Analyzing Parliamentary Elections Based on Voting Advice Application Data

Jaakko Talonen and Mika Sulkava

Aalto University School of Science,
Department of Information and Computer Science
P.O. Box 15400 FI-00076 Aalto, Finland
talonen@cis.hut.fi, mika.sulkava@tkk.fi
http://www.cis.hut.fi/talonen

Abstract. The main goal of this paper is to model the values of Finnish citizens and the members of the parliament. To achieve this goal, two databases are combined: voting advice application data and the results of the parliamentary elections in 2011. First, the data is converted to a high-dimension space. Then, it is projected to two principal components. The projection allows us to visualize the main differences between the parties. The value grids are produced with a kernel density estimation method without explicitly using the questions of the voting advice application. However, we find meaningful interpretations for the axes in the visualizations with the analyzed data. Subsequently, all candidate value grids are weighted by the results of the parliamentary elections. The result can be interpreted as a distribution grid for Finnish voters' values.

Keywords: Parliamentary Elections, Visualizations, Principal Component Analysis, Kernel Density Estimation, Missing Value Imputation.

1 Introduction

The parliamentary elections in Finland were held on 17 April 2011. In the elections, 200 members of parliament (MP) were elected for a four-year term. Every Finnish citizen, who has reached the age of 18 not later than on the day of the elections, is entitled to vote. The number of people entitled to vote was 4 387 701 and the voting turnout was 70.5% [1].

Helsingin Sanomat (HS), the biggest newspaper in Finland, published the voting advice application (VAA) [2] about a month before the elections. Questions from different topics were available for the candidates to answer in advance. VAA provides a channel for candidates to tell their opinions to various topical issues. Voters express their views on these issues. The output of VAA is a ranked list of candidates. On 6 April 2011 HS released the data of the candidates' answers [3].

The aim of this research is to model the opinions, ideals and values (later just values in this paper) of Finnish citizens and the MP. Two databases were combined: the VAA data provided by HS [3] and the results of the parliamentary

J. Gama, E. Bradley, and J. Hollmén (Eds.): IDA 2011, LNCS 7014, pp. 340–351, 2011.
© Springer-Verlag Berlin Heidelberg 2011

elections 2011 [1]. These data are explained in more detail in the Section 2.The methodology is presented in Section 3. In Section 4 the data is preprocessed and converted to a high-dimension space. Then it is reduced to two dimension by Principal Component Analysis (PCA). Next, the value grids are weighted by the results of the parliamentary elections 2011 [1]. The value grid is produced with a kernel density estimation method without explicitly using the questions of the VAA. Thus, the visualization results presented in this paper do not depend on what was asked and which answer options were available. As a result a distribution grid for Finnish voters' values is presented. Finally, more visualizations and possibilities for use are introduced. A discussion in Section 5 concludes this paper.

2 Voting Advice Application Data and Parties in Finland

Questions with a large amount of factual information were selected for the VAA [2]. These questions were divided into nine subgroups according to the topic, which were: (general) questions 1–5, pensions 6–8, economy 9–11, taxes 12–15, defense 16–17, foreign countries 18–21, domestic 22–26, localities 27–30. In question 31 three parties were asked to select for desired government. In question 21 it was asked that if Finland was in Facebook, which three countries should be her friends. So total 35 answers were assumed for each candidate. The list of registered parties is presented in Table 1.

Long and precise questions enhance the reliability of the answers. It can be assumed that candidates use more time with the questionary than, e.g., citizens in gallups. For example, one of the shortest questions was "Differences in income have increased very fast in Finland after mid-1990s. How should it be approached?". The answer options were

- The differences in the incomes should be reduced.
- The differences in the incomes have to be reduced slightly.
- The disparity of incomes is now at an appropriate level.
- Income inequalities can increase moderately.
- Income inequalities should grow freely.

In the VAA it was also possible to give weights for the questions depending how important the questions were. The options were small, medium and the great importance. A typical way to convert qualitative data to quantitative is to give a numerical value for each answer in each question [4]. Surely, some results can be obtained by analysing this type of data, but the question is that can we really trust the results. Should "The differences in the incomes should be reduced." be quantized as one and "The differences in the incomes have to be reduced slightly." as two or maybe three? In traditional analysis, the question importances and multiple choice questions are usually omitted. In this paper, a solution for proper VAA data preprocessing is introduced.

