ANALYSIS OF LEGAL AND POLITICAL DATA

Aleks Jakulin
"Jožef Stefan" Institute
Jamova 39, SI-1000 Ljubljana, Slovenia
jakulin@acm.org

Abstract With the abundance of publicly available data registering the judgements at supreme courts and parliamentary votes, we can employ various data mining techniques to identify interesting patterns. For example, we can identify explicit and implicit voting blocs, which may or may not agree with official party affiliations. We can assess the political strength of those blocs. We can examine the vote of which particular senators is the most representative of the final outcome of the vote. We can employ text mining and visualization tools to cope with a large number of issues discussed in parliaments. While the paper primarily acts as a survey, it demonstrates the utility of several techniques that have not yet been used in the context of law.

Keywords: voting, statistical analysis, visualization, data mining

Introduction

The present paper will attempt to introduce the topic of computer-based statistical analysis in a legal and political context. There is plenty of data from these areas, and if it is appropriately structured into the form of a matrix or an ontology, it yields easily to automated analysis. With data analysis it is no longer necessary to rely on subjective interpretation of events or institutional characteristics. Instead, summaries are generated automatically and provide a quantitative basis for most kinds of study. Often these summaries are pictorial and aesthetically pleasing, a distinct evolutionary step since the days of textual descriptions of models.

Not just that the quantitative basis is less subjective, data mining algorithms are exhaustive in their analysis. Traditional statistical methods were based on postulating a hypothesis a priori. Afterwards, the data was collected independently of the hypothesis. Finally, the hypothesis was tested. This approach is referred to as confirmatory data analysis, because we attempt to confirm or refute the hypothesis. Data mining, instead, follows the concept of exploratory data analysis. There is no hypothesis set up in advance. Instead, interesting

claims are generated directly from the data. These claims may be confirmed or refuted through subsequent research.

Of course, automated analysis expects the data to be presented in a structured way. The structured representation also aids the problem of information retrieval, so analysis is not the only purpose. A considerable amount of research has been dedicated to proposing appropriate structures for a variety of phenomena. A particular structure is simply a real number. It is not natural, but instead we must think on how to map a particular mental property into a number. Another structure is a set of categories. For example, we collapse the whole variety of kinds of weather into $\{rain, sunny, cloudy\}$ as to be able to analyze it statistically. Not just that we need to think about structuring properties, we also need to identify objects that act as carriers of the properties. What do we attribute the category to? We can attribute numbers to objects, to time intervals, to events. Sometimes the objects are not independent, and we represent the dependencies with links. In summary, identifying structure in the world comes down to defining properties and objects.

This article is structured as follows. In Sect. 1 we will show a structure of the Spaeth database of the US Supreme Court. We will also mention how the database could be further enhanced through the use of ontologies. In Sect. 2 we will focus on the analysis of voting, and describe the fundamentals of the methodology as applied to the US Supreme Court. In Sect. 3 we will describe hypothesis testing in the legal context. In Sect. 4 we display the issues discussed in the US Senate with the use of the *TextGarden* toolkit[1].

1. Structure of Data

The US Supreme Court Judicial Database (Spaeth, 2005) contains a wealth of information about each issue that was discussed at the court between 1953 and present. Each case ('docket') is described with almost 250 variables, including the votes of each judge (majority vs. dissent), the type of the vote, a description of the parties involved, indication of an alteration of precedent, type of the issue, authority for decision, and so on. More specifically, there are the following variables:

- **Identification variables:** case citations, docket number, unit of analysis, number of records per unit of analysis.

- **Background variables:** manner in which the Court takes jurisdiction, administrative action preceding litigation, three-judge district court, origin of case, source of case, lower court disagreement, reason for granting cert, parties, disposition of case by court whose decision the Supreme Court reviewed, direction of the lower court's decision.

