The Pennsylvania State University
The Graduate School
College of Information Sciences and Technology

POLITICAL OPINION IDENTIFICATION, MINING AND RETRIEVAL

A Thesis in
Information Sciences and Technology
by
Lei Zhu

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2010

The thesis of Lei Zhu was reviewed and approved* by the following:

Burt L. Monroe
Associate Professor of Political Science
Thesis Co-Advisor

Donald R. Shemanski
Professor of Practice

Madhu C.Reddy
Associate Professor of Information Sciences and Technology
Director of Graduate Programs

*Signatures are on file in the Graduate School.

Abstract

I provide a critical literature review on Computational Political Science in this thesis, which summarizes studies of political science issues utilizing computational techniques. Text analysis and Network analysis, the two main sub-fields in computational political science are discussed in detail, and the usage of miscellaneous computational techniques in political science is also addressed.

I present my studies on the problem of Political Spectrum Analysis, namely text-based ideal point estimate, in Chapter three as an example of computational political science. Political Spectrum refers to a multidimensional opinion space where each geometric axis models one political dimension. Political opinion mining shares some characteristics with product reviews mining [39] [14] while introducing new challenges to opinion identification, modeling and representation. The study starts from the congressional political domain. I show the importance of multidimensional opinion representation in the congressional context combining domain knowledge and results from three different dimensionality analysis methods. Several regression models are trained to get ideology scores from the text, based on both Bag-of-words feature sets and Topic-based feature sets. I also transfer to the civic political domain by studying a tagged blog space with the learned regression models from the congressional domain.

Real world applications of both political opinion mining and political opinion retrieval are discussed in the last chapter and several user scenarios are proposed to conclude the contribution of my studies and reflect future potential.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction to Computational Political Science

Political Science is a branch of social science concerned with the theory and practice of politics and the description and analysis of political systems and political behavior. The study methodologies used in political science usually include Formal Theory Building, Narrative Analysis, Quantitative Analysis and Survey-Based Analysis [1] etc. As Lee Sigelman pointed out in the review paper of the 100 years of publication history in the American Political Science Review [79], the presentation of empirical results is the primary purpose of most papers, while quantitative analysis as the featured methodology has a "dramatic upsurge" during the last half of the 20th century. Quantitative analyses again evolve into computational analyses of big data in the 21th century [86]. The trend is repeatedly confirmed by top researchers in the social sciences [54, 9], calling it "the coming age of computational social science". Data-driven computational social science bears the capacity to collect and analyze massive amounts of information. With computational technologies, social science studies including political science studies extend their scope from individual studies to group interactions and society studies [54].

Political science is usually divided into the following major sub-fields: American

[1] http://en.wikipedia.org/wiki/Political_science

politics, political theory, public policy, international relations, and comparative politics. In this summary, I will explore computational political science, which covers studies in all the sub-fields of politics. I will not review those studies that simply use a computer for basic computation but relate to no computational science research. Some studies, such as Computer-assisted survey research and Computer-based qualitative analyses, use existed computer tools to facilitate the analysis, but they barely contribute to current computer/information science studies, thus they would not be covered in this survey.

I also want to differentiate computational political science with statistics-enhanced political analyses. Statistical methods like numerical analysis, regression, statistical modeling are ubiquitously applied in all the social science studies, and political scientists use them to explain almost all political phenomena about voters, elections, policies etc. These studies are usually not directly related to computational science, although the underling statistical models may need to be implemented using computational methods. In this sense, I will also not cover them.

Computational Political Science, as defined by the recruiters in the Department of Political Science at the University of Massachusetts Amherst, encompasses both the analysis of computer-generated data from the web, sensors, communications, electronic media or digital databases and the use of computational formalisms and languages to describe and analyze political phenomena. Computational techniques of particular interest in this survey include social network analysis, text analysis, agent-based modeling, dynamic relational or clustered modeling, qualitative data mining, simulations of social processes based on models with realistically complex assumptions, and statistical analyses of very large sets of relational or clustered data. The papers reviewed in this survey will show that these techniques are usually chosen according to the nature of the dataset utilized or the political problems targeted.

Chapter 2

Computational Political Science

I will categorize this survey into three sections. The first section features studies using text analysis methods. The second section focuses on network analysis, especially emerging social network analysis techniques. And the third section summarizes all the other techniques including agent-based models, mathematical logic, web 2.0, geographic information system, global positional system, cloud computing etc. In general, all these studies deal with objective data automatically collected and analyzed by computer programs, which best represent the fundamental belief of computer science: let human beings be relieved from everything but thinking.

2.1 Text Analysis

Scientists usually use the most appropriate computational techniques to "play with" the corresponding format of data. "Data" herein refers to all the possible information carriers, like image, sound, video, roll-call records, bills, survey results, polls etc. Among all data formats, text is the most common and useful information resource format that attracts research interest from both disciplines, i.e. political science and computational science. "Text is arguably the most pervasive and certainly the most persistent artifact of political behavior", Monroe and Schrodt [64] wrote, "the possibility that the analysis of texts could provide insights into the political processes has a long pedigree."

Text analysis is usually referred to as "computer annotation" or "automated content analysis" [1]. The later term derives from "classic quantitative content analysis", where communication content (speech, written text, interviews, images, etc.) is analyzed and categorized by human coding. Computer scientists usually use "text mining", or "text data mining", to refer to the process of deriving high-quality information like patterns and trends from text automatically.

Text mining has many applications in multiple fields, e.g. spam mail filtering, sentiment analysis of customer reviews, medical records management, and detection of terrorist activities. In political science, it is applied to analyze election campaigns and voter profiles, and for "determining ideological position from texts, coding political interactions, and identifying the content of political conflict" [64], etc.

Content analysis typically works at the word/sentence level under the "bag-or-words" model. Words carry information, and statistics like word frequency, prominence of words or expressions, distinctive/representative terms are processed. These statistics can thus provide information for further text mining needs. Typical text mining tasks [22] include text classification/clustering, information (concept/entity/relation) extraction, topic modeling, sentiment analysis, document summarization, etc. Natural Language Processing usually supports text mining from the perspective of linguistics.

Modeling the text as words or n-grams, although it remains to be the dominant method in text mining, obviously bears the risk of losing any syntax or semantics information in the text [64]. Even if the latest developments in text mining partially solves the problem with probabilistic techniques like topic modeling [2], hidden Markov models[3], conditional random fields[4] etc., machine intelligence on interpreting texts still cannot be compared with human intelligence in the sense of information lost during the process.

I will briefly review studies on several different tasks of political text analysis in

[1] Text Annotation for Political Science, Journal of Information Technology and Politics, Volume 5 Issue 1 2008

[2] http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

[3] http://en.wikipedia.org/wiki/Hidden_Markov_model

[4] http://en.wikipedia.org/wiki/Conditional_random_field

the following subsections.

2.1.1 Fighting Words

Political scientists fight with words almost always to extract ideological positions from texts. The pivotal study in this area is Benoit and Laver's Wordscores [52] [5]. Simply speaking, it is a technique to generate word scores from reference texts with a priori known positions, and then score each virgin text using the generated word scores. Note that here the authors used the term "reference text" and "virgin text" to refer to the two classes of text with respectively known and unknown information of their categories on them. In computer science, this kind of technique is named "supervised learning", and "reference text" is called "training set", while "virgin text" is called "testing set". The word-scores method be applied to various datasets. For example, the authors estimated policy positions on party manifestos in Britain, Ireland and Germany, as well as the legislative speeches. The technique can be easily used on different datasets as long as a reasonable sample of "reference text" exists.

The Wordscores technique inspired a considerable amount of further studies [64]; nonetheless, it was also inevitably challenged by other researchers. Slapin and Proksch in their WORDFISH paper [80] pointed out several concerns with it: it deeply depends on the reference texts; it uses exactly the same reference texts for multiple dimensions; it weights all the words the same; and time-series estimation is problematic with it. Aiming to address these issues, Slapin and Proksch [80], as well as Monroe and Maeda [65], proposed their "unsupervised learning" techniques to estimate ideal points of the legislators. Unsupervised learning helps to estimate time-varied ideal point scores since the time factor is automatically incorporated in the statistical models; in supervised learning, reference text on various times much be provided for the method to work properly across time.

Just as supervised and unsupervised methods are both pervasively utilized, the unsupervised techniques mentioned [80, 65] herein also has their limitations. For example, since they must propose their own models for estimating ideology, they are inherently entangled with the underlying ideology theories in their proposed models.

In wordscores, the ideology theory is independent of the technique, and given by the ideology experts, ex ante. But in unsupervised learning, the scoring methods themselves need to take into account the definition of ideologies.

In another typical "playing-with-words" paper, Monroe et al. [63] discuss a variety of different approaches to the problem of feature selection and feature evaluation. They examined the lexical differences between Democrats and Republicans in the United States Senate. Fader, et al.[21], give an interesting paper on simulating a network analysis study on pure textual data. They use the TF-IDF cosine similarity [5] to construct a similarity graph, and calculate lexical centrality on the graph for identifying Influential Members of the US Senate. They show how a connection can be established between text analysis and network analysis to transform textual data into network data.

2.1.2 Classification and Clustering

Having the features of words detected, the typical next step is to categorize the texts using the word features. Classification and clustering are two similar methods in machine learning to categorize texts. Classification categorizes items/documents to some pre-defined classes according to their common attributes; and clustering groups together items/documents that have similar characteristics. The biggest difference between the two is: the former is a supervised learning method, while the latter is an unsupervised learning method. So for the task of classification, human-coded examples must be given as the training data.

Purpura and Hillard [75] designed a system for automated classification of congressional speeches. They wanted to classify the speeches into one of 226 subtopic areas. The training data is easy to obtain in this case, because the same task has been done by human coders before, and the standards for classifying the texts have long been established [6]. They used a popular classification tool in machine learning—the Support Vector Machine algorithm to achieve their goal. The experiment result is

[5] http://en.wikipedia.org/wiki/Tf-idf

[6] Policy Agendas Project: www.policyagendas.org

not surprising since SVM has shown its consistent performance in various classification tasks: they found that "the automated system is about as effective as human assessors, but with significant time and cost savings."

Yu et al. [89] also looked at the congressional speech data with the help of the same SVM algorithm, but they did not classify it into subtopics; instead they want to classify party affiliation from the texts. In other words, they want to identify the ideological polarity in congress. The task can also be done by the wordscores-style methods, but machine-learning techniques help reduce the complexity of the problem, and provide another way for automatically determining the weights of single words.

Hopkins and King [38] also proposed their own classification/rating methods, and they even tested their methods in multiple data sources like blogs, movie reviews.

Grimmer and King's effort in Clustering is presented in the book "The Future of Political Science: 100 Perspectives" [47], in which they applied their method to cluster the 100 essays of 100 political scientists talking about the future of political science. The essays were divided into several clusters, each with different focus of future directions. They also illustrated and tested their clustering algorithm on the Press Releases [32].

2.1.3 Sentiment Analysis

Sentiment Analysis aims to determine the attitudes of speakers/writers with respect to some topic in text. Generally speaking, the studies on policy positions I discussed in the "Fighting words" section; and the studies on ideal point estimation of roll-call data (I will briefly mention this thread of studies in the discrete data analysis subsection later) also fall in the field of "sentiment analysis". Classic sentiment analysis studies should provide fine-grained analysis on the words from the linguistics view, i.e. studying the sentiments on word and phrase level and identifying the emotional bias of representative terms; such as detecting ideology preferences of words on the personalized habit of language usage. For example, Thomas, et al. [83], and Diermeier, et al. [15], studied language and ideology issues on congressional speeches by investigating the contributions of words on ideology.

However, it is worth noting that it is not always appropriate to simplify the political opinion mining problem into a classification task, especially when the target of study is informal political discourses like posts in online political forums. As Mullen and Malouf [66] realized, "posts made in direct response to other posts in a thread have a strong tendency to represent an opposing political viewpoint to the original post." In this case, web forum posts with totally opposite opinions might overlap a lot in contents, which makes the task of automated word-backed classification extremely hard. They also pointed out that "difficulty with analysis of informal text, for example, is dealing with the considerable problem of rampant spelling errors. This problem is compounded when the work is in a domain such as politics, where jargon, names, and other non-dictionary words are standard."

Mullen and Malouf thus concluded that "traditional text classification methods will be inadequate to the task of sentiment analysis in this domain, and that progress is to be made by exploiting information about how posters interact with each other." They are suggesting applying network analysis to the problem of political sentiment analysis, which I will discuss in the next chapter.

2.1.4 Topic Modeling

Probabilistic Topic Models, as tutored by Steyvers and Griffiths, "are based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. A topic model is a generative model for documents: it specifies a simple probabilistic procedure by which documents can be generated." [81] They are another unsupervised learning tool to analyze text.