Table 1. The list of registered parties and approximation for party type *(left/right) [5,6]. Parties are listed by the results of the Parliamentary elections 2011 (given seats). Names in English are unofficial translations [1]. **a new member of parliament from Åland Coalition is added to RKP.

Abbreviation	Party	Seats-07	Seats-11	left/right*
KOK	National Coalition Party	50	44	7.2
SDP	The Finnish Social Democratic Party	45	42	3.5
PS	True Finns	5	39	6.4
KESK	Centre Party of Finland	51	35	5.8
VAS	Left-Wing Alliance	17	14	2.4
VIHR	Green League	15	10	3.7
RKP	Swedish People's Party in Finland	10**	10**	6.3
KD	Christian Democrats in Finland	7	6	7.2
SKP	Communist Party of Finland			
SEN	Finnish Seniors Party			
KTP	Communist Workers' Party			
STP	Finnish Labour Party			
IPU	Independence Party			
KA	For the Poor			
PIR	Pirate Party of Finland			
M2011	Change 2011			
VP	Liberty Party - Future of Finland			

2.1 Data Preprocessing

For our experiments, the qualitative data was first converted to a matrix $\mathbf{X}$.

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,k} \\ x_{2,1} & x_{2,2} & \dots & x_{2,k} \\ \vdots & \vdots & & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,k} \end{pmatrix}, \tag{1}$$

where n is the number of candidates. k is defined as a sum of all answer possibilities in all questions Q_s, $k = \sum_{s=1}^{n_q} \#options_s$, where n_q is the number of questions. A matrix element $x_{i,j}$ is zero or N_s. All answer options are equally weighted, if $N_s = 1 \; \forall s$. In the experiments, different definitions for variable N_s were tested.

The candidate's selection for the importance of each answer was stored in a weight matrix $\mathbf{W}(n \times k)$, where a matrix element $w_{i,j}$ is $1-a$ where $a \in [0,1]$ (candidates selection for his/her answer is not so important), 1 (default) or $1+b$ where $b \geq 0$ (important). With larger b, radical candidates are projected far from the center by questions which really divides opinions between the candidates. For example, if $a = b = 0.1$, candidates with strong opinions are not clearly separated from other candidates. And in practice it means that we lose important information about the importance of the questions. A candidate answer matrix for further analysis is defined by elementwise product as

$$\mathbf{X_b} = \mathbf{X} \cdot \mathbf{W}. \tag{2}$$

It is possible to find out which questions are important in general by selecting proper values for a and b. Finally, matrix $\mathbf{X_b}$ is mean centered before further analysis. More details about selection of variable N_s and parameters a and b in this study are explained in Section 4.

3 Methods

3.1 Principal Component Analysis

There is often redundancy in high dimensional data. However, none of the measurements is completely useless, each of them delivers some information. The solution for this is dimensionality reduction. [7] Feature extraction transforms the data on the high-dimensional space to a space of fewer dimensions. The data transformation may be linear, as in principal component analysis (PCA), but many nonlinear dimensionality reduction techniques also exist [8].

Principal component analysis (PCA) is a useful tool for finding relevant variables for the system and model. It is a linear transformation to a new lower dimensional coordinate system while retaining as much as possible of the variation. It is selected as the main method in this paper, because results are taken into continued consideration. In addition, orthogonal axes make visualizations easier to read. The PCA scores is range scaled [9] from -100 to 100 as

$$score_i = \frac{score_i - min(score_i)}{min(score_i) + max(score_i)} \cdot 200 - 100, \tag{3}$$

where i is a component number. Scaling of the score values make a value grid creation in Section 3.2 easier.

3.2 Kernel Density Estimation and Missing Value Imputation

Kernel density estimation is a non-parametric fundamental data smoothing technique where inferences about the population are made, based on a finite data sample [10]. In our experiments, several matrices (grids) were defined for densities. A value grid for each candidate is defined as

$$\mathbf{B_c} = \begin{pmatrix} h_{1,1} & h_{1,2} & \dots & h_{1,s} \\ h_{2,1} & h_{2,2} & \dots & h_{2,s} \\ \vdots & \vdots & & \vdots \\ h_{s,1} & h_{s,2} & \dots & h_{s,s} \end{pmatrix}, \tag{4}$$

where $h_{i,j}$ corresponds the candidates probability for certain opinion (i, j). s is selected to get decent accuracy for the analysis. In this paper PCA scores were scaled, see Eq. 3. Therefore the best choice for the size for each cell in the grid is 1×1. Then the grid size is defined as

$$s = max(score_{.,1}) - min(score_{.,2}) + 1. \tag{5}$$

A similar matrix $\mathbf{B_p}$ for each party or an individual candidate without a party is defined similar way as in Eq. 4.