- **Chronological variables:** date of oral argument, reargument date, decision date, term of Court, chief justice, natural court.
- **Substantive variables:** legal provisions considered by the Court, multiple legal provisions, authority for decision, issue, issue areas, direction of decision, direction of decision based on dissent.
- **Outcome variables:** type of decision, disposition of case, unusual disposition, winning party, formal alteration of precedent, declarations of unconstitutionality, voting and opinion variables, the vote in the case, vote not clearly specified, the individual justice's votes, the individual justice's opinions, special opinions with which the individual justice's agreed, direction of the individual justiceŠs votes, majority and minority voting by justice, majority opinion assigner, majority opinion writer, minimum winning coalition.

The structure of the Spaeth database could be an interesting resource for developers of legal ontologies. Many of the above variables have an intricate internal structure: such as a detailed hierarchical taxonomy of issues or parties. For example, consider the following hierarchical taxonomy of issue topics from the First Amendment domain:

401 First Amendment, miscellaneous
*410 Speech
- 411 commercial speech
- 415 libel, defamation: defamation of public officials and public and private persons
- 416 libel, privacy: true and false light invasions of privacy

*420 Federal security
- 421 legislative investigations: concerning internal security
- 422 federal internal security legislation: Smith, Internal Security, and related federal statutes

430 loyalty oath or non-Communist affidavit
- 431 loyalty oath, bar applicants
- 432 loyalty oath, government employees
- 433 loyalty oath, political party
- 434 loyalty oath, teachers
- 435 security risks: denial of benefits or dismissal of employees for reasons other than failure to meet loyalty oath requirements

*440 "Governance"
- 441 conscientious objectors to military service
- 444 campaign spending: financing electoral costs

*440 Freedom of Expression
- 451 protest demonstrations: demonstrations and other forms of protest
- 455 free exercise of religion

*460 Religion
- 461 establishment of religion

- 462 parochiaid: government aid to religious schools, or religious requirements in public schools

*470 Obscenity

- 471 obscenity, state: including the regulation of sexually explicit material
- 472 obscenity, federal

Some of the codes do not exist in the Spaeth taxonomy, and they were marked with '*'. The reason why we list them is to strengthen the internal hierarchy which can be seen only through the numerical encoding. Several of the entries contain references to related concepts. For example, it is important to distinguish 444 (campaign spending) as a part of First Amendment category from 650 (corruption) as a part of the Economic Activity category. Such references anticipate potential misclassifications and provide additional guidance in a way that resembles differential diagnosis in medicine. Nevertheless, there is a unique classification of each issue.

In spite of the above complexity, each issue is a natural object or instance. For that reason, the Spaeth database is structured as a spreadsheet. Each column corresponds to a particular property, and each row to a particular docket. When a property is unknown for a particular docket, the value of the property is considered to be a missing value. This representation yields easily to statistical analysis.

2. Analysis of Voting

The Democratic Senator G. Miller of Georgia was not voting like other Democrats in the US Congress in the year 2003. Instead, his votes were typical for a Republican senator. Similarly, the Democratic Senators Kerry, Lieberman and Edwards did not vote very often: Senator Kerry only cast 195 votes out of 459, abstaining from voting in all others. These senators were Democratic presidential candidates at the preliminaries. All these findings were obtained through automated analysis of the US Senate roll calls. Roll call data does not appear only in political science. Instead, majority voting is a part of the judicial procedure at various courts, including the US Supreme Court or the European Court of Human Rights.

Roll call data is a formal record of the voting actions in a parliament. Because the voters are the representatives of the citizens, their actions and opinions should be fully transparent. But even if they are transparent, few people will make the effort of examining the results. For that reason, automated analysis can provide easy-to-understand summaries.

