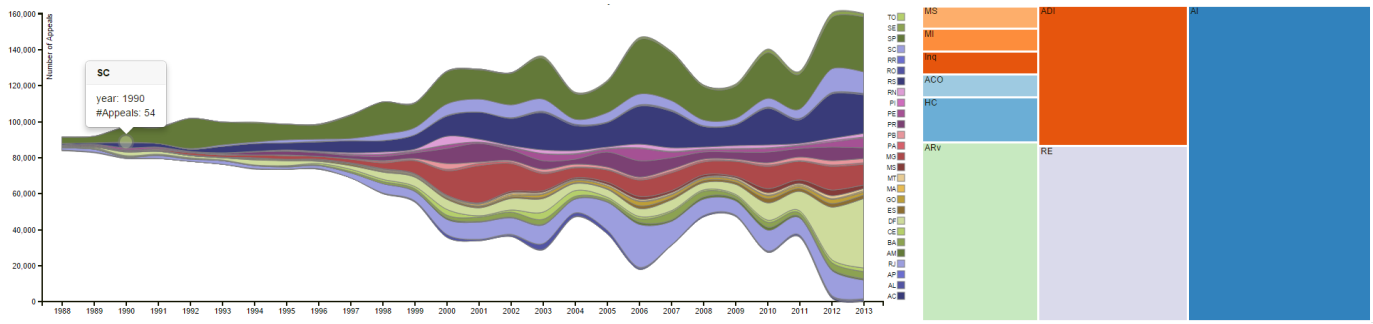

Fig. 2. Our approach depicts in linked-views all the appeals from the Brazilian Supreme Court database grouped by states during 1988-2013. (Left) An arbitrary point of the streamgraph is selected, which indicates the state of Santa Catarina (SC) in 1990. (Right) Once the user selects a point on the streamgraph, a treemap illustrates the frequency for each type of appeal w.r.t. the chosen point.

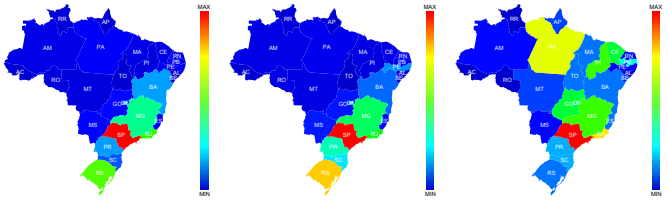

Fig. 3. Three different geographical heatmaps obtained from querying the same justice in 2007 for: (left) *"Agravo de Instrumento"*, (middle) *"Recurso Extraordinário"* and (right) *"Reclamação"*.

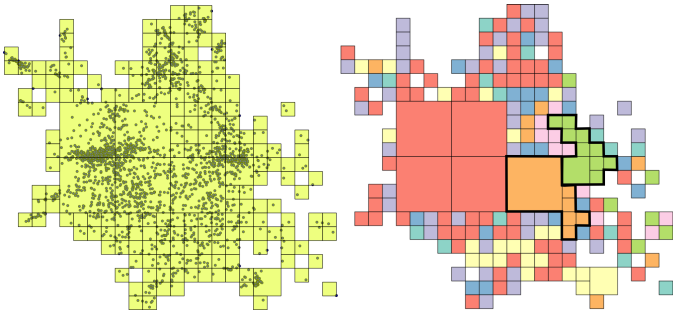

Fig. 4. Showing 2412 cases from 2008. Lawsuits are mapped to the visual space based on the text of the rulings. Highlighted green region depict lawsuits with the same decision dates. Orange region also correspond to lawsuits with the same decision date.

applications of multidimensional projection are semantic-preserving layout arrangement [24], [25], space-filling strategies [26] or scientific data exploration [27] In this work, we use LSP projection method [28] to map each instance (lawsuits) on a 2D visual space. We choose LSP due to its good performance in terms of distance preservation and low computational cost, however, any other method with similar properties can be used. We further process the layout resulting from LSP using a density-based grid, which allows to limit the similarity search space among instances. The process of refinement is carried out by the density of points inside a region, similar to a unbalanced quad-tree. Users can interactively defines a number of equal size cells for the first level of the grid. Then, the largest density value d_{max} is calculated and finer levels generated from the following conditions: if density in the cell d_{ij} is lower than $\frac{d_{max}}{k_1}$ then refine it one level, if density in one of these new refined cells d^*_{ij} is lower than $\frac{d_{max}}{k_2}$ then refine one level more, and so on, where k_m is the m-st value of the serie $k = \{2^1, 2^2, 2^3, ..., 2^n\}$. Fig. 4 illustrates the refinement process using three levels.