A value grid for each candidate is constructed based on PCA score results by two-dimensional Gaussian function as

$$\mathbf{B_{c(i,j)}} = Ae^{-\left(\frac{(i-i_0)^2}{2\sigma_i^2} + \frac{(j-j_0)^2}{2\sigma_j^2}\right)}, \tag{6}$$

where the coefficient A is the amplitude, i_0 and j_0 is the center (in our case $score_1$ and $score_2$). A distribution volume for each candidate is scaled to one. In practice it means that if the center is near the border of the value grid $\mathbf{B_c}$, coefficient A is larger.

After all available data is used, missing data m is approximated by the distribution of the party value grids as

$$\mathbf{B_{c=m}} = v \cdot \mathbf{B_p}, \tag{7}$$

where $v = 1/\sum_{i=1}^{s}\sum_{j=1}^{s} \mathbf{B_{p=m_{party}}}$. If a missing candidate is not a member of any party, a sum of the distribution matrices is defined as

$$\mathbf{B_{tot}} = \sum_{i=1}^{\#parties} \mathbf{B_p} = \sum_{i=1}^{n} \mathbf{B_c} \tag{8}$$

and it is used as $\mathbf{B_p}$ in Eq. 7. So, all candidates have some kind of distribution in the value grids $\mathbf{B_c}$ (Gaussian, party distribution or total distribution) with the volume of one.

4 Experiments and Results

The concept of data mining process varies depending on the current problem. Many decisions for example in the parameter selections are based on intermediate results. In this particular context the flow of the experiments is presented in Fig. 1. Different candidate classifications and visualizations by histograms, scatter plots, factor analysis, PCA, Self-Organizing Map (SOM) and Multidimensional scaling (MDS) have been published after the data was released [11,12]. In these analysis a lot of information is just dropped off. In addition, the results are not combined with voting application data. In this paper, new visualizations and guidelines for future research are presented.

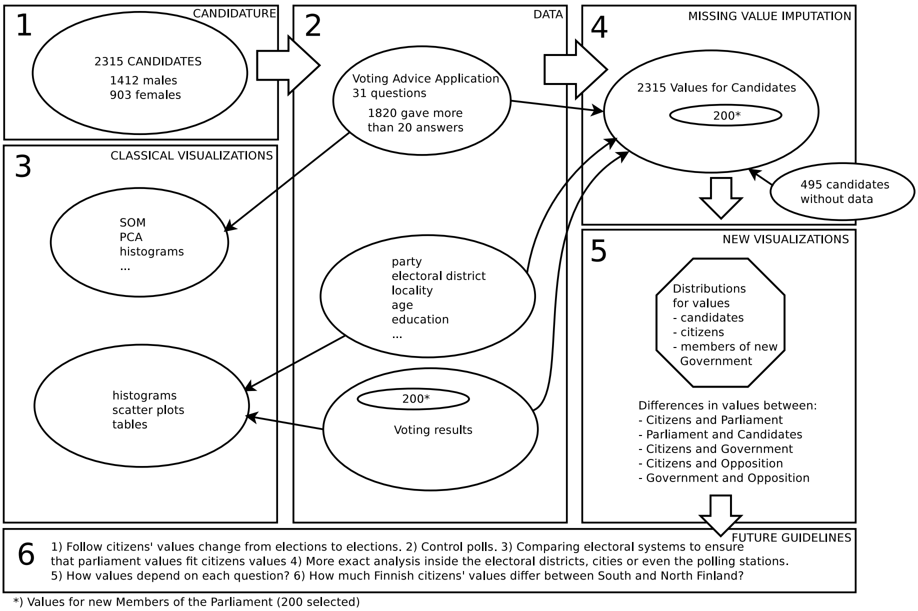

Fig. 1. The six stages of parliamentary data analysis

4.1 Preprocessing of Voting Advice Application Data

HS sent invitation for all candidates and 469 persons did not answer the VAA [2]. In two of the questions a candidate had to choose three options, so a total of $29 \cdot 1 + 2 \cdot 3 = 35$ answers were expected for each candidate. Although the instructions were clear, 39 candidates gave more answers than were asked. 1682 candidates gave 35 answers. 85 candidates did not answer to questions 33–34. 26 candidates gave 21–32 answers. 14 candidates gave only 1–20 answers.