Quinn, et al., [76] applied topic modeling to the United States Senate speeches; they obtained 42 topics from the data and labeled the topics to construct meaningful categories. This study can be compared with the aforementioned study by Purpura and Hillard [75] on classifying congressional speeches. Purpura and Hillard classified the speeches into 226 pre-defined subtopic areas; and Quinn, et al., [76] automatically obtain the topic division with no effort made on coding the training samples. Grimmer [31] also applied this technique to press releases from senators in 2007. One possible

problem with this method comes from the assignment of labels to the topics. Because the topics are all machine-generated; thus they have mistakes and can be stained by noisy information in the texts.

Topic modeling and opinion mining can be connected by some newly designed unsupervised models. Chen et al. [11] proposed a generative model to automatically discover the hidden associations between topic words and opinion words. With this mixed model, the authors successfully extracted statements which best express politicians' standings on certain topics. This method is a recent development in computer science, and has the potential to be applied to more political science studies.

2.2 Network Analysis

Network theory is an area of computer science and network science. The theoretical foundation of network analysis is based on the mathematical conceptualization and abstraction of real network/graph structures. Thus most of the techniques need come from graph theory, where objects are modeled as vertices or nodes in the graph, and relations are modeled as edges that connect pairs of vertices.

Network Analysis has application in many disciplines including particle physics, computer science, biology, economics, operations research, and sociology. In some disciplines, network analysis is the dominant research methodology due to the inherent nature of the problems, such as in traffic analysis or in file transfer protocol design. In political science, statistical modeling and causal inference are usually the key to interpreting political phenomena and answering particular research hypotheses, where the researchers believe in the "cause and effect" philosophy, that political behaviors can be explained by some known/unknown, explorable/unexplorable features. The focus is on the explanation of the phenomena, instead of explaining/understanding the process of network interactions.

However, politics as well as other social sciences, studies the ways human beings live, compete, cooperate, and compromise. Interactions are everywhere in both the elite political circles (e.g. the judiciary, the legislative, and the executive branches of government), and the informal political discourse (as evidenced by polls, surveys

etc.), and all these interactions contribute to the basic elements of network analysis, the network structure.

Political science network analysis also usually utilizes graph theoretical techniques. However, the novelty of their network structures is of special interest to us. The studies are usually data-driven; the research methods and outcome are highly dependent on the network data. I would like to summarize the studies in this sub-field into three categories. The first, also the easiest one to think of, is the social networks. Here a social network refers to a specially defined network, that is, "a social structure made of individuals", as compared to a second type of networks, the "general social network" defined on groups (including cliques and cohesive blocks), organizations, nation states, web sites, citations between scholarly publications, etc. In this study we name the general social network "Politic Networks". The third type of networks is rooted on the internet. It includes online SN websites like Facebook and Twitter[26], and web blogosphere. Research topics include the impact of web 2.0 to politics, social media, blogosphere, and internet users. These studies, as discussed in the next subsection, resemble the social media studies in computer sciences [7], sharing the feature that they are all focusing on the novelty of the problems instead of the technologies involved. Resources for political network studies include The ITP section of APSA, the Harvard/Duke politics network conference, the journal of political analysis, and the OpenSIUC political networks paper archive, etc.

Studies that we are interested must study the network directly, instead of just studying the interviews conducted on social network websites or surveys on the usage of SN tools. I have noticed some papers use social networks to get dependent/independent variables, and they conduct some feature-based regression analysis. These studies [34][27], although also important for understanding social networks in politics, don't really utilize the dynamics of the network structure, and thus contribute little to computer and network science. In rare cases, network data is obtained by non-computational methods as in [60], where the data was surveyed; after gathering the data, they analyze the network. In most cases, network analysis requires a big dataset, which can only be collected with the help of computers instead of gathered

[7] http://www2010.org/www/program/papers/

manually in surveys.

2.2.1 Online Political Blogosphere Mining

Many studies have been pursued on the online political blogosphere. As claimed by Ackland in [1], "Weblogs are now a key part of online culture, and social scientists are interested in characterizing the networks formed by bloggers and measuring their extent and impact in areas such as politics." In this study, Ackland asked a simple and political behavior related question: "Are Conservative Bloggers More Prominent?" It might be an arbitrary conclusion that the conservative bloggers are more prominent, but the computer algorithm the author used is really very prominent. He used Kleinberg's HITS algorithm to measure the blogs' web visibility by calculating the authority and hub score, defined iteratively over the inbound and outbound links of the blogs. The authorities and hubs score, as well as the Google PageRank score, provide a way to inspect individual nodes' visibility/importance integrating knowledge from the overall network. Hargittai et al. [36] also looked at the conservative/liberal blogs, but they focused on the "cross-ideological discussions", defined by the number of links within the same wing (internal links) and between the two wings (external links). They also looked at the changes over time, and came to the conclusion that "information technologies will NOT lead to more isolation and insularity", as against Sunstein's theory about political fragmentation and polarization.

The seminal work of this thread is the "Divided They Blog" study [2], in which the authors also collected and ranked the conservative and liberal blogs and analyzed their linking strategies. They also examined links from blogs to the media pages, and showed the interaction of the blogs with mainstream media. Kelly et al. have a similar link analysis study on the political discussion network of Online discussion groups [46].

Another thread of studies [85, 45] target at "blog-mining" the blogosphere in countries under tight control, surveillance, blocking/filtering of the government. Figure 2.1 shows the levels of internet filtering over the world [8]. As shown in this figure, Chinese

[8] the OpenNet Initiative http://opennet.net/, an institution aiming at investigating Internet censorship and surveillance.

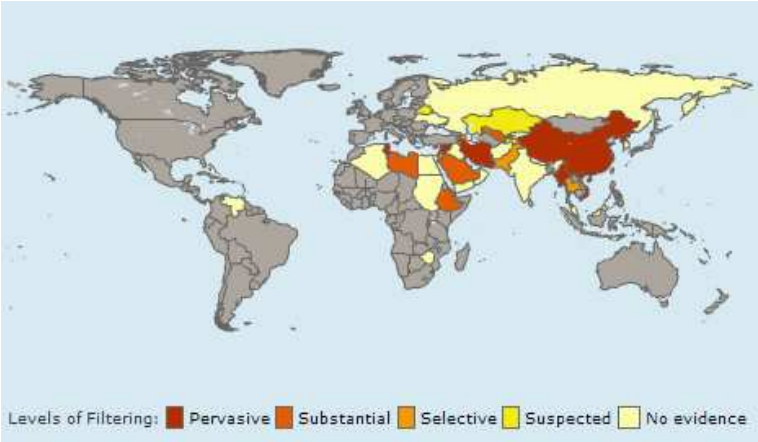


Figure 2.1: Levels of internet filtering

bloggers suffer from pervasive internet filtering, as well as bloggers from some other counties concentrated in Asia and Africa. Malaysia and Iran are two other countries where the internet is controlled.

Brian Ulicny examined the Malaysian blogosphere in [85]. He points out, "Recent confrontations between Malaysian bloggers and Malaysian authorities have called attention to the role of blogging in Malaysian political life. Tight control over the Malaysian press combined with a national encouragement of Internet activities has resulted in an embrace of blogging as an outlet for political expression." Through categorization and link analysis of the Malaysian blogs, he first identified the most influential blogs with the highest in-degree centrality, then compared the behavior of social/political bloggers with ordinary bloggers in their engagement with news sources, and third, he found out that the number of active Malaysian social/political bloggers is on the order of 500 to 1000 blogs, rather than the potential millions suggested by another study. He also introduced the usage of information retrieval technology for building a search engine to blog-mining in the Malaysian blogosphere.

The Persian Blogosphere in Iran was studied by John Kelly and Bruce Etling in this Berkman Center research paper [45]. The political blogs are divided into the clusters of Secular/Reformist and Conservative/Religious, instead of conservative/liberal in the United States. The authors employed a Fruchterman-Rheingold 'physics model' algorithm to model the network structure. Also they used human

and automated content analysis for topic coding, term and list frequencies calculation, and out-link analysis. Some more advanced text analysis techniques should have the potential to be applied to replace the human-coding in this study. The authors discovered that "Blocking of blogs by the government is less pervasive than we had assumed. Most of the blogosphere network is visible inside Iran, although the most frequently blocked blogs are clearly those in the secular/reformist pole. " they also advocated the "peer-to-peer architecture of the blogosphere" against "the older, hub and spoke architecture of the mass media model" in light of the fact that "blogs may represent the most open public communications platform for political discourse given the repressive media environment in Iran today".

There is room for improvement in the two studies described above with respect to the employment of computational network analysis techniques, but they serve as the best examples for directing a systematic and analytical examination on the political blogosphere of a specific country, and they help readers from another culture/political system to understand the target country. Some really good comparative politics studies can be expected to be drawn from this thread of research.

2.2.2 Interpersonal Social Network Analysis

The aforementioned blogosphere studies usually feature exploratory analyses using graph theory and link counts; this is in contrast to traditional political science analysis where explanatory variables with political importance are identified and collected and regressions are conducted to predict the dependent variables as an attempt to find the relationships between variables/phenomena. I will show that studies on traditional fields like congressional studies [23, 33, 49] usually combine both the exploratory social network analysis and regression analysis to obtain more convincing arguments than what either techniques allows in isolation.

I will first introduce two papers on the congressional co-sponsorship network study. In the co-sponsorship networks, legislators are connected if they co-sponsored bills; the list of co-sponsors in each bill and the final vote of each legislator are usually also available for analysis. I name this section "Interpersonal Social Network Analysis"

because here I adopt the narrow definition of social network analysis that each node in the network should be some individuals like the members of congress.

The primary research questions in James Fowler's "connecting the congress" [23] are "Who are the most connected legislators?" and "Does connectedness correspond to influence?" To answer the first question, he used a number of statistics to describe the legislators in the co-sponsorship network such as the quantity of legislation sponsored and cosponsored by each legislator, and the graph theoretical measures of closeness, betweenness, and eigenvector centrality. Fowler also proposed a new measure "connectedness", "which uses information about the frequency of co-sponsorship and the number of cosponsors on each bill to make inferences about the social distance between legislators." All these measures answer the first question to some degree.

The second question is answered by a general linear regression between the connectedness and the legislators' voting choices controlling for the ideological score (DW-Nominate Score) and partisanship. Fowler found a positive relationship in both the House and the Senate, which proves that "connectedness corresponds to influence".

Justin Gross investigates the U.S. senators' propensity to support one another's proposals in the co-sponsorship network [33]. It proposed a new measure "WPC": the weighted propensity to cosponsor. It pursues a similar exploratory analysis as that pursued by Fowler and uses GLMM, the generalized linear mixed model, to "examine how such social factors as homophily, proximity, and institutional role are associated with varying odds of cosponsorship among senators."

Although co-sponsorship might be the most open and canonical collaboration relationship in the congress, there are still many more connections among the legislators that are not necessarily known or noticed. Robbins provides an interesting observation paper [77] on the "leadership political action committees" (PAC) network in the U.S. congress. The PACs are the campaign donation committees created and chaired by individual legislators that directly contribute to them for their campaigns, as well as to the party and any other PACs. "Leadership PACs provide a way for redistributing fundraising wealth from safe, well-funded legislators to challengers and competitive races." The cash flow from one PAC to another thus creates the ties in the

network. The author studied this network overtime with the betweenness centrality measurement.

The Lobbyists' Donations [49] constructs another donation network to the members of congress. The number of common donors between legislators indicates the ties between them. In this study, the authors try to explain the number of common donors with factors like the party, state, committee, vulnerability in the next election etc. They used the Random Intercept Poisson Model to accommodate the explanatory variables.

Interpersonal social network also includes the presidential nomination co-endorsement network [68]. Analysis based on this network is expected to "give insight into who is important, what groups are stable, and what characteristics lead the endorsers to act together", etc.

2.2.3 Intergroup Social Network Analysis

Intergroup Social Network Analysis refers to studies on group/organization networks, such as parties [48], interested groups [6], and NGOs [8]. As Koger, et al., posit: "SNA techniques are especially useful for studying political parties and interest groups since these actors are best understood as networks of co-operating allies." [48]

Koger et al. [48] provides an exploratory study of the "Partisan Web". This study identifies links between formal party organizations and informal networks of interest groups, media and 527s (tax-exempt political organizations). The links are specially defined on the "donor name list exchange" relationship: one group exchanges the name list of their donors with other groups; thus they play the role of bridges to connect these political organizations together in a single network. The authors used the usual network measures (graph theory, exploratory analysis) to analyze this network, and applied the NETDRAW software [9] for network virtualization as well as many other SNA studies.

Another exploratory study [6] examines interest group networks. The network is constructed on the cosigner status to United States Supreme Court amicus curiae,

[9] http://www.analytictech.com/downloadnd.htm

or friend of the court briefs; the co-signer relationship, as believed by the authors, sheds light on "the interest group coalitions formed to impact governmental decision making and policy." The authors adopted the standard network measures, calculated the statistics of the graph, and detected the most central interest groups; they also examined the respective egocentric networks of the central interest groups.

The network can also be big enough to include all the important actors in world politics, including states, IGOs, NGOs, transnational corporations, academic institutions, news media, municipalities, think tanks, and private individuals. In the ongoing study of the cooperative response network constructed after the Indian Ocean tsunami [44], the authors find the relations of all these actors in their financial transactions and cooperative interactions, and apply social network analysis techniques to this system level network to explore their co-operations during and after disasters. It demonstrates "how a new way of thinking about the constitution of system level world politics can produce knowledge not available to traditional methods", and thus demonstrates the power of social network analysis.