The study of voting data is not new. Many of the ideas in the field go back to (Rice, 1928), but in recent years the techniques of (Poole and Rosenthal, 2000) have been regularly used to analyze the situation wherever voting is used: in national parliaments, in United Nations, in the European Parliament and at supreme courts. Most of these methods attempt to create a *spatial model* where

Issue	Breaux (D-LA)	Frist (R-TN)	Kerry (D-MA)	Kyl (R-AZ)	Levin (D-MI)	McCain (R-AZ)	Miller (D-GA)	Voinovich (R-OH)	Outcome
To provide additional funds for certain homeland security measures.	Yea	Nay	NV	Nay	Yea	Nay	Nay	Nay	Amendment Rejected
To provide additional funding for innovative programs at the state and local level.	Nay	Yea	NV	Yea	Nay	Yea	Yea	Yea	Amendment Agreed to
To provide additional funding for education.	Yea	Nay	NV	Nay	Yea	Nay	Nay	Nay	Amendment Rejected
To provide agricultural assistance.	Yea	Yea	NV	Yea	Nay	Yea	Yea	Yea	Amendment Agreed to
To improve health care under the medicare and medicaid programs.	Nay	Nay	Yea	Nay	Yea	Nay	Nay	Nay	Motion Rejected
A bill to prohibit the procedure commonly known as partial-birth abortion.	Nay	Nay	NV	Nay	Yea	Nay	Nay	Nay	Motion Rejected
To redirect $1.214 trillion in revenues that would have been lost by implementing the President's entire tax cut agenda into a reserve fund to strengthen the Social Security trust funds over the long-term.	Yea	Yea	Nay	Yea	Nay	Yea	Yea	Yea	Motion to Table Agreed to
To prevent consideration of drilling in the Arctic National Wildlife Refuge in a fast-track budget reconciliation bill.	Nay	Nay	Yea	Nay	Yea	Yea	Nay	Nay	Amendment Agreed to

Figure 1. A small subset of votes and senators in the US Senate expressed as a spreadsheet.

each representative can be seen as having a particular *ideal point* on some ideological scale or space. There is also an easy-to-use and freely available software, VoteWorld. An alternative approach is to examine the correlations between individual voters, and then interpret the correlation as a measure of proximity. The proximity of the voter to the outcome can be considered a measure of how influential that voter is, under some assumptions.

The roll call data is normally also represented in a spreadsheet format. Each row of the spreadsheet corresponds to a particular issue voted upon. Each column corresponds to an individual voter. The majority vote can also be listed in a special column. Thus, the opinions of individual voters can be seen as properties of a particular issue. An example of the roll call table is shown in Fig. 1.

The vote of each justice is recorded in much detail in the Spaeth database. We distinguish the vote of the justice, his opinion and special opinion, and the writer of the majority opinion. Specifically, there can be the following outcomes of the vote: a) voted with majority or plurality, b) dissent, c) regular concurrence (agreement with the Court's opinion as well as its disposition), d) special concurrence (agreement with the Court's disposition but not its opinion), e) nonparticipation, f) judgment of the Court, g) dissent from a denial or dismissal of certiorari (literally and only such a dissent), or dissent from summary affirmation of an appeal, and h) jurisdictional dissent (disagreement with the Court's assertion of jurisdiction without addressing the merits, or without providing the parties oral argument).

We summarize this information by only distinguishing agreement (agreement, regular or special concurrence, or the judgement of the Court) and dis-

sent, and ignoring nonparticipation (nonparticipation, jurisdictional dissent, dissent from a denial or dismissal of certiorari). We thus obtain a property that can have two different values, or the value can be missing. Furthermore, we will focus on the period 1994-2005, as the only two justices have not been present throughout the whole period. For that reason, the database has already been used for similar analysis (Sirovich, 2003; Lawson et al., 2003).

To analyze the data, we will employ a part of the methodology of (Jakulin and Buntine, 2004). Considering two justices and ignoring the cases when at least one of them did not cast a vote, there can be four joint outcomes: (1) yy - both agreed with the majority, (2) nn - both dissented, (3) yn - the first justice agreed, the second dissented, and (4) ny - just the opposite. We will use the count $\#nn$ to indicate the number of roll calls with outcome nn, while the sum of counts for all four outcomes is N. We do not include the roll calls where either of the justices did not participate: their relationship cannot be analyzed in such a case.