For several parameters, some assumptions had to be made. One question is how many answers are needed to take the candidate in the further analysis.

If many candidates without data is taken in the analysis, PCA results (other candidates' scores, latents, loadings) have bias. These candidates with missing data would have approximately the same scores even if they were from different parties.

In our experiments, the candidates who gave more than 20 answers were selected. This selection is based on distribution of given answers by candidates. Only 14 of 1834 candidates were omitted and selected candidates have more than 60% of the answer information. In 31 questions there was total $k = 173$ answer options, see Eq. 1. A matrix $\mathbf{X}$ in the new higher dimension has $n = 1820$ rows and $k = 173$ columns.

Next question is how to select a and b for matrix $\mathbf{W}$ in Eq. 2 and what do "high" or "low" importance mean. For example, when parameters were selected as $a = 0.5$ and $b = 2$, PS, VAS and VIHR parties were separated from other parties. This example result can be explained by the information of PCA loadings. In the first component VAS and VIHR were separated, because candidates in these parties have the following strong opinions (absolute high loadings): negative attitude to nuclear power, the differences in the incomes should be reduced and development co-operation should be improved. PS was separated from the other parties because their candidates are against giving adoption right for gay couples and the role of Swedish language should be reduced in Finland.

The grades of membership functions have close association between the concept linquistic terms such as very low, low, medium (0.5), high, very high, etc. Ranges are over the interval $[0, 1]$. [13]

In our case matrix $\mathbf{W}$ elements are over the interval $[0, \infty[$. Therefore in our experiments, a good selection for parameters are $a = b = 0.5$. Experiments indicated that party positions were relatively similarly placed even if a and b were varied. Moderate values for a and b are justified because our goal is to visualize a distribution of the values, not to separate radical candidates from others.

The number of answer options varies between the questions and therefore, different weights for answers were tested. If weighting is not used, $N_s = 1 \; \forall s$ in Eq. 1. Questions with two options get less weight in the PCA projection. Next, the answers of each question were weighted with the number of options in the question. N_s is inverse of the number of options in each question, in Eq. 1.

It is evident that two variables are not enough to describe the variation measured by the 31 questions creating 173 different answers. The data was centered before PCA and after each weighting experiment. The coefficient of determination for the first principal component was 16.5% and for the second 9.48%. The problem is that questions with many options were underweighted, because there are many ways to approach it. Therefore the loadings on the first and the second component mainly depended on questions:

– In 2009, Parliament approved a law that allows adoption for registered gay and lesbian couples within the family. Should gay and lesbian couples have the right to adopt children outside the family?
– In the spring of 2010, the government granted two nuclear power plant licenses. The third candidate, Fortum, was left without, but hopes the next government would authorize the Loviisa power plant to replace two old reactors. Should new government give a license?
– In elementary school, there are two compulsory foreign languages, one of whom must be a second national language. Should the study of second national language be optional?

The next suitable experiment on the selection of variable N_s should be between an inverse of the number of options in each question and one. So, inverse of the square root of the number of options in each question is justified (Eq. 1). The coefficient of determination for the first principal component was 12.7% and for the second 7.51%. Answers in the three questions above still had meaningful loadings (not as large as with $N_s = 1/\#answers_s$). These questions Q_s have often been discussed in media and these questions divide some parties into different sections. Although the coefficients of determination were smaller there were more meaningful PCA loadings in the model. Therefore, $N_s = 1/\sqrt{\#answers_s}$ was used in subsequent experiments.

4.2 Values of Candidates and Parties

A value grid $\mathbf{B_p}$ for each party was computed. The cell size in the grid was 1×1, so the grid size is 201×201 (see Eq. 5). In total 1820 PCA scores were inserted to the party value grids using a two dimensional Gaussian curve as follows. A single candidate was inserted to the grid by mean $\mu(x) = score_1$, $\mu(y) = score_2$. In our experiments a variance $\sigma^2 = \sigma_i^2 = \sigma_j^2$ was used. The values of variance σ^2 was selected so that it would produce clear visualizations. In practice this means that candidates of the same party were mainly joined together in the value grid $\mathbf{B_p}$. with $\sigma^2 = 0$ the visualization looks like the original PCA score plot. With higher values of σ^2 a wider group of citizens voted the candidate. With $\sigma^2 = \infty$ a flat surface is produced. With the selected scale and grid size clear results were achieved with $4^2 < \sigma^2 < 12^2$, so we used $\sigma^2 = 64$ in our experiments.

