Making Computers Understand Coalition and Opposition in Parliamentary Democracy

Matthias Steinbauer, Markus Hiesmair, Gabriele Anderst-Kotsis

Making Computers Understand Coalition and Opposition in Parliamentary Democracy

Matthias Steinbauer, Markus Hiesmair, and Gabriele Anderst-Kotsis

Johannes Kepler University of Linz, Institute of Telecooperation,
Altenbergerstrasse 69, 4040 Linz, Austria
`markus.hiesmair@tk.jku.at`
`{matthias.steinbauer,gabriele.kotsis}@jku.at`
`http://www.tk.jku.at/`

Abstract. In recent years a tremendous raise in the establishment of Open Data initiatives can be observed, aiming at more transparency in government and public institutions. One facet of this trend are data from legislative bodies, including records and archived transcripts of plenary sessions as a measure of transparency and accountability. In this paper the system design and a prototypical implementation of an information system that makes use of these data is presented. From session transcripts naive metrics such as when and how often representatives participate in political discourse but also network metrics as in with whom representatives engage in consenting and opposing discourse can be derived. The objective of the system is to make those relationships visible and accessible to the user in an intuitive way. The system neither can nor attempts to interpret the data, this is left to the user. This paper discusses how data analytics, data visualisation, and network analytics can be facilitated to make the transcripts of legislative bodies more accessible for this purpose. The findings are underpinned by first observations over a proof-of-concept prototype which exploits data available from the Austrian parliament.

Keywords: Data Visualisation, Open Data, Network Analytics

1 Introduction

Technological, political, and sociological developments in recent years are leading to a situation where public bodies, governments, but also many other organisation with extensive influence on the general public aim for higher transparency in their management. In democracy a certain degree of transparency is achieved by making the legislative process public and in other bodies key data is made available to the public through Open Data platforms.

A tool for larger transparency in parliamentary democracy are the transcripts of debates in the different legislative bodies. These transcripts are created during debates by stenotypists, are then typeset and published as continuous volumes. Depending on the country these volumes are distributed to policy makers

and subscribers via mail, and are available for reading in libraries. Today many legislative bodies provide such debate transcripts as part of their Open Data initiatives.

Although this already allows very detailed and good insight into the democratic decision making process the transcripts bear some detriments. First of all the sheer volume of text and data found in these transcripts makes it time consuming to analyse the political discourse such that the general public needs to rely on more condensed information formats as provided by daily political news papers and news shows. Direct analysis of the transcripts remains in the domain of professionals such as political analysts, researchers, and journalists.

Further, also professionals might find it hard to analyse simple metrics in the given data. The extraction of information interesting to the general public, such as how often their elected representatives engage in discussions and which political positions they take, will often require manual analysis of large volumes of transcripts.

Finally, from the mere text interesting structures in the political landscape are hard to observe. Revealing the structure of the political discussion and finding links between individual policymakers would usually require analysts with informed background.

In the given work the authors claim that by lending methodologies from automated information extraction, data modelling, and graph analytics one is able to generate structured data about the political discourse in parliamentary democracy. The structured data allows to objectively compute metrics over the observed system and by creating agreeable visualisation allows clear insight into the political system for the general public.

It is believed that by employing a process which (1) continuously retrieves transcripts from a legislative body, (2) extracts the relationships between actors in the discourse, and (3) visualizes the results, an important contribution to political transparency can be made. Politically interested are able to make more informed decisions. These will still be base based on information as received through media, however, arguments can be questioned and verified in the available data.

A public software system that follows the process described above can provide simple metrics on the members of political parties and the participants of political discourse. The system can provide data on how often individual members attended sittings of a council, and if attended how actively they took part in discussions.

Through taking part in discussions policymakers induce relationships. These relationships can be tracked in a network (sociogram) which can be modelled as a temporal graph. By visualizing this graph relationships between politicians and political parties become visible, groups of politicians (supposeably with similar attitudes) can be found, and formal groups (parties, coalition, and opposition) can be analysed for their homogeneity.

Future legislative periods in democratic systems might become rateable; the electorate might assess the performance of the delegates through political perfor-

mance indicators just as companies now use key performance indicators in their informed decisions. With mathematical models and machine learning approaches it might even be possible to make predictions.

The remainder of this paper is structured as follows. In section 2 inspiring work from automated analytics of political systems and relevant foundational material from the relevant fields such as information retrieval and temporal graph analytics is discussed. Section 3 discusses the system design of a software system as proposed in this work, and section 4 presents a real-world prototypical implementation in greater detail. In section 5 the use of the system on the example of the Austrian parliament and first observations are presented. Finally section 6 provides pointers towards future work and concludes the paper.