There are two basic probabilistic models that describe the voting process of two justices. In the first we assume that the justices are not voting independently, either because of similar judgement, similar opinion or an explicit agreement. As an example, the probability of outcome nn in the correlation-assuming model is estimated as $p_{nn} = \#(nn)/N$. The second model assumes that the votes of both justices are independent. The probability of a joint outcome nn, p_{nn} is therewith a product of the probability that the first justice voted n, $p_{n*} = p_{nn} + p_{ny}$, and the probability that the second justice voted n, $p_{*n} = p_{nn} + p_{yn}$. The correlation-assuming model predicts the probability of the joint outcome nn as $\pi_{nn} = p_{nn}$, while the one assuming no correlation as $\phi_{nn} = p_{n*}p_{*n}$.

The difference between the two models quantifies the amount of correlation between justices X and Y. We compute it with the following formula for mutual information:

$$I(X;Y) := D(\pi||\phi) = \sum_{x\in\{n,y\}} \sum_{y\in\{n,y\}} \pi_{xy} \log_2 \frac{\pi_{xy}}{\phi_{xy}}. \tag{1}$$

We control for individual justice's propensity to agree or dissent, which does not affect the correlation. We furthermore transform the mutual information into Rajski's metric (Rajski, 1961) (also known as de Mántaras' distance (López de Mántaras, 1991)):

$$d(X,Y) := 1 + \frac{I(X;Y)}{\sum_{xy} \pi_{xy} \log_2 \pi_{xy}} \tag{2}$$

To summarize the dependencies between justices' votes we employ agglomerative hierarchical clustering (Kaufman and Rousseeuw, 1990) to summarize the dissimilarity matrix composed of inter-justice Rajski's distances.

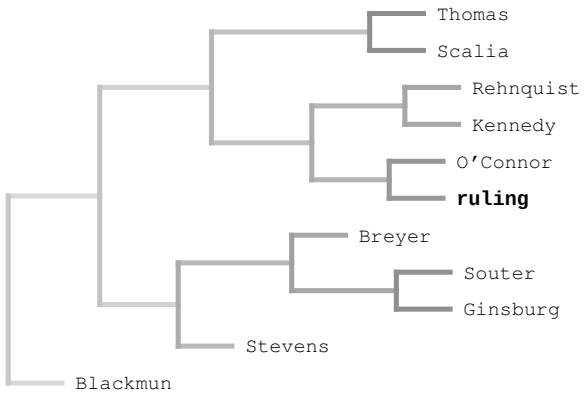

Figure 2. In this dendrogram, the justices that tend to agree more appear closer in the tree structure. Offset to the right and the darkness of the connecting lines is indicatory of higher mutual information between two justices' votes.

The result is shown in Fig. 2, and identifies two blocs of justices. Our method yields the roughly same result as did the SVD-based analysis by (Sirovich, 2003), and the noncommutative harmonic analysis of (Lawson et al., 2003): there is the majority bloc of Thomas, Scalia, O'Connor, Rehnquist and Kennedy, along with the final court's opinion; the minority bloc is formed by Ginsburg, Stevens, Breyer (previously Blackmun) and Souter. Generally, the majority bloc tends to be conservative and affiliated with the Republican party in the Senate, whereas the minority bloc is closer to the Democratic party.

Another approach to visualization the similarities between justices is the family of spatial models, frequently used in political science. They normally postulate a model of rational decision making. Each justice is modelled as a position or an *ideal point* in a spatial model of preferences (e.g., (Davis et al., 1970)). We have employed the binary PCA model (de Leeuw, 2003), and the results are shown in Fig. 3. The votes for a particular docket would be explained by a line separating one set of justices from the other. The vast majority of votes would be represented with horizontal lines: there were only a small number of votes that separated the more moderate from the more extreme justices.