Missing data was approximated by the distribution of the party value grid $\mathbf{B_p}$. The head of The Finnish Social Democratic Party (SDP) Jutta Urpilainen, for example, did not answer the VAA. In this study, it is supposed that her opinions and values have same distribution as her party $\mathbf{B_{c=Urpilainen}} = v \cdot \mathbf{B_{p=SDP}}$, where $v = 1/\sum\sum \mathbf{B_{p=SDP}}$. Most probably her values are same as the mode of $\mathbf{B_{p=SDP}}$.

The minimum and maximum values in the principal components were between -1.96 and 1.91 (1. component) and -1.74 and 1.68 (2. component). These axes were scaled, see Eq. 3. This means that the second axis is stretched by 12.7% compared to the first one. Therefore, all contour plots are scaled into 1:1.127. A sum of value grids $\mathbf{B_{tot}}$ (see Eq. 8) as a contour plot is visualized in Fig. 2. All value distributions were multiplied by $2\pi\sigma^2$ to enhance the readability of labels in the contour plots.

In the visualizations and in the analysis of the principal component coefficients it seems that the first axis depicts the economic left wing vs. right wing classification (cf. Table 1). The second axis seems to reflect at least partially liberal vs. conservative opinions. The same definitions for the axes were achieved in the previous candidate classifications [6,11,12].

Some remarks can be made such as the candidates of SKP, VAS, KTP and STP had similar opinions. Candidates of KOK had the most different opinions comparing to these candidates. An interesting remark is that they got the most seats, see Table 1. This conclusion is based on Fig. 2 and the first and second coefficients of determination in principal components without committing what is left/right or liberal/conservative. The bimodal candidate distribution have two different modes $(40, 0)$ and $(-22, 23)$. In this paper, all votes were combined with the candidate value grids $\mathbf{B_c}$ and completely new visualizations are created.

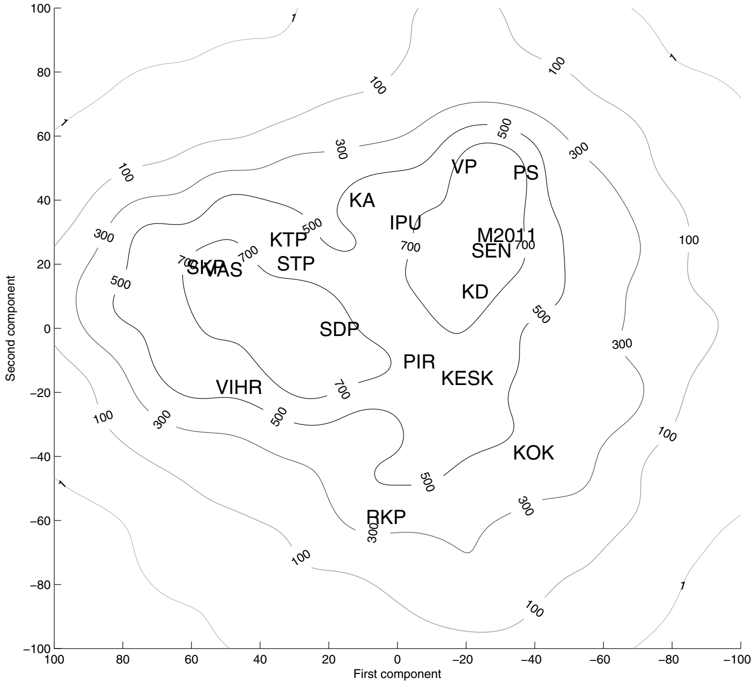

Fig. 2. The values of the candidates. The visualization can be used to get an estimate how the median values of candidates in each party (*labels*) are related to each other. The numbers in contour lines describe the density ρ of candidates. $V = \frac{\rho}{2\pi\sigma^2} B_{tot} = 2315$.