2.2.4 Network Analysis Methodologies

Citation network analysis is a long-established and vibrant field in computer science, especially in digital library research [28]. It's also possible to find this kind of data in formal politics. Fowler, et al., [24, 25] constructed a citation network using the opinions written by the U.S. Supreme Court and the cases that cite them. Therefore, all usual citation analysis methods [67, 4] can be applied to this Supreme Court dataset. For example, a recent development in citation analysis is the automatic recommendation of new citations based on the old citations and the contents, using which one can naturally extend Fowler's studies on the Supreme Court.

Another study utilizes the relational data from HROs (human rights international non-governmental organizations) [8]. This study is a good example for showing the three of the standard methods usually adopted in political network analysis. The first

is the utilization of graph theoretical network measures, including Betweenness, Centrality, Closeness, Eigenvector centrality etc., as well as some user-defined problem-specific network measures. Most political network analyses employ these measures as part of the exploratory analysis. Some studies [85, 45] end with fine-grained exploratory analysis, while some others continue to propose hypotheses related to the targeted political problem as the second step, and detect/introduce explanatory variables to test the proposed hypotheses. The tests can be either exploratory or model-based. Only some of the studies [33, 49] apply the third methodology. They want to either justify the discoveries from the network analysis or utilize the discoveries for explaining other phenomena. Usually, they introduce new dependent variables, and fit the network variables into some statistical models to best represent the internal logic/co-relationship between different phenomena. All these three methods can be found in this HRO study.

The goals of social network analysis on web science differ with different target websites. Researchers use different data analysis algorithms for different websites, like Wikipedia, Flickr, E-commerce sites, news sites, forums. Classic tasks in this thread of studies include "Community Structure Discovery" [7], "Folksonomy Construction" [72], "Product Rate Prediction"[61] etc. Sometimes because the size of the network is bigger than the size of the computer memories, or the operations defined on the network require more computation power than the computers can provide, specific algorithm optimization issues arise respectively.

Comparatively speaking, studies in political network analysis usually have broader definition of the networks, and are more structured, rigorous and complete in analyzing the target problems. However, studies in web/computer science [72, 61, 7] have much more flexibility in choosing/designing methods/algorithms for different datasets. In a word, both computer scientists and political scientists can definitely learn from each other to inspire thoughts and strengthen their studies.

2.3 Miscellaneous Techniques and Further Potentials

The application of computational techniques in political science is far from restricted in text/network analysis. Agent-based modeling is an interesting method using computer programs to simulate the actions and interactions of "autonomous agents". An agent could be individuals or groups acting in a restricted/idealized space. Interactions among agents can be simulated using evolutionary programming. It is easy to introduce randomness and simulate evolution in programs that provide a way to inspect, observe and even predict the appearance of complex phenomena. In political science, the phenomena could be policy consequences, election outcome and trade changes, etc. An early paper of Kollman, et al., (1992) [50] best explains the advantages of agent-based modeling. They modeled parties as rationally bounded adaptive actors/agents, and employed different algorithms as representing different behaviors of parties to explore election results. This method is able to create a virtual environment to try even the most impossible hypotheses in theory, and thus helps to avoid the disadvantage of non-experimentality in social sciences. Other interesting applications include simulating terrorists and wars, etc.

We have seen the usage of the "E"-prefix for its popular usage on almost everything. A simple search on Google returns us a lot of "E"s on politics, "E-Politics", "E-Activism", "E-Governance", and "E-Campaigning" — It is hard to give a comprehensive definition of "E-Politics", at least I cannot find one on Wikipedia. Generally speaking, E-Politics studies the power of the internet on real world politics and how politics can be improved/transformed on the internet. E-Politics does not necessarily utilize network analysis or other computational techniques, but it usually inevitably borrow research methodologies from similar computer science branches. Jiang and Xu explored the Chinese government portals to study the status of citizen political participation and government legitimation in China [42]. This study uses the same methodology with the studies of Human-Computer Interaction in computer science. Another interesting paper takes the full advantage of information collected on Twitter [35] to study the public opinion on Obama and his health care reform. In this study,

the authors tracked and collected data like the number of click-throughs of some Twitter profiles, trends in the distribution of Re-tweeted messages; and information obtained from other Web 2.0 media like Youtube, Facebook can also be analyzed in this way.

The domination of Google on the internet is studied in the well-cited paper [37], which is also the creator of the word "Googlearchy". This paper is an outcome of the collaboration of one political scientist and two computer scientists; it proves to be a solid interdisciplinary study with its usage of advanced computer science methods like web crawler and a machine learning classifier (SVM), and its comprehensive political analysis relying on the experiment results. As the authors pointed out, "Though comprehensive analysis of link structure requires software and hardware resources not commonly available to social scientists, it is nonetheless much easier and cheaper to perform than alternative approaches such as large-scale surveys", in this paper, Hindman et al. downloaded and analyzed millions of web pages in different categories of political information, it would be an impossible task without the help of computer algorithms. They discovered the "Googlearchy" phenomena, that "political information may remain highly concentrated even in the online world" as "dominated by a handful of heavily-linked sites". The authors continued to explore the impact of "Googlearchy" to politics. They smartly paraphrased Orwell to describe the phenomenon as "on the Web all sites are equal, but some sites are more equal than others"; and provided a detailed analysis of the Web's impact on "media balkanization", "democratic deliberation", and "the competence of ordinary citizens". The analysis sheds light on understanding the political consequences of the information age. Matthew Hindman, the political scientist author has been continuing his study on digital government after this paper.

Mathematical Logic and Game theory as a sub-field of computer science is also utilized in social science studies like in the "Computational Social Choice" [12] theory, especially for Voting theory studies in political science. Clustering methods are used throughout [41] for analyzing the bloc structure of the 2003 U.S. Senate. And discrete data analysis algorithms have also been developed in Roll-Call data studies, like the prominent DW-Nominate algorithm [73], and the Clinton-Jackman-Rivers spatial

voting model [13].

Many new computational techniques are still being developed. An incomplete overview of recent political science literature finds us a handful of early-bird studies. In [40], the authors explored the nature and potential of cloud computing, the policy issues raised; and research questions related to cloud computing and policy. Cederman, et al., use data from the Geographic Information System [10]. The authors constructed and analyzed a dataset of geo-referenced politically relevant ethnic groups, covering the entire world during the period from 1951 through 2005. They show that "the conflict probability of marginalized groups increases with the demographic power balance compared to the group(s) in power." And "the risk of conflict increases with the distance from the group to the capital, and the roughness of the terrain in the group's settlement area". In [51], the authors used Global Positioning System (GPS) for sampling the migrants. Although the paper focuses on the comparison of sampling methods and analyzing the survey results, they showed the advantage of sampling in GPS, as well as the potential to use data from the IT industry to enhance political science studies.

As a large scale dataset becomes more and more available, and the cost for learning and implementing computation tasks is significantly reduced, and as the research interest of social science moves from individuals to societies [54], computational methods have the potential to be applied more and more in political science studies. New technologies, such as Video Surveillance, Image Recognition, Distributed Systems, Semantic Web, although not seen a lot in the previous political science studies, will definitely be used for addressing political issues. Especially, technical developments on Data Mining, Machine Learning, and Information Retrieval show a great future for interdisciplinary studies, because they provide a comprehensive way of investigating data, organizing information, and generating knowledge.

Chapter 3

Political Spectrum Analysis

3.1 Introduction

We live in a world overflowing with opinions and we can easily be tempted to simplify the opinions we receive and label the opinion holders directly as "black" or "white", "left" or "right", without giving an accurate and comprehensive measurement of the opinions. The goal of "opinion mining" is exactly to help us understand the opinions surrounding us better, to discover the unseen, and to explain the inherent complexity of opinions. However, the ubiquitous nature of opinions decides their complexity of dimensionalities. In reality, opinions are always complicated and composed of multiple perspectives: We seldom choose our positions relying on one single aspect; we actually make decisions under comprehensive considerations. "Opinion mining" helps to discover and reproduce opinions without losing their explanatory power of the world.

Most existing researches in opinion mining focus on some uni-dimensional issues, e.g. in product/customer review [93] [92], they usually feature in polarity analysis on the positive-negative dimension or the objective-subjective dimension; In political opinion mining[19] [18], computer science researchers usually simplify the opinion space into the single left-right or liberal-conservative dimension. They all have the problem that the single dimension bears restricted capability to explain the realistic opinion distribution.

In political opinion mining, opinion space is even more multidimensional. Single dimension of "liberal" and "conservative" works fine when all we want is the results of a binary classification; but challenges lie in that many realistic political phenomena, such as a Conservative votes for a Democratic candidate in the previous presidential election, cannot be explained in this simplified model. Aware of that, political scientists proposed the concept of Political Spectrum, which refers to a multidimensional opinion space where each geometric axis models one political dimension, and each dimension represents one importance perspective of political ideology. Examples of the so-said dimensions include the traditional left vs. right, free trade vs. protectionism, contrary attitudes towards personal freedom, different religious beliefs, etc. And 2-Axes or 3-Axes coordinate systems based on these dimensions were also proposed as possible representations of the political opinion spectrum.

Congressional, judicial, and presidential opinion domains have always been the focuses of study in political science since they play the role of three supporting poles of the "separation of powers"; and there are also numerous well-organized data published by the authorities in these domains. The civil opinion domain, however, is comparatively less studied. The reason is not only because of the lack of high-quality data, but also because that political orientations and party affiliations are usually not clearly expressed in the civil domain. In this study, I take advantage of congressional data to analyze the editor-tagged personal blogs. I apply the regression models learned from the congress dataset to the blog dataset, and evaluate the learned political standing scores of blogs comparatively. The experimental results show that my proposed methods retain the explanatory power after transferring from the congressional domain to the blog domain.

In this study, I also formally identify and define the problem of political spectrum analysis, compare it with the other sentiment analysis and opinion mining tasks, and justify the importance of multiple political dimensions, measuring it against the explanation power of formal political behaviors such as voting records in the congresses. I answer the questions of why we need political spectrum analysis and how better-off it will be when more dimensions are brought into the opinion analysis space. I introduce the DW-Nominate scores, the dominant quantification method

of political standings in political science, and take them as the target variables for the regression model. The method helps to test the correlation between the speech records of politicians and the voting records of them. I discuss the differences of feature selection between political spectrum analysis and normal opinion mining tasks and prove that political standings, or political ideologies, are reflections of the opinion holders' attitudes towards various political issues, which indicates the rationality and feasibility of taking probabilistic topics as features in the regressions.

3.2 Related Work

3.2.1 Sentiment Analysis and Opinion Mining

Sentiment Analysis slightly differs from opinion mining [70] in that it focuses more on issues related to natural language analysis like subjective-objective detection, negative-positive detection, term orientation identification [20] [88] [14] and other special linguistic phenomenon [87] [82]. Specialized sentiment analysis requires more than just computational techniques and in-depth understanding of the linguistic background in the targeted context is usually the prerequisites for a good study. In the context of political spectrum analysis, this is also a basic requirement.

Instead of just indulging in the beauty of linguistic phenomenon, researchers are more interested on the application of sentiment analysis, and attempt to find real knowledge through the sentiment analysis process, namely to mine opinions. The most prevalent combination of sentiment analysis and daily applications is in the domain of E-commence, where Customer Reviews [39] [56] and Product reviews [93] [92] become the best carrier for computer scientists to exert their wisdom and help promoting the development of online business. The most common way to organize the discovered knowledge is through summarization and integration [59] [58], while sentiment-enriched topic modeling [62] [84] [55] became the state-of-the-art technique of opinion representation in recent years due to its effectiveness in capturing the hidden semantics in the opinionated texts.

The charm of opinion mining lies in that opinion exists ubiquitously in the human

world and there are always demands for some specific opinion mining tasks. Pang and Lee [69] explored how to rate reviews with stars; Liu et al. [57] challenged predicting sales performance through opinion mining on blogs; Jindal and Liu [43] was attracted to study the problem of opinion spam detection. Political opinion mining [18] is also a fashion toy for lovers of opinion mining, Adamic and Glance [2] analyzed the network structure in the political blogosphere; Efron [19] dug into conspiracies between the liberal media and the right-wing.

3.2.2 Political Spectrum Analysis

Political Spectrum Analysis originates from the concept of spatial model in political science proposed by Downs [17] in 1957. The model holds the assumption that each voter will vote for the candidate or party that is closest to his or her political position, which is the essentially the incentives of political spectrum analysis for political scientists. Political spectrum theory evolved in the following years and many explanations of the dimensions were proposed in attempt to more precisely describe the complete picture of political orientations and ideologies. Greenberg [29], as a recent example, posited a model comprising the standard left-right axis and an axis representing ideological rigidity. And Grendstad [30] focused on the role of cultural and individual characteristics in the decision of political orientations while surveyed their findings in the Europe. Poole and Rosenthal [74] brought roll call analysis into political spectrum analysis, indicating a transition in methodology from qualitative to quantitative analysis.