It is also possible to perform the traditional kind of statistical analysis. For example, each vote of the court is classified as having a liberal or conservative direction (variable DIR). The direction is considered to be liberal when when the court decided in favor of an individual in disputes between an organization or an individual, or in favor of the federal government in disputes between state and federal government. The amount of correlation between a justice and the outcome can be interpreted as the amount of influence a justice exerts

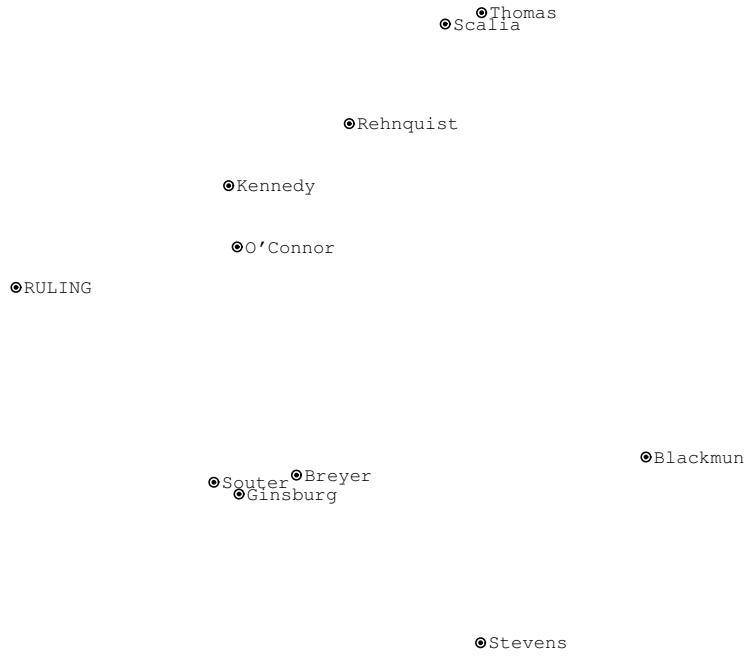

Figure 3. The binary PCA model shows the ideal points of justices that best explain the resulting votes. The similarities between individual justices correspond to the euclidean distances between the corresponding bull-eyes: they parallel the results of Fig. 2. The arrangement is rather simple, and a one-dimensional spatial model would be sufficient to capture the nature of the votes.

on the outcome; this is controversial, but if we want to attach a number to this characteristic, there is no other way: all we can infer from the data are correlations after all. The particular quantification we use is the ratio between the mutual information between a justice and the outcome, and the Shannon entropy of the outcome. But a related measure is simply the probability of dissent. The results are summarized in Table 1.

It is interesting to notice that the probability of dissent is closely related to the influence. For that reason, the probability of dissent is a more intuitive measure of the same concept. Also, moderate justices from either side were the most influential. Nevertheless, the conservative justices exerted more power on the outcome in this period, overall.

Table 1. The justices are sorted on the Liberal-Conservative dimension. It is also shown how much influence they exert on the outcome, along with the probability of dissent and the proportion of dockets they did not express their opinion on. All numbers in the table are percentages.

Justice	Liberal Votes	Influence	Prob. of Dissent	Absence
Thomas	24.8	21.1	25.6	1.5
Scalia	26.2	24.3	23.6	1.2
Rehnquist	28.8	34.4	18.5	0.7
Kennedy	36.2	48.0	12.2	0.5
O'Connor	38.8	53.8	11.9	2.5
Breyer	56.8	30.8	22.8	2.2
Souter	58.1	41.9	17.7	0.4
Ginsburg	58.9	37.4	19.9	0.7
Stevens	66.8	21.3	30.2	1.6
Blackmun	71.2	13.7	39.2	2.3

3. Analysis of Interactions

In the previous section we analyzed the voting behavior of US Supreme Court justices. Each justice is described with a variable. Our objective was to organize a large number of variables, summarize their similarities and basic characteristics. The goal of the present section is not summarization of the correlations among many variables but instead examination of a single correlation in more detail.