V. Results and Discussion

In this section, we discuss obtained results from the proposed methodology. All the results were produced using an Intel(R) Core(TM) i7-3537U processor with 8GB RAM and Google Chrome browser.

Fig. 1 shows an overview of our methodology. In this case, the entire database has been grouped according the judge who reviewed the appeal. To exemplify this task, we select the judge labeled as 32 and the year 1996 in the streamgraph, retrieving a total of 2655 appeals reviewed during the selected year by the selected judge. The treemap illustrates how the 2655 appeals are distributed according their type, being the largest number of cases of the type *"Agravo de Instrumento"* (AI), followed by *"Recurso Extraordinário"* (RE) and *"Habeas Corpus"* (HC), respectively. Moreover, the geographical heatmap provides an intuitive vision of the distribution of a particular type of appeal (AI) over the country. The type AI was chosen for being the most frequent one in São Paulo state, followed by Rio Grande do Sul (RS), Rio de Janeiro (RJ) and Minas Gerais (MG), respectively. As it can be notice, the enabled interactive mechanism allows us to obtain an intuitive and clear analysis of how the appeals were reviewed, and examine the performance of each judge during in the last 2 decades.

A similar example is presented in Fig. 2, where the streamgraph describe the increment the appeals for each state by year. A simple exploration shows that Brasília (DF), São Paulo and Rio Grande do Sul are the states with the largest number of reviewed by the Supreme Court. In addition, the chart reveals that São Paulo presents a more uniform trend since 1990-1991, compared to Brasília and Rio de Janeiro, which were

dramatically increased in 2012 and 2006, respectively. The treemap (on the right side) point out the distribution for Santa Catarina State (SC) in 1990, where most appeals are of type (AI) and (RE).

Fig. 4 shows how the similarity map groups a sample of 2412 cases according to the similarity in the text of their respective rulings (on the left). Each case ruling is a dot and the background layout merely frames the arrangement produced by the text similarity. The tool then repeated the same grouping but displayed information on the date when the case was decided. This revealed the characteristics and extent of the overlap between opinion text and the date when the ruling was issued. It thus indicates how well the two variables could be used to identify similar cases.

VI. Conclusion, Limitation and Future Work

In this work we present a methodology that combines a global view of time-varying legal data using linked-views and a detailed view mapped by similarity using a multidimensional properties allowing the appeal exploration from Brazilian Supreme Court database. Although our methodology is focused on the Brazilian Law, it can be used for any other legal dataset.

In our implementation, treemaps represent data structurally improving the hierarchical aspect of the data categories. However, in some cases – e.g. displaying a few number of hierarchy levels – a lot of empty space is unused, it can be leveraged by another visual resource.

In the future, we plan enrich the exploration experience by adding some components. First, in order to understand differences among two or more cases on all their steps across the review process, a time-varying visualization to show the order of relevant events should be implemented. Second, as previously commented, a network is an valuable resource to analyze data, we think on generate a similarity aware network to represent parties/documents involved to improve our analysis.

Acknowledgment

The authors acknowledge the financial support from FAPESP (#2011/22749-8, #2013/00191-0 and #2014/16857-0) and CNPq (#302643/2013-3).

References

[1] "Sketchlex: infographies juridiques." [Online]. Available: http://www.sketchlex.com

[2] "Ravel." [Online]. Available: https://www.ravellaw.com

[3] "Le fabrique de la loi." [Online]. Available: http://www.lafabriquedelaloi.fr/

[4] J. Falcao, P. Cerdeira, and D. W. Arguelhes, "I relatorio supremo em numeros," *FGV Direito Rio*, 2011.

[5] J. Falcao and I. A. M. Hartmann, "Acesso ao supremo: Quando os recursos são parte do problema," *Revista Dialogos sobre Justica*, 2014.

[6] H. Park, J. Yoon, and K. Kim, "Identifying patent infringement using sao based semantic technological similarities," *Scientometrics*, vol. 90, no. 2, pp. 515–529, 2012.

[7] C. Holton, "Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem," *Decision Support Systems*, vol. 46, no. 4, pp. 853 – 864, 2009, {IT} Decisions in Organizations.

[8] Q. Lu, J. G. Conrad, K. Al-Kofahi, and W. Keenan, "Legal document clustering with built-in topic segmentation," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, ser. CIKM '11. New York, NY, USA: ACM, 2011, pp. 383–392.