Political scientists also brought text analysis into their studies to extract domain knowledge they need. Laver et al. [53] published a famous paper featuring in extracting policy positions from political texts using word count information. Monroe et al. [63] further discussed the lexical feature selection problem in political analysis, especially proposed the Bayesian Shrinkage and Regularization method for bipartisan analysis. Diermeier et al. [16] analyzed the ideology in Congress using legislative speech records in the U.S. Senate. My work is most similar to this, but the primary target of their paper is the classification problem in the left-right dimension, while

mine is the regression models on multiple dimensions and I also issue the domain-transfer problem. Yu et al. [90] [91] also explored the characteristics of opinion expressions for political opinion classification but they focused more on the sentiment analysis of political corpus without considering the spectrum and ideology problem.

3.3 Problem Formulation

3.3.1 Political Spectrum

The spectrum of political opinions is far more complex than what can be captured by a few simple labels: conservative or liberal, rich or poor, and even good or bad. The complexity of political spectrums is ultimately the source of uncertainty and unpredictability in real politics. There are multiple dimensions that dominate our political perceptions and decisions, e.g. one could be a social liberal but a financial conservative, or a domestic "dove" but a diplomatic "hawk".

We usually construct our perceptions of political standings of either our friends or politicians, from our general interpretations of their talks and political activities. It is always "labor-saving" to simply label others as liberals, conservatives or environmentalists and then judges whether or not they are in the same trench with us. But also as the saying goes, oversimplification causes misunderstandings: sometimes we might want to know the exact coordinates of somebody in the political spectrum. So the questions come as: Whether we can determine political orientation automatically, especially when available data is restricted? Can we compare political opinions of different people or social groups numerically? And are their textual records useful enough for mining, or more behavioral records are necessary to be included?

In this section, I formally define the key concepts in Political Spectrum Analysis.

Definition 1 (Opinion Entity): An opinion entity E could be an individual, a public media, a social organization or a website which expresses its opinions to the public represented in a collection of document C = C$\{d_1, d_2, ... , d_n\}$, where d_i is one

of this opinion entity's retrievable textual records composed of a bag of words T$\{t_1$, t_2, ... , $t_m\}$. Opinion entities share a hierarchical structure: they either contain or belong to other entities, e.g. an individual as an opinion entity is also part of a higher level opinion entity: the Nation. Examples of opinion entities include legislators, blog authors, news sites, parties, and nations.

Definition 2(Political Spectrum): The political spectrum S is a multidimensional coordinate system where each geometric axis models one political opinion dimension and each opinion entity is positioned in the spectrum as a point according to its political orientations on each dimension. Given k opinion dimensions, S is defined as $\{D_1, D_2, ... , D_k\}$ where D_i is one dimension and the coordinates of an opinion entity E is defined as $S_e\{s_1, s_2, ... , s_k\}$. S also yields to a weight vector W$\{w_1, w_2$, ... , $w_k\}$ constraining the strength and significance of each dimension which primarily takes role in the calculation of distances between opinion entities.

Definition 3(Domain Transfer in Political Spectrum Analysis): Domain-Transfer problem is a common problem in NLP-related tasks due to the heterogeneous nature of human language in different context. In political spectrum analysis, Domain-Transfer problem refers to the different habit of term usage and linguistic expression among different groups of stakeholders in politics. Examples include spatial and temporal divergences among people in different region and era, or the colloquial-literary divergence between politicians and other individuals.

3.3.2 DW-Nominate Score

Political scientists have studied the political standings of electors and politicians through different ways according to their different degree of involvement in the real

politics. For electors and citizens, they[1] [2] [3]conduct polls or questionnaire-based opinion surveys designed to score participants according to their responses to the questions. And politicians, especially legislators are usually rated through some theory-based scoring procedures with respect to their roll-call voting records in the congress.

One dominant scoring system in the political spectrum analysis, which is widely used in academic political science studies, is the DW-Nominate scores developed by Poole and Rosenthal[74]. It features in estimating the position of legislators in the United States Congress using a weighted utility model of their roll call voting records. This score is presented in two dimensions: the first dimension is the well-know liberal-conservative dimension, usually interpreted as the various attitudes of legislators in government intervention of the economy; the second dimension is a weighted dimension compared in significance to the first one with dynamic weights varying along congresses. It used to be more significant in the early days of the United States before the Civil War representing the North-South divergence; and now the left-right dimension has become more dominant so the second dimension is more like a supplementary dimension today, with the weight 0.4163 compared to weight 1.0 of the first dimension, picking up political standings in the view of traditional values, social lives, and civil rights. I want to prove in this study that correlation exists between the political addresses and political behaviors. And opinion mining in opinionated political corpus not only provides a way to predict the formal political behaviors of politicians like roll-call voting in the congress, but also helps to probe and explain the political orientations of other individuals like bloggers even when there is an endogenous shortage of their behavioral records.

Also I want to show that due to the diversity of opinions in the political spectrum, it is not enough to just attribute legislators into a uni-dimensional measurement such as the straight left-right or liberal-conservative wings. I provide a dimensionality analysis in the next section to show that the two-dimension scores outperform the single dimension scores in the capability of explaining the roll-call voting records measured in the coupling degree of distance matrix.

[1]http://www.politicalcompass.org/
[2]http://politics.beasts.org/
[3]http://politicalsurvey2005.com/

3.4 Dimensionality Analysis

In this section, I provide three examinations of the Roll-Call voting records and the DW-Nominate score to justify the existence of multidimensional opinion space in the congressional domain and thus the necessity of political spectrum analysis.

I collect the necessary data for dimensionality analysis. I get the DW-Nominate scores for the 100 senators of the 110th United States Senate from the voteview website. In this data, each senator has a pair of scores for each dimension which are ranged in [-1.0, 1.0]. I scale the scores to [0.0, 1.0] and take them as the dependent variables of the regression models. Figure 3.1 illustrates the distribution of DW-Nominate scores in the Senate: the red squares in the figure are republicans, blue circles are democrats, and yellow triangles are independent senators. The two adjunct histograms count the number of senators in a series of ranges on each dimension. We can see that the senators are distinctly divided into two camps on the first dimension while diversely distributed on the second dimension.

I also collect the roll-call voting records of each senator during the 110th Senate sessions, which is composed of 442 voting of bills. For each bill, the value 1 means the senator voted "Yea", -1 means the senator voted "Nay", and 0 means the senator was absent in the voting. So the voting records are presented in a vector of 442 elements. I thus construct the bill-senator decision making matrix for this data.

3.4.1 Principal Component Analysis

Principal component analysis (PCA) finds a smaller set of synthetic variables called principal components from the original set of variables, the high-dimensionality of variables is reduced but the variance information is retained in the new set. Principal components are extracted in decreasing order of importance measured in the proportion of variance, and are orthogonal to each other. I apply PCA to the roll-call voting record, taking the voting of bills as observations, and the senators as variables. The goal of this analysis is to find out what voting patterns exist in the record, and the patterns guide the voting choices of the senators. Thus linear combinations of the senators are expected to be found out in which each senator plays either a positive or

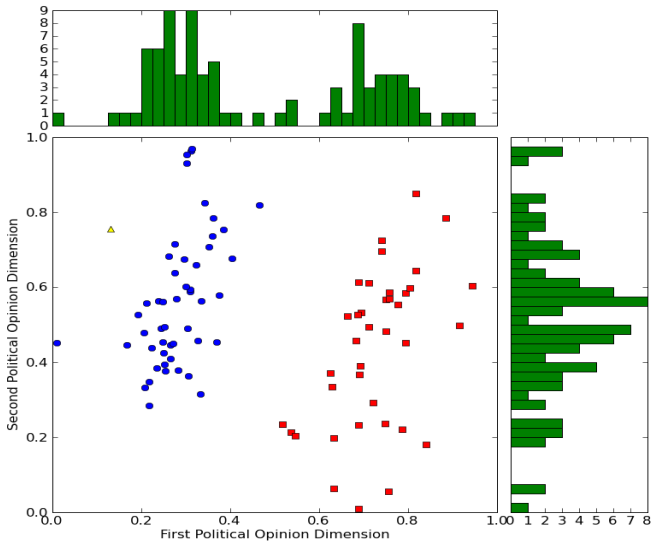


Figure 3.1: DW-Nominate Scores of senators

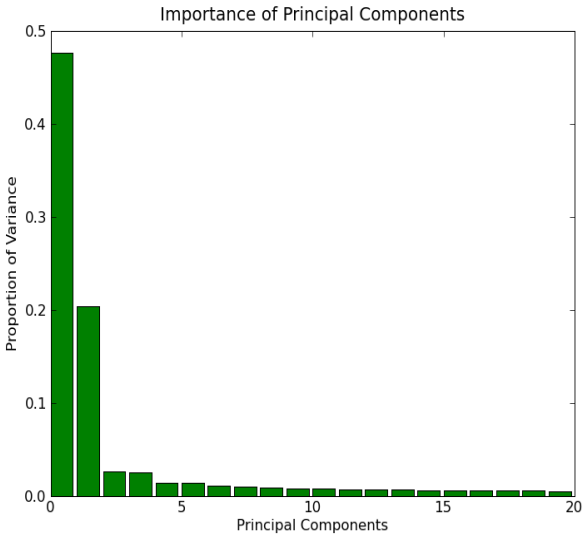


Figure 3.2: Principal components and their proportion of variances

negative role towards the final pass of the bill. So each selected principal component of this data set is exactly a linear combination of the senators which represents one voting pattern, that is, one perspective on which the senators hold to decide their standings of the bills.

Figure 3.2 shows the discovered principal components and their proportion of variances. The first principal component yields a proportion of 0.476, and the second principal component yields a proportion of 0.204, and the proportion of variance diminishes rapidly from the third principal component. The cumulative contribution rate of the first two principal components is 0.680.

The results justify the existence of the multidimensional political spectrum in the Senate. The first two components, explained as two voting patterns, are especially significant compared to the others although they still cannot provide complete coverage of the spectrum. Here note that one interpretation of the first dimension is the left-right voting pattern, and the second dimension is the social rights voting pattern, but different interpretations are still possible to exist and be reasonable, while interpretations of the other small principal components become more and more difficult and usually cannot bring any more benefits for the studies.

3.4.2 Coupling degree of distance matrix

I design the Coupling Degree of Distance Matrix ($CDDM$) as the criteria to test the explanatory power of the dimensional scores on the roll-call voting records.

Suppose we have n legislators, m voting of bills, thus the voting record matrix V is n*m dimensional, V_{ij} is the voting choice of legislator i for the j_{th} bill:

I define all the distances in the Normalized Euclidean space in $CDDM$.

- First, I calculate the voting record distance matrix VD. I define the distance between legislator A and legislator B in the voting records context as the Normalized Euclidean distance between V_a and V_b. So VD is n*n dimensional,

where VD_{ab} is the distance of legislator A and B.

$$VD_{ab} = \frac{\sqrt{(V_{a1} - V_{b1})^2 + (V_{a2} - V_{b2})^2 + ... + (V_{am} - V_{bm})^2}}{\sqrt{2 * 2 * m}}$$

- Second, I calculate the uni-dimensional score distance matrix U1 for the first dimension and U2 for the second dimension. U1 and U2 are also n*n dimensional, where $U1_{ab}$ is the Normalized Euclidean distance between two legislators in the first dimension and $U2_{ab}$ is the Euclidean distance between two legislators in the second dimension.

$$U1_{ab} = \sqrt{(U1_a - U1_b)^2}$$

$$U2_{ab} = \sqrt{(U2_a - U2_b)^2}$$

- Third, I calculate the two-dimensional score distance matrix S for the DW-Nominate scores in two dimensions. Since the two dimensions are not exactly equal in the significance to describe political opinions, I give the second dimension a weight to balance the Euclidean space. Suppose two legislators A and B have scores (a1, a2) and (b1, b2). With weight w, the distance between A and B is:

$$S_{ab} = \frac{\sqrt{(a1 - b1)^2 + w^2 * (a2 - b2)^2}}{\sqrt{1 + w^2}}$$

- Fourth, I calculate the Mean Squared Error σ^2 between U1, U2, S and VD.

$$\begin{aligned} \sigma_{u1}^2 &= \frac{\sum(U1_{ij} - VD_{ij})^2}{n^2} \\ \sigma_{u2}^2 &= \frac{\sum(U2_{ij} - VD_{ij})^2}{n^2} \\ \sigma_s^2 &= \frac{\sum(S_{ij} - VD_{ij})^2}{n^2} \end{aligned}$$

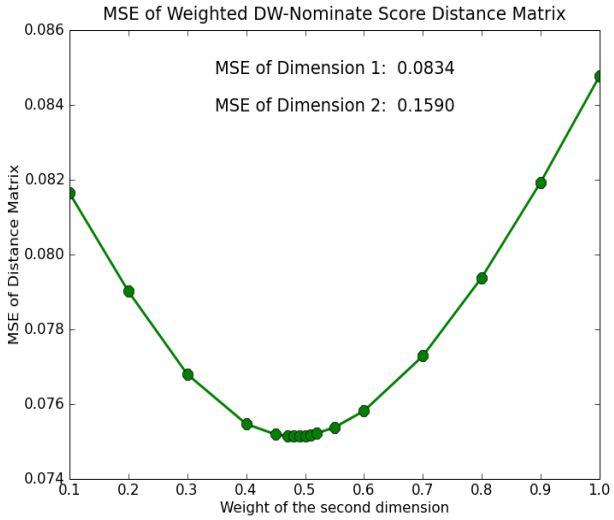


Figure 3.3: Weighted Distance Matrix

The MSE σ_{u1}^2 between U1 and VD is 0.08339, and the MSE σ_{u2}^2 between U2 and VD is 0.15897. I tuned the weight w from 0.0 to 1.0 to find the best w with the lowest MSE for S and VD.