4.3 Values of Finnish Voters

It is assumed that a citizen gave a vote based on his/her values by selecting a candidate with the same opinions and values. The values of the Finnish voters are defined as a sum of candidate value distributions weighted by votes as $\mathbf{B_{Finn}} = \sum_{c=1}^{2315} votes_c \cdot \mathbf{B_c}$. The sum of these distributions is an estimate of Finnish voters' values. This is illustrated in Fig. 3. The modes of the party supporters can be classified to three groups, see *party labels* situated near the modes of $\mathbf{B_{Finn}}$. Finnish citizens who voted VAS, VIHR and SDP were classified as left-wing, see Table 1. PS, KESK, KD, SEN and M2011 supporters are more conservative than other citizens based on coefficient a and b testing results in Section 4.1. Those who voted KOK and RKP are more liberal and right-wing than PS and KESK.

Finnish citizens who have the same opinions and values as KD voted more probably for either PS or KESK than KD. KESK suffered the heaviest defeat in the elections, so it can be speculated that some citizens who are not very religious but conservative voted PS rather than KD.

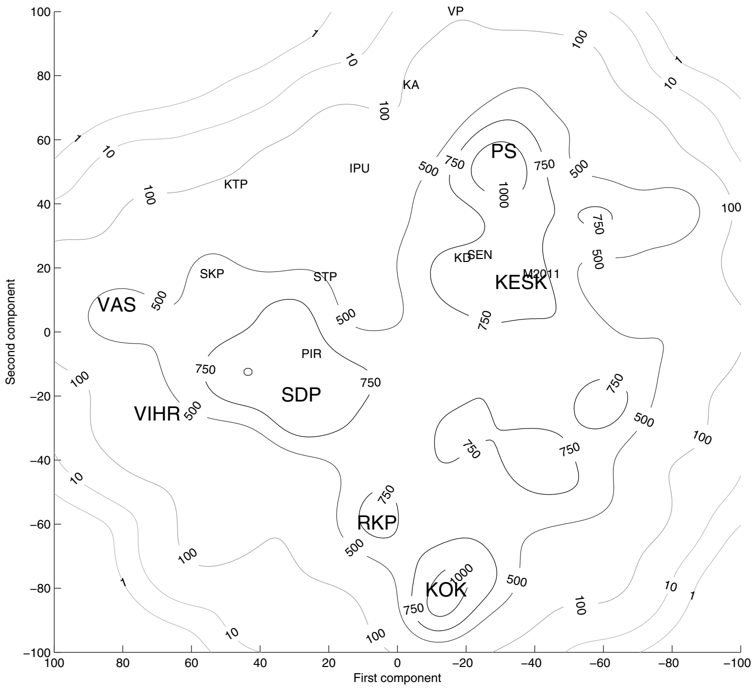

Fig. 3. The values of the Finnish voters. This visualization can be used to get an estimate for one party how the opinions of the supporters are related to the supporters of other parties. Each *party label* is a mode of its supporters and those are in smaller font, if the density of the party supporters (in the mode) is less than some other partys density. The *numbers* in contour lines describe the density of citizens in thousands.

4.4 The Values of the Parliament

The 200 MPs' values and the value distribution of each party in the Parliament is shown in Fig. 4. The distributions are not sums as in earlier contour plots.

The increase in the number of seats for the True Finns (PS) from five to 39 is a change of historic proportions. It is expected that PS picked up votes from the voters of other party with the same opinions such as KESK. KESK have supporters with rather different values. They seem to have lost liberal right supporters to KOK, liberal left to SDP and conservative KESK supporters to PS. They suffered the heaviest defeat and lost 16 seats, see Table 1.

The form of the new government was an open question on the beginning of May, 2011. Speculations of the next government have been published in local and international media [14]. Mostly majority governments $KOK + SDP + PS$, $KOK + SDP + PS + RKP$ and $KOK + SDP + VIHR + RKP + KD$ have been mentioned. Visualization of the values of the Parliament can be used for supporting the formation of the new coalition government.