2   Related Work

The analysis of political debate and reflection upon the performance of public bodies are key tasks of political science and social sciences. Traditionally these sciences afford high expert involvement. Manual review of literature, transcripts, and datasets are often used as methods.

Recent progress in computer science, the boom of the social web, and transparency efforts towards Open Data lead to a spiked interest in political analysis from other fields of research such as computer science. Ultimately nurturing efforts towards automated analytics of political structures starting with pure lexical analysis of political debate [13] and stopping at structural analytics of Big Data resources [14].

2.1   Open Data

Open Data in the context of public bodies is defined as data which is non-privacy-restricted and non-confidential that was generated with public money. It becomes Open Data when made available without any restrictions on its usage or distribution. It is assumed that Open Data closes gaps between public organisations and citizens thus nurturing discourses and the exchange between public bodies and citizens is seen as constructive. Open Data can coarsely be categorized into political and social data, economic data, and operational and technical data [8]. The data used in this paper falls into the first category.

Although governments worldwide are at different levels of installing Open Data initiatives some early adopters can already look back at a history and lessons learned from Open Data. For instance Shadbolt et al. [9] are able to reflect on the benefits gained from the linked open government data platform `http://data.gov.uk` installed in the United Kingdom. As an important finding it becomes clear that the state transforms to a service provider. The vision of an Internet of linked open data and thus also linked open government data makes us believe that systems such as the one described in this paper will become easy to implement in the future.

2.2 The Social Web

However, not only official state bodies provide political analytics with data. Also a vast amount of services which invite their users to social interaction form another pool of information. In general social networks, their structure and especially information diffusion are topics which are very well studied [10]. In the context of this paper political discourse in online social networks are of particular interest.

Exemplary studies that address political discourse on the popular sites Twitter and Facebook are given in [11, 12]. In Hsu et al. the micro-blogging service Twitter was used to scrape information on a distinct political topic in South Korea. The study shows that a limited number of opinion leaders are the main drivers in the political discussion around this topic. This is clear through the fact that thousands of users interact with the artefacts on the site which were created by the opinion leaders. From the 20 identified key users several results were derived. (1) The users were categorized and some of the most popular key users refer to large Korean media outlets which are already opinion leaders in other media (print, TV, radio broadcast, etc.). (2) Central keywords were derived from the discourse and clusters were derived from them such that the political position of the key users becomes visible. (3) Finally, the keyword clusters and key users were visualized as network diagrams such that the links between them become visible [11].

In contrast Kushin et al. discuss the computer mediated communication possible in online social networks. These systems have been criticised for isolating disagreeing persons from engaging in discussions and for fostering atmosphere of uncivil behavior due to a perceived feeling of anonymity and distance between the actors. Although political discussions on the web have been taking place since the very beginning of the public Internet and thus also the analysis of it is a long standing topic of interest, systems such as Facebook allow for deeper insight. Whereas in the past discussions where scattered over many different platforms such as web forums and Usenet groups some of which accessible only to a technically proficient audience, now systems like Facebook and Twitter are used by a wide demographic. In online social networks different aspects of political engagement are possible and according to Kushin et al. will lead to different reactions. Users can be-friend politicians, can express their interest in political content posted on sites, and can directly comment on political content posted by other users.

2.3 Structural Analytics

In fall 2013 Renzo Luicioni created several graph visualisations that highlight voting relationships between US senators from the 101st congress throughout the 113th congress [1]. The data was scraped from GovTrack.us [4] converted to graph structures which were then automatically layouted by an implementation of the ForceAtlas algorithm [7] as found in the Gephi graph visualisation workbench. The results impressively document how the political landscape in

the US morphed from a collaborating scene towards a polarized political landscape. In the recent visualisations one can get the impression that the two major forces (Democrats, and Republicans) are almost dictating the voting schemes. The work of Luicioni was picked up by Yahoo News [2] and since it spiked large interest was later featured in a short piece in The Economist [3].

Although the work of Luicioni gained much public attention there has been earlier work in the field of structural analysis of political networks. Naturally the field of graph analytics has interest in this area. Well known metrics such as centrality measures, graph partitioning, and graph clustering can also be applied on political networks. In 2005 Porter et al. [5] were able to successfully demonstrate the application of graph clustering algorithms on data originating from the U.S. house of representatives. The outcome of their studies are dendrograms representing the hierarchical structure of the different communities within the political bodies. Their results also underpin the visual results of Luicioni as the clusters in their data show a high degree of separation.