Figure 3.3 shows that when w is 0.48, the MSE $\sigma_s^2 = 0.07514$ is the lowest. Note that σ_s^2 is also much lower than σ_{u1}^2 and σ_{u2}^2 which proves that two dimensional scores outperform any single dimension in the coupling degree with the roll-call voting records.

3.4.3 Correlation of two dimensions

The Pearson correlation coefficient is a measure of the strength of linear dependence between two sample variables. It is always between -1 and 1. I calculate the Pearson correlation coefficient for the DW-Nominate scores in the two dimensions, it is -0.16326. This shows that the two dimensions have low or no association and for the purpose of this work, I consider them orthogonal.

3.5 Regression Models

It is a basic assumption of this study that the secrets of the senators' political standings and voting choices hide behind the languages they use. In order to achieve the best regression effect, we need to first answer one question: what is the difference between a general political ideology and a concrete political opinion?

For a concrete political opinion, it is similar with the attitude towards a specific product in business opinion mining, which is usually modeled in the opinionated words used in the product review, and the opinionated words are usually adjectives or adverbs. For example, a concrete political opinion could be "to support or oppose a tax reduction", or "to prefer which candidate", and a customer attitude could be "to suggest buy this product or not", or "to choose which product", they are similar in that they all involve binary or multiple choices which are comparatively easy to be detected only with opinionated words.

A political ideology, or a political spectrum as we call it in this paper, is not only an issue of choices. It is the underlying guideline for political opinion holders, which forms during a complicated process of interaction between numerous psychological and social elements. It can be taken as built up upon many concrete political opinions. Those ideology surveys are also derived from the thinking that a person's political spectrum can be detected through integrating his or her responses of several political issues. So although the textual records have not been human-coded into several political issues, we can take advantage of the state-of-the-art probabilistic topic modeling to help detecting the political spectrum.

I also use the Part-of-Speech tagged words as features to compare with the topic-features and to test the special role of nouns and verbs in the task of political spectrum analysis.

3.5.1 Lemmatization and POS Tagging

Unlike other opinion mining tasks, the definition of "Opinion" in the political spectrum analysis is special. Opinions in this context are not necessarily directly detectable based on the subjective-objective lexical feature selection methods. Instead,

any words, either subjective or objective, noun or adjective, could be the indicator of political orientations. For example, nationalists might be more likely to cite the name of the historical national heroes and wars in their addresses, and environmentalists might be more likely to talk about greenhouse gases, animals, forests with elegant literary style. So what is the most powerful feature set for political spectrum analysis? And are adjectives still the dominant opinion indicator as in other opinion mining tasks? I address these questions in the experiments.

I run Lemmatization instead of Stemming on the feature set. Stemming usually results in a chop-off of the ends of words into the stem form which is usually not even a real word. It helps to sum up derivatives but inevitably loses the part-of-speech information which is crucial in our analysis. Lemmatization does a similar job with Stemming but will return the dictionary form of a word and save the part of speech information. Examples of lemmatization include:

$$\begin{aligned} am, are, is &\Rightarrow be \\ better, best &\Rightarrow good \\ striking &\Rightarrow striking \\ loved, loves &\Rightarrow love \\ cars &\Rightarrow car \end{aligned}$$

I use the MontyLingua[4] Python Natural Language Processing library for lemmatization and pos-tagging in this study. I select lemmas tagged as Noun, Verb, Adjective or Adverb as four of the candidate bag-of-word feature set, and all the four sets together as the fifth feature set.

3.5.2 Topic Features

Probabilistic topic models are based upon the idea that documents are mixtures of various topics, where a topic is a probability distribution over words. In Latent Dirichlet Allocation, the document-topic distribution is assumed to have a Dirichlet

[4]http://web.media.mit.edu/ hugo/montylingua/index.html

prior, while the assumptions vary in different models. Topic Modeling helps to capture the more subtle latent semantic information of the documents, and the inferred parameters of the document-topic distribution can be easily transformed to be the features of the regression model built on the same corpus.

A similar idea of the topic features is to use the clustering of words as features in other text mining tasks. In [3] [71], the authors use distributional clustering of words as features for text classification. In this study I use LDA to generate topic features for the regression, which is a more fine-grained semantic detection tool than the distributional clustering method, and also I use the features for a regression task instead of text classification. As far as I know, this is a novel method being introduced in opinion mining, partly because of the unique nature of political spectrum analysis.

I ran the LDA tool of Mallet on the senators' speech record dataset, and controlled the number of topics to 20, 50, 100, 500, and 1000 respectively. For k topics, I get the topic distribution of the senators in a vector of k elements of probability numbers, and I take the k elements as k features, while they always sum up to 1.

3.5.3 Regression Algorithms

The choice of regression algorithm is also derived from the nature of the regression problem. I have a large bag-of-word feature set with more than 10,000 lemmatized words ranked by their TF*IDF scores, and a topic feature set with up to 1000 features. At the same time, the training set is quite small since there are at most 100 senators in the Senate, and I get at most 100 training instances. This problem restricts the choice of regression models.

Although I aim at the regression of the DW-Nominate scores, the task shares some properties with traditional text classification tasks. Thus it is natural to choose the Support Vector Regression (SVR) as the primary regression algorithm. SVR shares the same mathematical foundations with SVM, thus the model produced by SVR only depends on a subset of the training data because outliers that lie beyond the margin controlled by threshold ϵ will be ignored to contribute to the cost function. It helps to reduce the noise information in the dataset. Also SVR handles the potentially

non-linear relationship between features and scores with the usage of different kernel functions.

I experiment on three kernel functions. The Radial Basis kernel function (RBF), the Polynomial kernel, and the linear SVR. The hyperparameters of them are trained through cross-validation. I use the SVR tool provided by libsvm, and I also use Multiple Linear Regression (MLR) as a baseline model. Experiments are repeated on five lexical feature sets and 5 topic feature sets.

Chapter 4

Experiments

4.1 Data Collection

I collected four different types of data in this study:

- DW-Nominate Scores of senators.
- Roll-Call voting records of senators.
- Floor statement records of senators.
- Political Blogs tagged in different political orientations.

Congressional opinion space

I first downloaded the DW-Nominate scores for the 100 senators of the 110th United States Senate. Each senator has a pair of scores for each dimension which range in [-1.0, 1.0]. I also downloaded the voting records for each senator, which is composed of 442 voting of bills. For each bill, the value 1 means the senator voted "Yea", -1 means the senator voted "Nay", and 0 means the senator was absent in the voting. So each senator's voting record is presented in a vector of 442 elements, and the data is organized as a 442*100 bill-senator voting record matrix.

Statements	15512
Stetements per senator	176
Token Count	3320283
Tokens per senator	37730

Table 4.1: Statistics of the floor statement dataset

State	Senator	Statements	Tokens
Arizona	MCCAIN	708	105212
Massachusetts	KERRY	459	81856
Kentucky	MCCONNELL	139	17773
Pennsylvania	SPECTER	136	74859
West Virginia	BYRD	23	8225

Table 4.2: Samples of the floor statement dataset

I also crawled the floor statement records of each senator through the Project Vote Smart website[1], which is a non-profit organization that provides detailed information for all the incumbent members of pubic offices in the United States. Floor statement is the best speech records representing the political orientation of senators because it is consistent among senators and is regularly and formally given in the Senate. The basic statistics and samples of this crawled floor statement dataset are summarized in table 4.1:

Table 4.2 shows samples of the floor statement dataset.

I then remove those senators who are no longer incumbent from the senator dataset, since their floor statement records are usually not publicly available after their retirement. I also remove those who did not serve the full term of office to reduce the absence of voting rate in the dataset. The remaining dataset features 88 senators who have at least served 6 years in the Senate.

The three datasets above contribute to the study on the congressional opinion space. With them I want to prove that the formal political behaviors of legislators are to a certain degree explainable by their public speeches using the regression models while the measurement is based on the two dimensional DW-Nominate scores.

[1]http://www.votesmart.org/

Catalogue	Blog Number
Anarchism	63
Democratic	1,358
Green Politics	99
Humor	206
Independent	1,114
Liberalism	181
Libertarian	666
Moderate	159
News Only	266
Religious	165
Republican	1,059

Table 4.3: 6,373 Political blogs cataloged by BlogCatalog

Internet opinion space

I want to study the domain-transfer problem in the context of political spectrum. Politicians have their own language. Even if they share the same political orientation with their supporters, they do not speak in the same way as the electorate. In political science, the transferring of opinions among domains can be modeled as the result of two phenomena: the shifting phenomenon and the scaling phenomenon. The shifting phenomenon refers to the move of opinion centre among domains, e.g. the opinion centre varies dramatically among countries with different social systems and religions. The scaling phenomenon refers to the change of scale when the domain changes, e.g. the electors in one state are more diverse in the distribution of opinions then the electors in another state. So one interesting problem in this paper is to compare the formal and informal political opinion space and observe the shifting phenomenon and scaling phenomenon.

I crawled the personal political blog data from Blogcatalog[2], a social blog directory that maintains lists of tagged political blogs according to their contents. It is of my interest to see the applicability of the regression models when transferred to the informal blog opinion space. BlogCatalog lists 6,373 political blogs at the time I accessed it, of which I selected and crawled 88 blogs tagged in 8 categories with

[2]http://www.blogcatalog.com/directory/politics/

Feature Set	SVM Regression			MLR
	RBF	POLY	LINEAR	
Noun	0.05178	0.06553	0.05292	0.09027
Verb	0.05590	0.05820	0.06338	0.20058
Adj	0.04723	0.06323	0.06313	0.93407
Adv	0.05748	0.05814	0.08886	0.23944
All	0.05107	0.06626	0.05077	0.09007

Table 4.4: Regression Performance on the first dimension

Feature Set	SVM Regression			MLR
	RBF	POLY	LINEAR	
Noun	0.04109	0.04673	0.04384	0.09760
Verb	0.03992	0.04403	0.04512	0.17667
Adj	0.03891	0.04705	0.09479	0.51629
Adv	0.04387	0.04543	0.06793	0.33888
All	0.04066	0.04782	0.04062	0.09130

Table 4.5: Regression Performance on the second dimension

different political identity. The tags I use are: Anarchism, Democratic, Green politics, Independent, Liberalism, Libertarian, Moderate, and Republican. Table 4.3 contains the tagging statistics of these blogs.

I filtered the blog data with HTML tags removal and Noise webpages reduction. I than extract out the word features from the blogs with respect to those used in the learned regression models, and inferred the blog-topic distribution of the blogs based on the trained topic models of the congressional data, using the topic inference tool provide by MALLET.

4.2 Experimental Results

4.2.1 Floor Statement Dataset

I run the Leave-one-out cross-validation on both dimensions separately. The regression performances on the two dimensions are summarized in Table 4.4 and Table 4.5

Number of Topics	Score 1	Score 2
1000	0.0578449804565	0.0422864402028
500	0.0558501167672	0.0414904845701
100	0.0459193126132	0.0395444962401
50	0.0438124975075	0.0353361251982
20	0.0476832813083	0.0352560522689

Table 4.6: Scores for different number of topics

respectively. I use two evaluation criteria here to measure the performance of the regression models. The first is the Mean Squared Error, which measures the average of the square of the difference between the estimator and the quantity to be estimated. And the second is "Rosset Ranking"[78], a ranking-based evaluation method specially designed to evaluate regression models by Saharon Rosset etc. "Rosset Ranking" calculates the number of ranking order switches among the estimated results and the true values of the testing set. In this study, ranking is based on the legislators' ideal points on each dimension. The result of this evaluation is presented in Table 4.7, it shows consistent evaluation results with the residual-based measure of mean squared error.

As shown in the experimental results, the regression models achieve good performances in estimating the senators' DW-Nominate scores on both dimensions using their floor statement records. I also have the following observations:

- Nouns play an important role in the regression; they steadily achieve small MSE and even outperform all the other three separate feature sets in SVR-Linear and MLR models. This proves that nouns are indispensable in the political spectrum analysis as the key indicators of political orientations. People are different in their political standings not simply because that they have different way of expressing political beliefs, but rather because they don't share the same interests and topics at all at the very start.