[9] C. L. Boyd, D. A. Hoffman, Z. Obradovic, and K. Ristovski, "Building a taxonomy of litigation: Clusters of causes of action in federal complaints," *Journal of Empirical Legal Studies*, vol. 10, no. 2, pp. 253–287, 2013.

[10] L. de Colla Furquim and V. de Lima, "Clustering and categorization of brazilian portuguese legal documents," in *Computational Processing of the Portuguese Language*, ser. Lecture Notes in Computer Science, H. Caseli, A. Villavicencio, A. Teixeira, and F. Perdigo, Eds. Springer Berlin Heidelberg, 2012, vol. 7243, pp. 272–283.

[11] D. M. Katz, M. J. B. II, and J. Blackman, "Predicting the behavior of the supreme court of the united states: A general approach," *CoRR*, vol. abs/1407.6333, 2014.

[12] R. Guimerà and M. Sales-Pardo, "Justice Blocks and Predictability of U.S. Supreme Court Votes," *PLoS ONE*, vol. 6, no. 11, pp. e27 188+, Nov. 2011.

[13] M. J. Bommarito, D. M. Katz, J. L. Zelner, and J. H. Fowler, "Distance measures for dynamic citation networks," *Physica A: Statistical Mechanics and its Applications*, vol. 389, no. 19, pp. 4201–4208, 2010.

[14] C. L. Studies, "3d-hi-def visualization of the time evolving citation network of the united states supreme court," 2014. [Online]. Available: http://computationallegalstudies.com/2014/01

[15] R. Boulet, A. F. Barros-Platiau, and P. Mazzega, "35 years of multilateral environmental agreements ratification: a network analysis," in *Proceeding of 2nd. International Workshop "Network Analysis in Law"*, December 2014.

[16] N. Lettieri, S. Faro, L. Vicidomini, and A. Altamura, "Nets of legal information connecting and displaying heterogeneous legal sources," in *Proceeding of 2nd. International Workshop "Network Analysis in Law"*, December 2014.

[17] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "Themeriver: visualizing thematic changes in large document collections," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 8, no. 1, pp. 9–20, Jan 2002.

[18] L. Byron and M. Wattenberg, "Stacked graphs – geometry & aesthetics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1245–1252, Nov. 2008.

[19] B. Shneiderman, "Tree visualization with tree-maps: 2-d space-filling approach," *ACM Trans. Graph.*, vol. 11, no. 1, pp. 92–99, Jan. 1992.

[20] B. B. Bederson, B. Shneiderman, and M. Wattenberg, "Ordered and quantum treemaps: Making effective use of 2d space to display hierarchies," *ACM Transactions on Graphics*, vol. 21, no. 4, p. 833854, 2002.

[21] K. Onak and A. Sidiropoulos, "Circular partitions with applications to visualization and embeddings," in *Proc. of the 24th ACM Symposium on Computational Geometry*, 2008, p. 2837.

[22] M. Bruls, K. Huizing, and J. van Wijk, "Squarified treemaps," in *In Proceedings of the Joint Eurographics and IEEE TCVG Symposium on Visualization*. Press, 1999, pp. 33–42.

[23] M. Friendly, "The history of the cluster heat map," *The American Statistician*, vol. 63, no. 2, 2009.

[24] E. Gomez-Nieto, F. S. Roman, P. Pagliosa, W. Casaca, E. S. Helou, M. C. F. de Oliveira, and L. G. Nonato, "Similarity preserving snippet-based visualization of web search results," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 457–470, 2014.

[25] E. Gomez-Nieto, W. Casaca, L. G. Nonato, and G. Taubin, "Mixed integer optimization for layout arrangement," in *Graphics, Patterns and Images (SIBGRAPI), 2013 26th SIBGRAPI Conference on*, 2013.

[26] F. Duarte, F. Sikansi, F. Fatore, S. Fadel, and F. Paulovich, "Nmap: A novel neighborhood preservation space-filling algorithm," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 20, no. 12, pp. 2063–2071, Dec 2014.

[27] J. Poco, A. Dasgupta, Y. Wei, W. Hargrove, C. Schwalm, R. Cook, E. Bertini, and C. Silva, "Similarityexplorer: A visual inter-comparison tool for multifaceted climate data," *Computer Graphics Forum*, vol. 33, no. 3, pp. 341–350, 2014.

[28] F. Paulovich, L. Nonato, R. Minghim, and H. Levkowitz, "Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 14, no. 3, pp. 564–575, 2008.