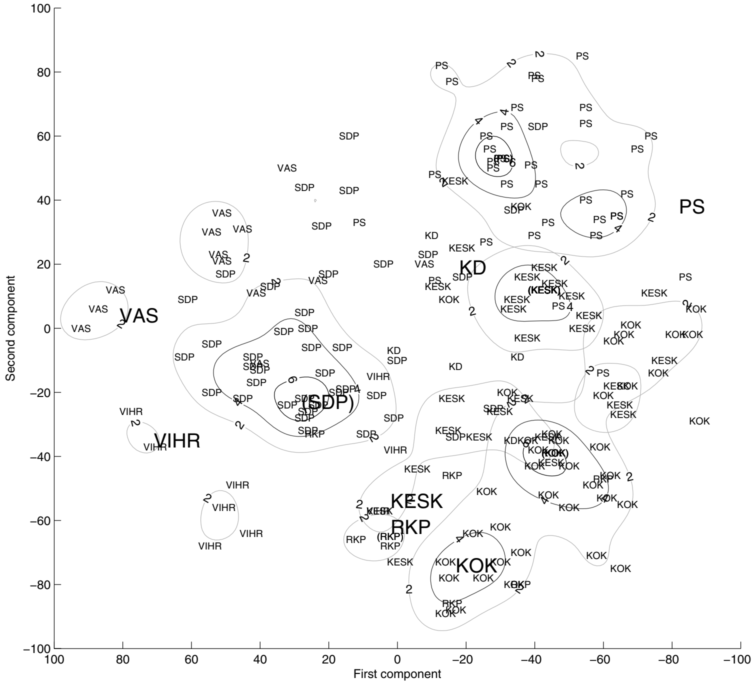

Fig. 4. MPs' values. Each *party label* corresponds to one politician. The president of the party is visualized by larger font. *Numbers* in contour lines describe the density of politicians in each party. MPs whose values are approximated (*labeled in parentheses*) are close to the point $max(\mathbf{B_p})$.

5 Discussion

A huge amount of new possibilities for data mining in political space was opened by combining the two data bases analyzed in this study. Using the analysis methodology, it is possible to answer, e.g., the following questions: How well the values of the citizens fit the MPs' values? Would they fit better with some other electoral system? Is there difference in value distribution between young and old candidates?

Areal information about value distribution can be achieved by limiting voting application data by electoral districts. The results for each candidate by cities or even the polling station is available. Citizens' values can be visualized on the map for example. Difference of value distributions between some polling station, such as different parts of a town can be used to explain some other measures, such as housing prices, criminal statistics, unemployment, etc.

We are working on an important research area concerning everyone in democratic countries. In this paper, dynamics (candidates → results → parliament → government) of only one election was investigated. Political data, election results, pollings in the parliament etc. are produced regularly. By combining results presented in this paper, much more complex analysis becomes possible.

A fundamental way to argue against our approach could be that "The best argument against democracy is a five-minute conversation with the average voter." – Winston Churchill. True or false, in this article we presented preprocessing and analysis methods for VAA data, the visualizations of value distribution of the political field that seemed to be relatively insensitive to small changes in the data. Also a number of new possibilities for future experiments in political data mining were presented.

References

1. The elections website of the ministry of justice, `http://www.vaalit.fi/` (retrieved on April 2011)
2. HS Voting Advice Application (in Finnish), `http://www.vaalikone.fi/` (retrieved on April 2011)
3. HS blog (in Finnish), `http://blogit.hs.fi/hsnext/hsn-vaalikone-on-nyt-avointa-tietoa` (retrieved on April 2011)
4. xls-dataset - Typical way to convert qualitative data to quantative, `http://www.loitto.com/tilastot/hsvaalikone11/HS-vaalikone2011_num2.xls` (retrieved on May 2011)
5. Manow, P., Döring, H.: Electoral and mechanical causes of divided government in the european union. Comparative Political Studies 41(10), 1349 (2008)
6. National election in Finland (March 18, 2007), `http://www.parlgov.org/stable/data/fin/election-parliament/2007-03-18.html` (retrieved at May 2011)
7. Belsley, D., Kuh, E., Welsch, R.: Regression diagnostics: Identifying influential data and sources of collinearity, vol. 546. Wiley-Interscience, Hoboken (2004)
8. Hair, J., Anderson, R., Tatham, R., Black, W.: Multivariate Data Analysis, 5th edn. Prentice-Hall, Englewood Cliffs (1998)
9. Pyle, D.: Data Preparation for Data Mining. Morgan Kaufmann, San Francisco (1999)
10. Scott, D.W.: Multivariate density estimation, 5th edn. Wiley Online Library (1992)
11. Data Rangers example (in Finnish), `http://extra.datarangers.fi/vaalit/` (retrieved on May 2011)
12. Candidate Map (in Finnish), `http://nypon.fi/Vaalit2011/` (retrieved on May 2011)
13. Zadeh, L.: The concept of a linguistic variable and its application to approximate reasoning–I* 1. Information Sciences 8(3), 199–249 (1975)
14. Daley, S., Kanter, J.: Finland's Turn to Right Sends Shivers Through Euro Zone. New York TImes (April 21, 2011)