- Adjectives bear horrible results in MLR but magically return to normal in the SVR models. This is quietly possibly due to the existence of outliers in the Adjective feature set. As discussed above in this paper, personal styles lead the

Number of Topics	Dimension 1	Dimension 2
1000 topics	0.692	0.456
500 topics	0.743	0.540
200 topics	0.761	0.650
100 topics	0.759	0.548
50 topics	0.756	0.620
20 topics	0.661	0.634

Table 4.7: Rosset Ranking Scores

way these legislators express themselves, this brings noise to the training set for simple regressions which screws up the MLR. But in support vector based regressions, outliers outside of the decision margin are easily got rid of.

- Three kernel functions achieve almost the same performances. RBF is the best and most steady one, while polynomial function is less supported by the results. This reveals the tendency of equality for all the opinionated words in the feature set. And SVR-Linear proves this regression problem is still near linear separable. This is also consistent with our knowledge that when the number of instances is far less than the number of features, the map of data to a higher dimensional space is not as necessary as in reverse.

- All the lexical features together beat any separate ones. This result points out the effectiveness of feature combinations and indicates that semantic structures have a potential to facilitate the political spectrum analysis.

- The topic feature sets are tested in SVR-Linear Kernel. Table 4.6 shows that topic features achieve a steady performance enhancement to the lexical feature sets, and the best selection of the number of topics for the congress data is from 20 to 50. It reveals that the congressional political spectrum is concentrated on a comparatively small number of topics, e.g. health care, tax, energy, education, homeland security, crime, international relationship etc.

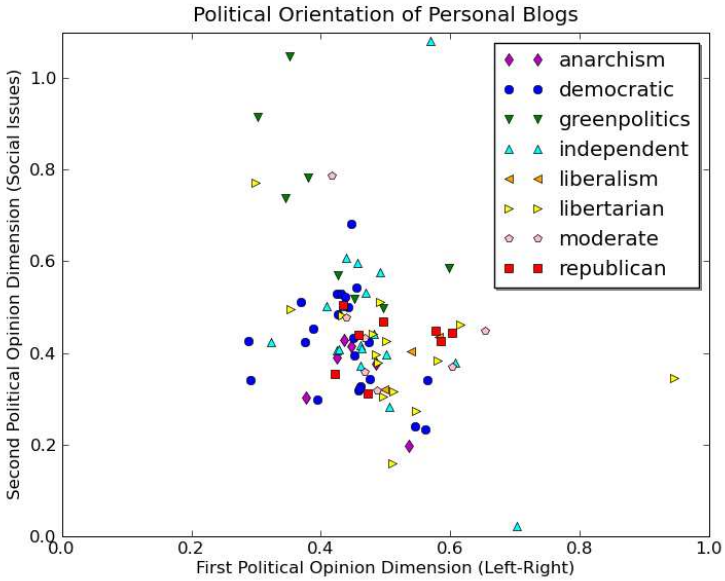


Figure 4.1: Blog Space

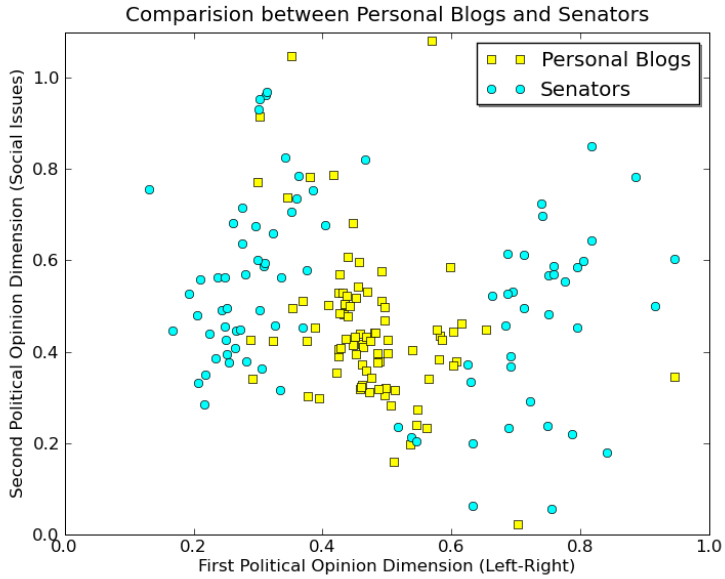


Figure 4.2: Mixed Space

Category	Score 1	Score 2
Anarchism	0.450894393733	0.35155748002
Democratic	0.438007263446	0.418215388959
Green Politics	0.418983545838	0.707009923273
Independent	0.479285712262	0.464233317667
Liberalism	0.539680934277	0.386308085369
Libertarian	0.516037775523	0.409941506657
Moderate	0.504935502215	0.456154045361
Republican	0.506026619653	0.42424342101

Table 4.8: Scores in each category

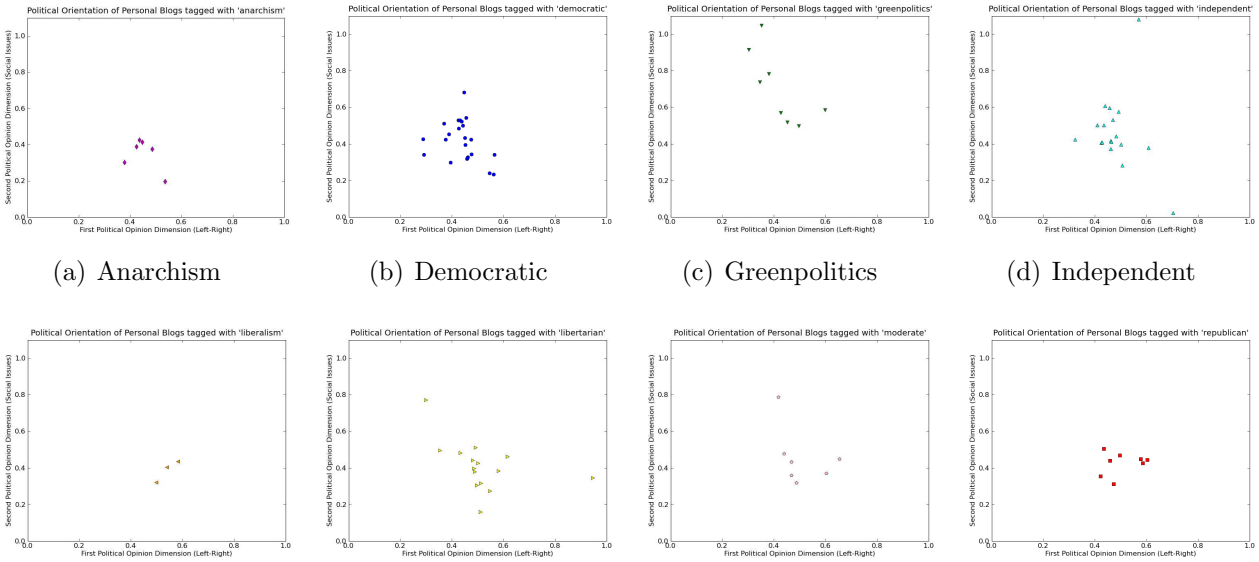
4.2.2 Blog Dataset

The blog dataset is a mirror copy of Blogcatalog, which was selected without any specific preferences, thus it is a reliable and representative sample for the political spectrum on the United States cyberspace. As shown in Table 4.3, these blogs are tagged in several categories that include the liberal-conservative dimension and the environmentalist, religion dimension etc. Also there are some independent, moderate or news blogs which might not contain opinionated contents.

What is interesting to us is the applicability of regression models learned in the formal political corpus when the domain transfers to an informal context like the cyberspace. Also I want to observe the distribution of political spectrum in the cyberspace.

I crawled all the blog entries of the 88 blogs and cleaned the data before the regression. These blogs are tagged by the editors of Blogcatalog in ideology, thus providing an easy way for us to evaluate the effectiveness of the regression on the blog dataset. Here the difficulties of evaluation lie in the lack of information to the political standings of the blogs on multiple dimensions. For a text classification task, it might not be hard for a human reader to tag the blogs and thus using the blogs as testing dataset. But in this study, the political spectrum scores are multidimensional and continuous. It is beyond our ability to manually assign any target scores for them.

So I adopt a comparative evaluation. I plotted the distribution map for each



(a) Anarchism (b) Democratic (c) Greenpolitics (d) Independent

(e) Liberalism (f) Libertarian (g) Moderate (h) Republican

Figure 4.3: Blogs tagged with different political orientations

editor-tagged category in Figure 4.3, and calculated the average scores of each category in Table 4.8. Although we don't know the political standings of each single blog, we do have some prior knowledge of each category. For example, we know the republicans should have bigger scores on the first left-right dimension. I put together all the blog scores into one blog score space in Figure 4.1 and also integrated it with the Senate score space in Figure 4.2.

I also experimented on both the lexical feature sets and the topic feature sets. In contrary to the congress dataset, in this blog domain the lexical feature sets outperform the topic feature sets. This is caused by the topic differences in two domains. As I discussed above, the political spectrum in the congress domain can be best modeled with 20 to 50 representative topics. But in the blog domain, since the bloggers are not professional politicians, they would inevitably talk about other topics than simply the congress issues. They share a much bigger topic divergence among each other, and usually pay less attention on any specific topics than the legislators. In other words, the topic features are not consistent among domains in the explanatory power of political spectrum. The lexical features also suffer from this problem. But

the information loss on the lexical features during the transfer of domain is less than on the topic features. I choose only the top 1000 words with the highest TF*IDF scores, the 1000-word feature set retains the power to explain the political spectrum after the transfer of domain.

4.3 Experiment Analysis

The ideology distribution of the blogs is consistent with some of our basic knowledge on politics, as we can see from the regression results of the blogs on Figure 4.1 and Figure 4.2:

- Republican blogs are on the right of the democratic blogs on the first dimension.
- Supporters of "green politics" favor strict control on civil rights the most. For them, environment protection always has highest priority, even higher than the rights of human beings.
- The bloggers tagged "moderate" are concentrated in the centre and the bloggers tagged "independent" are scattered.
- The Anarchists have no preference for economic policies, but strongly prefer less governmental and social control on civil rights.
- The Liberalists and Libertarians are even on the right of the republicans on the first dimension, indicating that they are bigger fans of small government.

We observe that the blog ideology positions are concentrated in the middle of the space, filling exactly the empties in the senate opinion space. One explanation of this could be that the bloggers are generally more moderate than most of the legislators or they just do not use their personal blogs as a training camp for their next election — they are just blogging for fun with less opinionated contents. A more pessimistic view of this is that the fitted regression model is not discriminating enough for the blog dataset, i.e. the lexical features picked up from legislative speeches are not the same set of words that the bloggers use to express their policy positions and ideologies;

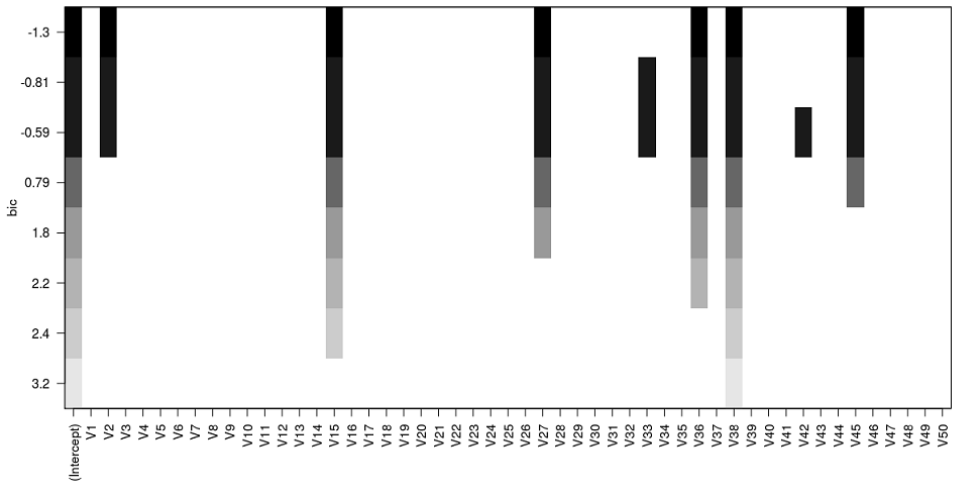


Figure 4.4: Model Selection on the first dimension

and the topic models learned from the words are also not consistent when domain of dataset transfers and language usage habit changes.

It is worth noting that the social dimension is comparatively more diversely distributed and controversial than the economic dimension, even if the regression model has discriminating issues. It is important here to point out again that the regression models are different on the two dimensions although we used the same set of features for them. The reason is that Support Vector Regression has the embedded step of model selection, in which only the most predictive features are selected as the support vectors in the regression model. So this indicates that the bloggers dispute more on social right issues than on economic issues. In other words, model selection is more successful on the social dimension.