Correlation as a concept has several meanings. Most often, correlation is considered to correspond to Pearson's correlation coefficient. The notion of an interaction generalizes upon the notion of a correlation, and we will use the term do distinguish ourselves from the assumptions and limitations of correlation, its Gaussian and linear nature. An interaction can be considered to be any kind of a dependency between two variables. Correlation is only a specific example of an interaction.

For example, we could wonder whether the time interacts with the ideological bias of a justice. If it does not interact, the ideological bias remains more or less the same throughout the whole period. If it instead interacts, the ideological bias will vary. This particular interaction has been addressed by several works, such as (Martin et al., 2005; Bafumi et al., 2005). They are of the type we examined earlier - they attempt to infer the ideological dimension as to explain the differences and correlation between individual justice votes. We will follow a different approach and make use of the DIR variable in the Spaeth database, computing an average for each year. The result is shown in Fig. 4. As it has been noticed before (Martin and Quinn, 2002), the justices' ideology does change with time.

In summary, most scatter or line plot visualizations intend to convey the nature of a particular interaction involving the variables corresponding to the x and y axes. All other variables are ignored. However, it is unclear whether the shifts in ideology can truly be attributed to justices' ideology: they could simply be explained through shifts in the characteristics of the dockets. In some sense the chart ignores all related temporal shifts.

4. Analysis of Text

The Library of Congress in Washington maintains the THOMAS database of legislative information. One type of data are the senate roll calls [2]. Every year, there are approximately 400 issues discussed in the US Senate. Over a couple of years, there is a staggering amount of data, and it is difficult to examine it manually. Manual categorization is quite time consuming, but we can employ automated methods for organizing the issues, based on the text of the law, nomination, etc., time, and so on. We then employ visualization techniques to present the issues, and software that allows exploring the database interactively.

For our experiment, we have taken 2700 issues, roughly from the period between Congresses 104 and 108 (roughly 7 years altogether). We then employ the visualization techniques, based on latent semantic indexing (LSI) (Deerwester et al., 1990), and followed by multidimensional scaling (MDS) (Borg and Groenen, 1997), as implemented in the TextGarden system. The main benefit of the Document Atlas approach is to squash a high-dimensional system of factors into a convenient two-dimensional form that can be examined interactively.

Each issue is represented with a yellow cross-mark in the diagram. The characteristic words insinuate the 'meaning' of issues in a particular area of the landscape. The lighter hues of blue indicate areas of high document density. The user interface allows the user to move a 'lens' around the landscape, while giving him the ability to adjust the scope (larger, smaller). Next to the