In order to look into the details of the topic feature regression, I did model selection for a simple linear regression on the legislative speeches dataset, using Bayesian information criterion (BIC) as the model selection criterion. Since it becomes harder to assign real meanings to the topics when the number of topics increases, I set the number of topics in LDA to 50, and I also set the maximum number of selected features to 10. The model selection results are shown in Figure 4.4 and Figure 4.5 respectively for the two dimensions:

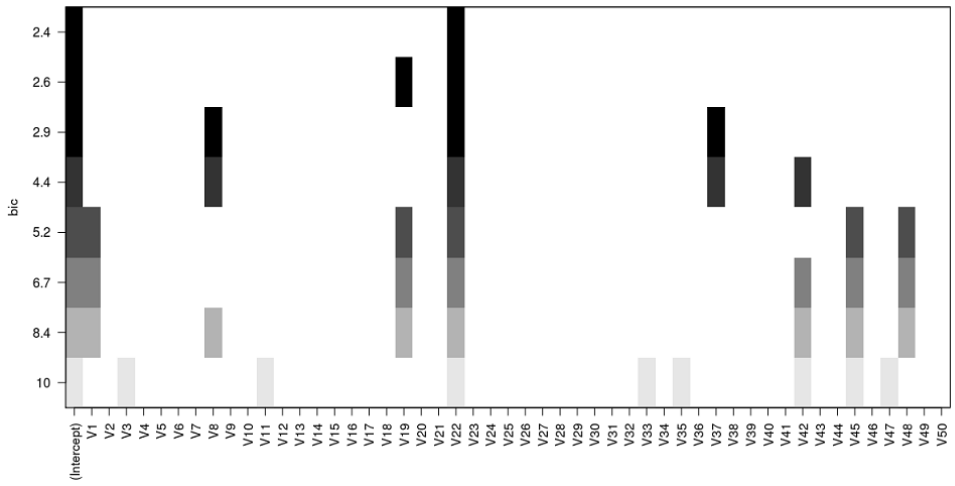


Figure 4.5: Model Selection on the second dimension

For the first dimension in Figure 4.4, BIC achieves the lowest when topics 2, 15, 27, 36, 38, 45 are selected as the predictors. And for the second dimension in Figure 4.5, topic 22 alone gets the lowest BIC. I looked at the word distribution for each of the topics above; they all clearly resemble real issues, correspondingly homeland security, justice nomination, stem cell research, abortion, mine, tax cut for topics in the first dimension and health care for topic 22. I am listing the top 20 representative words for each topic below:

- Topic 2 (Homeland Security): security, homeland, department, weapon, defense, threat, port, attack, chemical, guard, commission, facility, base, secretary, coast, nation, recommendation, force, dh, air
- Topic 15 (Justice Nomination): judge, court, nomination, nominee, president, justice, republican, supreme, circuit, committee, judiciary, district, confirmation, hearing, vote, vacancy, white, majority, record, bush
- Topic 27 (Stem Cell Research): cancer, disease, research, cell, asbestos, stem, vaccine, legislation, treatment, health, victim, blood, flu, compensation, exposure, institute, bill, breast, cord, life

- Topic 36 (Abortion): child, family, woman, parent, care, support, bill, mother, welfare, legislation, life, service, time, home, adoption, dc, colleague, day, abortion, infant

- Topic 38 (Mine): north, reid, dc, nevada, michigan, mine, south, nelson, dakota, waste, korea, harry, stabenow, coal, nuclear, miner, conrad, energy, nebraska, safety

- Topic 45 (Tax Cut): budget, cut, president, tax, spending, deficit, cost, economy, money, fund, increase, funding, debt, dollar, government, American, congress, time, stimulus, plan

- Topic 22 (Health Care): health, care, insurance, coverage, medicaid, cost, american, percent, plan, premium, program, access, benefit, legislation, family, business, employer, child, people, employee

- Topic 8 (guantanamo): attorney, department, justice, administration, law, white, detainee, president, investigation, executive, Guantanamo, counsel, gonzale, official, abuse, rule, war, committee, judiciary, document

- Topic 19 (Procedural Speeches): bill, legislation, break, transcript, resolution, statement, introduced, committee, federal, text, colleague, consent, dc, rise, congress, senator, nation, law, support, percent

- Topic 37 (Climate Change): research, technology, climate, science, development, change, national, world, innovation, investment, ocean, warming, nation, greenhouse, impact, emission, program, carbon, system, america

The topics in the first dimension are all controversial issues that are usually divided among the two wings of congress. Topic 45 (tax cut), 38(mine) are purely economic issues that are expectable as discriminating features for the economic dimension. Other topics (homeland security, justice nomination, abortion, stem cell research) reflect the partisanship status of the first dimension, that most of the ideology points in the first dimension can be explained by partisanship alone. For the

social dimension, fewer topics play as effective features, topic 22 (health care) unexpectedly is the most significant feature, followed by topic 8 (guantanamo), topic 37 (climate change) and topic 19 (procedural speeches). This result reflects that the second dimension is less prominent in the congress, and the legislators' policy positions on a few issues decide their ideology on the social dimension.

In order to improve the performance of the regression models on the blog dataset, some further studies can be explored. First, I need a bigger blog dataset that contains more blog entries that are purely commenting on political topics instead of on the bloggers' personal life. One empirical issue could be that it is hard to find such pure political blogs on the web; then some techniques of blog filtering must be adopted to clean the data. Second, the blogs may only have discussed some topics, that is to say, there are missing topics on the blogs compared with the legislative speeches. It is thus important to make sure that the significant topics in the legislative speeches are not missing; otherwise it would bring big performance loss for the regression models.

I assumed that the choice of words or the choice of topics indicate the underlining ideology by far. This is not necessarily true, especially when speakers of opposite standings fight on some specific issues and overlap a lot in the words they use to illustrate the issues. A more comprehensive way of looking at the generation of ideology from the text could then combine topic features with opinion features. The new assumption is that we know more about ideology by not only getting the topic distribution, but also the sentimental orientation on each topic. The topics reflect general policy interest, and the sentiments reflect attitude and preference of policies. A possible solution would be to look at nouns and other words differently when applying the topic model. Nouns can be modeled as topics, and the adjectives, adverbs, verbs can be divided into categories with linguistic or sentimental meanings to model opinions, and the topic features and opinion features can then be applied to a new regression model. If we set the number of opinion categories to be two, it simplifies into a positive opinion category and a negative opinion category. When we transfer the domain from the congress to the blogosphere, the topic features do not change, but the opinion features might need to be adjusted/re-modeled to reflect the changes of language use.

Chapter 5

Real World Applications and Future Work

Intuitively, the connection between the real world and opinion mining studies lies in the nature of human beings: we want to communicate with people around us and understand each other. Just like in Psychology, researchers study "Emotion" to explore our psychological changes and the follow-up effects on behaviors; in Opinion Mining, researchers study "Opinion" to explore our attitude changes or ideology changes as well as the follow-up effects on behaviors. For example in business opinion mining, the studies help to collect user experiences of products and customer feedback without extra labor cost to the companies; in this way, they provide information to the producers on how to improve their services and products.

5.1 Applications of Political Opinion Mining

Political opinion mining has similar objectives. Detecting the ideology of the politicians is nothing else but providing extra information of the politicians that will help the electorate to know more of the candidates in a super-explicit way. We admit that It is for sure a simplification of "ideology" to quantify it in the theoretical spatial models, and a simplification of politicians to judge and rate them just in some numbers. But ideology scores help to extract the objective traits of the politicians, as

compared to the personalistic traits that are much more likely to have been decorated and selectively presented by the politicians, as well as incompletely received by the electorate.

Political opinion mining also has the potential to be applied to collect public opinions. Instead of launching time and money-costly polls and surveys to get access to public opinions, the information era allows us to utilize the internet as one of the information resources to collect public opinions. For example, in the political blogosphere, the bloggers are interested in politics and they always keep track of the ongoing political news and discuss/comment on them. By applying the topic models to the blogosphere, we get knowledge of their policy interests in the form of the "topics"; and by detecting the opinion features on the topics, we then get their attitudes and standings on the policies. A study of this kind can then be taken as a replacement of the traditional polls in the sense that it is able to describe the policy positions of either a single blogger or the collective opinions for the overall community in an efficient and reliable way.

Time-series studies on the blogosphere can even reflect more on the dynamics of opinion generation and evolution. And with the help of machine translation to tackle multi-linguistic issues, we can even conduct comparative studies on political opinions, such as comparative studies on the distribution of political opinions with respect to region, time and different stages of democracy. These studies will help us to get a deeper understanding of how opinions are generated and why opinions diverge and conflict in different regions and cultures.

Another thought is that opinions evolve after interacting with others in the social networks. People post their political comments on the web, and interact with local or remote friends as they might agree or disagree on the comments. On the blogosphere, blogs link to their "friend" blogs with similar ideology; on twitter, people follow others and share tweets. Opinion is not just been "spread" on these social networks, it is actually also evolving at the same time. Although it is almost impossible for anyone to change their ideology overnight dramatically, living in the social networks means that they will inevitably be influenced by other thoughts and change will happen as long as the external influences have accumulated to a certain degree. So I would like

to study the political blogosphere and online media to discover how opinions initialize and change in this system, how the opinionated articles are spread in this system and how the users are involved in this dynamic opinion generation/evolution process.

5.2 Applications of Political Opinion Retrieval

Political opinion retrieval is a natural extension of opinion mining studies. Google shows us how a general search engine works in returning query-related web pages, for an opinion search engine, the queries will contain not only keywords, but also indicators of opinions to reflect the users' information needs of opinions. For a political opinion search engine, one application scenario could be the user gives a query "find me a legislator/blog that talks about health care reform, who/which holds a socially neutral but economically extreme conservative standing", and the opinion search engine would have to follow this instruction to return only the required results. A political opinion search engine like this is supposed to help users to search for reference texts on all the ideology dimensions and thus know better of the target issue after reading through all the different arguments.

Different information needs result in different opinion retrieval tasks and methods. Below I would like to propose four opinion retrieval application scenarios:

- User Scenario 1 (basic form of the topic-sentiment retrieval task):
  The users are interested in searching for opinions with a certain position on a certain topic. For example, they might ask: Give me the articles talking about abortion which take a mildly supportive attitude. The opinion position they query here is not necessarily their own opinion position. They might instead search for articles written by people with opposite standing on this topic. The format of queries in this scenario is also different from traditional queries. The new query is composed of two parts: the first part is a regular keyword-based query string describing the targeted topic/issue; the second part is a rate number (e.g. 3 in 1, 2, 3, 4, 5) or a float number (e.g. 0.5 as in the range of [0, 1]) to represent the specified opinion position on the query string. Since this retrieval

task asks for a topic first and then specifies the sentiment on that topic, I call it a topic-sentiment retrieval task.

- User Scenario 2 (an extension of the basic topic-sentiment retrieval task):
  Another user scenario might be that the user gives a regular query which is already opinionated, such as "I hate Bill Clinton". Then a topic-sentiment retrieval system should detect the opinion position in the query and reformulate the query into the format in User Scenario 1, e.g. a query string "Bill Clinton" and a number 0 as representing an extreme negative attitude on "Bill Clinton". In this way, we reformulate an opinionated query into a topic-sentiment query, thus transform the task into the one in Scenario 1.

- User Scenario 3 (basic interpretation of the ideology retrieval task):
  The users might also be interested in searching for Ideology other than Opinion/Sentiment. The term "Ideology" comes from Political Science, where the scientists believe that opinions are generated/derived endogenously from ideologies and ideologies are constructed in multiple dimensions. One typical setting is the economic-social ideology model, where a two dimensional coordinates can illustrate and represent the ideology space. As a simple example, for a two-dimensional ideology model, each opinion entity (authors, document owners) has two ideology scores to describe his/her positions on each of the two dimensions. Searching with an Ideology score means that the users want to search for articles on one topic written by someone having that Ideology score. So again, a keyword-based query string and a set of ideology scores together compose a query, while the query string is used to refine relevant documents, and the ideology score is used to refine authors instead of documents.

- User Scenario 4 (an extension of the basic ideology retrieval task):
  The users do not necessarily have to provide the ideology scores along with their query strings. We can take in the settings of personalized retrieval tasks here, assume that by studying search histories of the users, we can construct ideology profiles for the users, and calculate their preferred ideology scores automatically for them. A simple method to get ideology scores for the users might be adding

up ideology scores of the documents they have searched and browsed. In this scenario, as an ideology retrieval task, the users provide only the query strings, and the ideology scores come from search records of the users as background knowledge.

Bibliography

[1] R. Ackland. Mapping the u.s. political blogosphere: Are conservative bloggers more prominent? *BlogTalk Downunder 2005 Conference, Sydney*, 2005.

[2] L. A. Adamic and N. Glance. The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, New York, NY, USA, 2005. ACM.

[3] L. D. Baker and A. K. McCallum. Distributional clustering of words for text classification. In *SIGIR*, 1998.

[4] V. Batagelj. Efficient algorithms for citation network analysis. *CoRR*, cs.DL/0309023, 2003.

[5] K. Benoit, M. Laver, and S. Mikhaylov. Treating words as data with error: Uncertainty in text statements of policy positions. *American Journal of Political Science, 53(2):495-513*, 2009.

[6] J. M. Box-Steffensmeier and D. P. Christenson. Invaluable involvement: Purposive interest group networks in the 21st century. *Working Paper*, 2010.

[7] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. Network analysis of collaboration structure in wikipedia. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 731–740, 2009.

[8] D. V. Brewington, D. R. Davis, and A. Murdie. The ties that bind: A network analysis of human rights ingos. *the 50th Annual Convention of the International Studies Association*, 2009.

[9] W. P. Butz and B. B. Torrey. Some frontiers in social science. *Science. 2006 Jun 30;312(5782):1898-900*, 2006.

[10] L.-E. Cederman, H. Buhaug, and J. K. Rod. Ethno-nationalist dyads and civil war: A gis-based analysis. *Journal of Conflict Resolution, Vol. 53, No. 4, 496-525*, 2009.

[11] B. Chen, L. Zhu, D. Kifer, and D. Lee. What is an opinion about? exploring political standpoints using opinion scoring model. In *the 24th AAAI Conference on Artificial Intelligence*, 2010.

[12] Y. Chevaleyre, U. Endriss, J. Lang, and N. Maudet. A short introduction to computational social choice. *33rd Conference on Current Trends in Theory and Practice of Computer Science*, 2007.

[13] J. Clinton, S. Jackman, and D. Rivers. The statistical analysis of roll call data. *The American Political Science Review, Vol. 98, No. 2 (May, 2004), pp. 355-370*, 2004.

[14] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *WWW*, pages 519–528, 2003.

[15] D. Diermeier, J.-F. Godbout, B. Yu, and S. Kaufmann. Language and ideology in congress. *Midwest Political Science Association*, 2007.

[16] D. Diermeier, J.-F. Godbout, B. Yu, and S. Kaufmann. Language and ideology in congress. *Annual Meeting of the Midwest Political Science Association*, 2007.

[17] A. Downs. *An Economic Theory of Democracy*. New York: Harper and Row, 1957.

[18] K. T. Durant and M. D. Smith. Mining sentiment classification from political web logs. In *WEBKDD*, 2006.

[19] M. Efron. The liberal media and right-wing conspiracies: using cocitation information to estimate political orientation in web documents. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 390–398, New York, NY, USA, 2004. ACM.

[20] A. Esuli and F. Sebastiani. Determining the semantic orientation of terms through gloss classification. In *CIKM*, pages 617–624, 2005.

[21] A. Fader, D. Radev, M. H. Crespin, B. L. Monroe, K. M. Quinn, and M. Colaresi. Mavenrank: Identifying influential members of the us senate using lexical centrality. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 658C666.*

[22] R. Feldman and J. Sanger. *The Text Mining Handbook: Advanced Approaches to Analyzing Unstructured Data*. Cambridge University Press, 2007.

[23] J. H. Fowler. Connecting the congress: A study of cosponsorship networks. *Political Analysis, 14:456C487*, 2006.

[24] J. H. Fowler and S. Jeon. The authority of supreme court precedent. *Social Networks 30 (1): 16-30*, 2008.

[25] J. H. Fowler, T. R. Johnson, J. F. S. II, S. Jeon, and P. J. Wahlbeck. Network analysis and the law: Measuring the legal importance of precedents at the u.s. supreme court. *Political Analysis, 15:324C346*, 2007.

[26] D. Gaffney. iranelection: Quantifying online activism. In *Web Science Conf. 2010, April 26-27, 2010, Raleigh, NC, USA.*, 2010.

[27] R. Gibson and S. Ward. A proposed methodology for studying the function and effectiveness of party and candidate web sites. *Social Science Computer Review 2000; 18; 301-319*, 2000.

[28] A. A. Goodrum, K. W. McCain, S. Lawrence, and C. L. Giles. Scholarly publishing in the internet age: A citation analysis of computer science literature. *Information Processing and Management, v37 n5 p661-75 Sep 2001.*

[29] J. Greenberg and E. Jonas. Psychological motives and political orientation-the left, the right, and the rigid. *Psychological Bulletin*, 129:376–382, 2003.

[30] G. Grendstad. A political cultural map of europe. a survey approach. *GeoJournal*, 47:463–475, 1999.

[31] J. Grimmer. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Southern Political Science Association*, 2009.

[32] J. Grimmer and G. King. Quantitative discovery from qualitative information: A general-purpose document clustering methodology. *APSA 2009 Toronto Meeting*.

[33] J. H. Gross. Cosponsorship in the u.s. senate: A multilevel approach to detecting the subtle influence of social relational factors on legislative behavior. *Working Paper*, 2008.

[34] V. Gueorguieva. Voters, myspace, and youtube: The impact of alternative communication channels on the 2006 election cycle and beyond. *Social Science Computer Review 2008; 26; 288-300*, 2008.

[35] J. Han and Y. Kim. Obama tweeting and twitted: Sotomayors nomination and health care reform. *Working Paper*, 2009.

[36] E. Hargittai, J. Gallo, and M. Kane. Cross-ideological discussions among conservative and liberal bloggers. *Public Choice (2008) 134: 67-86*, 2008.

[37] M. Hindman, K. Tsioutsiouliklis, and J. A. Johnson. Googlearchy: How a few heavily linked sites dominate politics online. *Midwest Political Science Association*, 2003.

[38] D. Hopkins and G. King. Extracting systematic social science meaning from text. *Midwest Political Science Association*, 2007.

[39] M. Hu and B. Liu. Mining and summarizing customer reviews. In *SIGKDD*, pages 168–177, 2004.

[40] P. T. Jaeger, J. Lin, and J. M. Grimes. Cloud computing and information policy: Computing in a policy cloud? *Journal of Information Technology and Politics, 5: 3, 269 283*, 2008.

[41] A. Jakulin, W. Buntine, T. M. L. Pira, and H. Brasher. Analyzing the u.s. senate in 2003: Similarities, clusters, and blocs. *Political Analysis 17:291C310*, 2009.

[42] M. Jiang and H. Xu. Exploring online structures on chinese government portals: Citizen political participation and government legitimation. *Social Science Computer Review 2009; 27; 174-195*, 2009.

[43] N. Jindal and B. Liu. Opinion spam and analysis. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 219–230, New York, NY, USA, 2008. ACM.

[44] A. Z. Kamran. Human security networks in the indian ocean tsunami. *Harvard Political Networks Conference*, 2009.

[45] J. Kelly and B. Etling. Mapping iran's online public: Politics and culture in the persian blogosphere. *Berkman Center Research Publication No. 2008-01*, 2008.

[46] J. W. Kelly, D. Fisher, and M. Smith. Friends, foes, and fringe: norms and structure in political discussion networks. *Proceedings of the 2006 international conference on Digital government research: 412-417*, 2006.

[47] G. King, K. L. Schlozman, and N. Nie. *The Future of Political Science: 100 Perspectives (Editor Book)*. Routledge, 2009.

[48] G. Koger, S. Masket, and H. Noel. Partisan webs: Information exchange and party networks. *British Journal of Political Science, 39:633C653*, 2009.

[49] G. Koger and J. N. Victor. The beltway network: A network analysis of lobbyists' donations to members of congress. *American Political Science Association*, 2009.

[50] K. Kollman, J. H. Miller, and S. E. Page. Adaptive parties in spatial elections. *The American Political Science Review, 86(4):929-937*, 1992.

[51] P. F. Landry and M. Shen. Reaching migrants in survey research: The use of the global positioning system to reduce coverage bias in china. *Political Analysis 13:1C22*, 2005.

[52] M. Laver, K. Benoit, and J. Garry. Extracting policy positions from political texts using words as data. *American Political Science Review, 97:2:311-331*, 2003.

[53] M. Laver, K. Benoit, and J. Garry. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97:311–331, 2003.

[54] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. V. Alstyne. Life in the network: the coming age of computational social science. *Science. 2009 February 6; 323(5915): 721C723*, 2009.

[55] X. Ling, Q. Mei, C. Zhai, and B. Schatz. Mining multi-faceted overviews of arbitrary topics in a text collection. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–505, New York, NY, USA, 2008. ACM.

[56] B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *WWW*, pages 342–351, 2005.

[57] Y. Liu, X. Huang, A. An, and X. Yu. Arsa: a sentiment-aware model for predicting sales performance using blogs. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 607–614, New York, NY, USA, 2007. ACM.

[58] Y. Lu and C. Zhai. Opinion integration through semi-supervised topic modeling. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 121–130, New York, NY, USA, 2008. ACM.

[59] Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 131–140, New York, NY, USA, 2009. ACM.

[60] S. E. Masket, M. T. Heaney, J. M. Miller, and D. Z. Strolovitch. Networking the parties: A comparative study of democratic and republican national convention delegates in 2008. *APSA 2009 Toronto Meeting*, 2009.

[61] Y. Matsuo and H. Yamamoto. Community gravity: measuring bidirectional effects by trust and rating on online social networks. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 751–760, 2009.

[62] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *WWW*, pages 171–180, 2007.

[63] B. Monroe, M. Colaresi, and K. Quinn. Fightin words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis, 16(4):372-403*, 2008.

[64] B. Monroe and P. Schrodt. Introduction to the special issue: The statistical analysis of political text. *Political Analysis, 16:351C355*, 2008.

[65] B. L. Monroe and K. Maeda. Talks cheap: Text-based ideal point estimation. *Paper presented to the Political Methodology Society, Palo Alto, July 29C31, 2004.*, 2004.

[66] T. Mullen and R. Malouf. A preliminary investigation into sentiment analysis of informal political discourse. *In Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs.*

[67] S. Nerur, R. Sikora, G. Mangalaraj, and V. Balijepally. Assessing the relative influence of journals in a citation network. *Commun. ACM*, 48(11):71–74, 2005.

[68] H. Noel. A social networks analysis of internal party cleavages in presidential nominations, 1972-2008. *Harvard Political Networks Conference*, 2009.

[69] B. Pang and L. Lee. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[70] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135, 2008.

[71] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of english words. In *ACL*, 1993.

[72] A. Plangprasopchok and K. Lerman. Constructing folksonomies from user-specified relations on flickr. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 781–790, 2009.

[73] K. T. Poole and H. Rosenthal. A spatial model for legislative roll call analysis. *American Journal of Political Science, Vol. 29, No.2 (May, 1985),357-384*, 1985.

[74] K. T. Poole and H. Rosenthal. A spatial model for legislative roll call analysis. *American Journal ofPolitical Science*, 29:357–384, 1985.

[75] S. Purpura and D. Hillard. Automated classification of congressional legislation. *Proceedings of the 2006 international conference on Digital government research, 219-225.*

[76] K. M. Quinn, B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science, Vol. 54, No. 1., 209-228*, 2010.

[77] S. M. Robbins. Cash flows: Leadership pacs in the u. s. congress from 1992-2008. *43rd Hawaii International Conference on System Sciences*, 2010.

[78] S. Rosset, C. Perlich, and B. Zadrozny. Ranking-based evaluation of regression models. In *The Fifth IEEE International Conference on Data Mining (ICDM 05), Houston, Texas*, pages 370–377, 2005.

[79] L. Sigelman. The coevolution of american political science and the american political science review. *American Political Science Review (2006), 100:4:463-478*, 2006.

[80] J. B. Slapin and S.-O. Proksch. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science, 52(3):705-722*, 2008.

[81] M. Steyvers and T. Griffiths. *Probabilistic topic models. In T. Landauer, D.S. McNamara, S. Dennis, and W. Kintsch, editors, Handbook of Latent Semantic Analysis*. Erlbaum, 2007.

[82] S. Tan, G. Wu, H. Tang, and X. Cheng. A novel scheme for domain-transfer problem in the context of sentiment analysis. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 979–982, New York, NY, USA, 2007. ACM.

[83] M. Thomas, B. Pang, and L. Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing , 327C335.*

[84] I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 111–120, New York, NY, USA, 2008. ACM.

[85] B. Ulicny. Modeling malaysian public opinion by mining the malaysian blogosphere. *Book Chapter: Social Computing, Behavioral Modeling, and Prediction*, 2008.

[86] D. J. Watts. A twenty-first century science. *Nature 445, 489 (1 February 2007)*, 2007.

[87] C. Whitelaw, N. Garg, and S. Argamon. Using appraisal groups for sentiment analysis. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 625–631, New York, NY, USA, 2005. ACM.

[88] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *ICDM*, 2003.

[89] B. Yu, S. Kaufmann, and D. Diermeier. Classifying party affiliation from political speech. *Journal of Information Technology and Politics, 5(1):33-48*, 2008.

[90] B. Yu, S. Kaufmann, and D. Diermeier. Classifying party affiliation from political speech. *Journal of Information Technology and Politics*, 5:33–48, 2008.

[91] B. Yu, S. Kaufmann, and D. Diermeier. Exploring the characteristics of opinion expressions for political opinion classification. In *Proceedings of the 2008 international conference on Digital government research*, pages 82–91, 2008.

[92] Z. Zhang and B. Varadarajan. Utility scoring of product reviews. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 51–57, New York, NY, USA, 2006. ACM.

[93] L. Zhuang, F. Jing, and X.-Y. Zhu. Movie review mining and summarization. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50, New York, NY, USA, 2006. ACM.