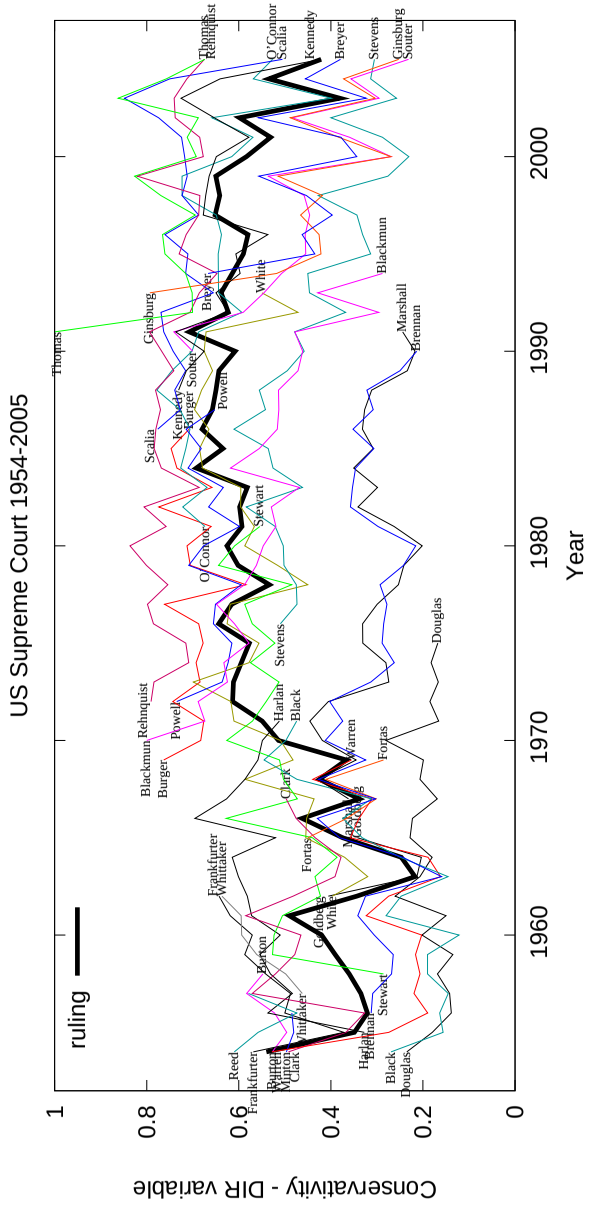

Figure 4. A chart of the yearly averages of the DIR variable for the ruling of the Court, and of each individual justice shows that the ideology is rather dynamic. We can conclude that ideology does interact with time, and the chart shows how.

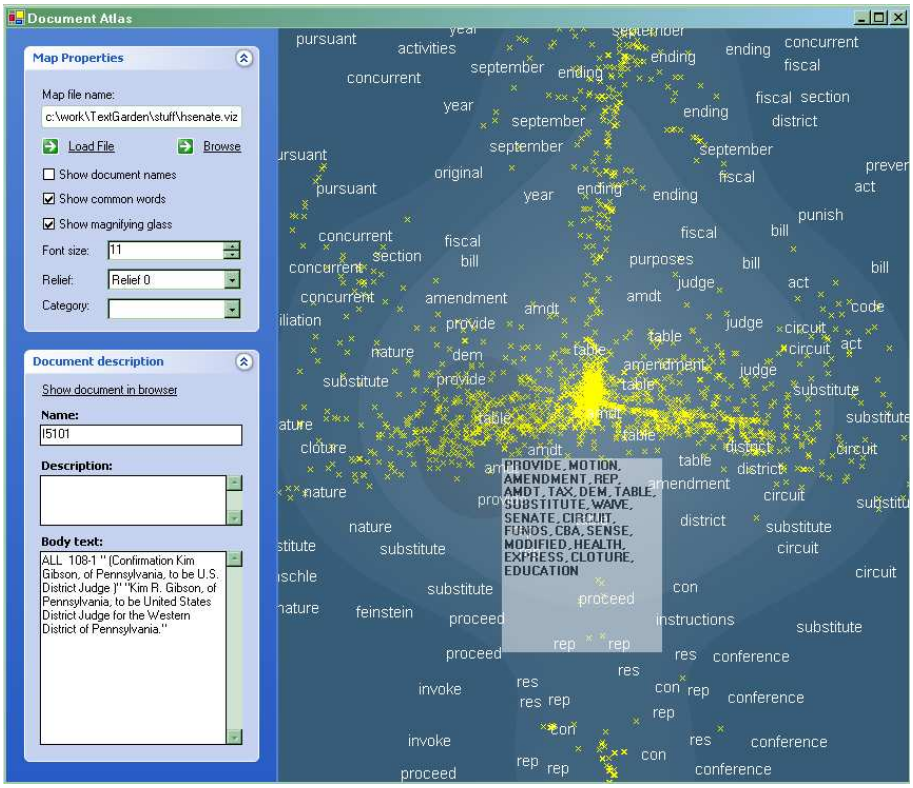

Figure 5. More than 2700 issues discussed in the US Senate are displayed in this Document Atlas diagram. In the bottom, there are various conference reports. To the left, there are various procedural issues. To the right, there are elections of judges. To the top, there are the budgetary issues. In the center, there are many bills that do not fit this scheme, but can be analyzed separately.

transparent lens, there is a semi-transparent box with a more detailed list of keywords. Furthermore, after a click, the closest issue is displayed in the tab to the left of the window.

The specific issue shown in Fig. 5 is an example of a notable battle between Democrats and Republicans in the 108th Congress: the Republican majority tried to elect a number of conservative judges, and the Democratic minority used the device of filibuster to block the attempt. Namely, these conservative judges would have influenced the judicial branch of government for a long time.

5. Conclusion

We hope that these simple examples have shown the power of automated analysis of legal and political data. Furthermore, we have shown the wealth of

information that can be found in the publicly available databases available on the World Wide Web. There are numerous observations that have been made: we can identify disputations in the parliaments, we can identify outliers, we can infer the blocs and their voting power.

The development of ontologies can also be inspired partly from how these databases are structured. Although most of them are simple spreadsheets, it is possible to perform sophisticated types of analysis. In fact, there are very few tools that would be able to work with data that is not a spreadsheet. For that reason, it might be preferable to pick an evolutionary approach of starting with spreadsheet-like data models, and later evolving to a more structured and intertwined representation of data.

There are distinct benefits to providing full and open access to the information of such kind. Except in the areas of national security, there are distinct dangers of having a non-transparent government, even if it is democratically elected. We suggest that the development of practical databases of such kind, combined with analytical tools should be a priority for the future.

Acknowledgments

This work was supported by the Slovenian Research Agency and the IST Programme of the European Community (Semantically Enabled Knowledge Technologies (SEKT) IST-1-506826-IP). The author wishes to thank Blaz Fortuna for his Document Atlas software, and Andrew Gelman for helpful comments to Fig. 4.

Notes

1. `http://www.textmining.net/`
2. Available online at `http://thomas.loc.gov/home/rollcallvotes.html`

References

Bafumi, J., Gelman, A., Park, D. K., and Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, 13(2):171–187.

Borg, I. and Groenen, P. (1997). *Modern Multidimensional Scaling*. Springer-Verlag, New York.

Davis, O. A., Hinich, M. J., and Ordeshook, P. C. (1970). An expository development of a mathematical model of the electoral process. *American Political Science Review*, 64:426–448.

de Leeuw, J. (2003). Principal component analysis of binary data: Applications to roll-call-analysis. Technical Report 364, UCLA Department of Statistics. `http://preprints.stat.ucla.edu/download.php?paper=364`.

Deerwester, S., Dumais, S., Furnas, G., Landuer, T., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*.

Jakulin, A. and Buntine, W. (2004). Analyzing the US Senate in 2003: Similarities, networks, clusters and blocs. Working paper. `http://kt.ijs.si/aleks/politics/`.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.

Lawson, B. L., Orrison, M. E., and Uminsky, D. T. (2003). Noncommutative harmonic analysis of voting in committees. `http://homepages.uc.edu/~lawsonb/research/noncommutative.pdf`.

López de Mántaras, R. (1991). A distance based attribute selection measure for decision tree induction. *Machine Learning*, 6(1):81–92.

Martin, A. D. and Quinn, K. M. (2002). Dynamic ideal point estimation via Markov chain Monte Carlo for the U.S. Supreme Court, 1953-1999. *Political Analysis*, 10:134–153.

Martin, A. D., Quinn, K. M., and Epstein, L. (2005). The median justice on the U.S. Supreme Court. *North Carolina Law Review*, 83:1275–1321.

Poole, K. T. and Rosenthal, H. (2000). *Congress: A Political-Economic History of Roll Call Voting*. Oxford Univ. Press.

Rajski, C. (1961). A metric space of discrete probability distributions. *Information and Control*, 4:373–377.

Rice, S. A. (1928). *Quantitative Methods in Politics*. Knopf, New York.

Sirovich, L. (2003). A pattern analysis of the second Rehnquist U.S. Supreme Court. *PNAS*, 100(13):7432–7437.

Spaeth, H. J. (2005). The original United States Supreme Court judicial database 1953-2003 terms. Distributed by the S. Sidney Ulmer Project for Research in Law and Judicial Politics.