Based on the findings of Porter et al., Amelio and Pizzuti [6] studied the voting behavior in the Italian parliament. In the first part of their study similar results are presented. Also the Italian parliament shows community structure which can be broken down into a dendrogram. However, further metrics such as the cohesion of political parties and the similarity in voting behavior were analysed. An interesting finding was that the cohesion within the governing parties decreased in relevant time-spans of the observed dataset. On the other hand cohesion within opposition increased. Ultimately the political landscape changed and government was not reelected. This leaves room for the interpretation that future automatic analytics systems might predict probabilities of government reelection.

Where the previously mentioned related work base their analysis on structured data of political systems, the work of [14] works in a larger context. The described software pipeline is able to detect election-related articles in large corpora of news articles and political information systems, parses them. After parsing key actors, objects, and actions are identified and used to form a network structure of political key players and topics.

3   System Design

In the following the overall system design of the analysis platform is discussed. The system lends its general processing structure from the well known ETL (extract, transform, load) steps as found in business intelligence applications. The ETL process is then continued by a processing and visualisation step. The process is outlined in figure 1.

The phase *Extract* is responsible to retrieve relevant data from a data source such as an Open Data repository. Depending on the actual implementation of the repository a variety of different methods can be used. For instance many large public bodies are starting to adopt data platforms such as CKAN[1] which

---

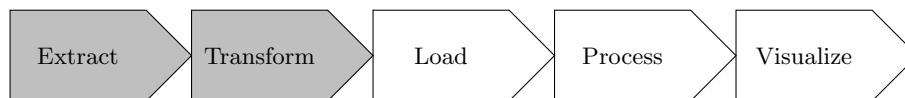[1] The Comprehensive Knowledge Archive Network (CKAN): `http://ckan.org`

Fig. 1: General structure of the processing pipeline

amongst others provides REST based APIs. Other data might have access paths based on the RDF Site Summary (RSS) framework or might be presented in other open or even proprietary formats. Hence the *Extract* component is tightly interlinked with the data resource it is bound to. This is indicated in the pipeline with grey filling of the box.

Also the *Transform* phase has a tight binding on the actual data source. Data about political debate is available in many different formats. For instance the transcripts of the Austrian parliament are available as HTML and PDF documents[2], the Italian parliament provides structured voting records on their site[3], and for the US the site govtrack.us[4] provides structured data and full text from many governmental bodies.

Observing the landscape of data sources it becomes clear that the two data-bound phases (*Extract* and *Transform* marked in grey) need to be adapted to specific data providers. However, for all of the resources it is possible to transform them into a set of structured data which contains representatives, and their voting and discourse patterns. This structured data is the input for the *Load* phase which uses the structured data and loads them to a query-able data repository such as a relational database management system. The *Load* phase reads input data in a generic data-format or through standardised APIs such that a general implementation of this phase can be used regardless the data-source.

On top of the loaded data model typical data analysis tasks can be run. Such as computing relevant metrics in the *Process*, and creating human-readable interpretations of the data in the *Visualize* phase. Metrics computed over the available data can be roughly discriminated into two groups. The first group are metrics that provide simple indicators over records found in the datasets. Exemplary indicators in this group for individual politicians are: the total number of years the representative is in service, degree of attendance in sessions, number of speeches and interactions in the plenary. We call these indicators naive indicators or metrics.

Further more complex indicators can be derived from the interaction network that is formed by representatives engaging in discussions with each other. Network metrics such as the node centrality, and betweenness centrality can be used to determine which actors are at the core of groups or who acts as a hub between individual groups. Further, methods from community detection can be applied to reveal the groups that form within the network. These are of partic-

---

[2] Austrian Parliament Session Transcripts:
  https://www.parlament.gv.at/PAKT/STPROT/
[3] Italian Parliament Open Data: http://parlamento17.openpolis.it
[4] govtrack.us: https://www.govtrack.us

ular interest if compared to formal groups that are expected to be found in the network such as coalition, opposition, and political parties.

Depending whether naive or network metrics are of interest different tactics can be applied to visualize the data for the user. Naive metrics can mostly be reflected through the use of standard charts such as bar-charts or scatter-plots. The network data can be visualized through automatically layouted graph representations. Additional information in this case must be color coded.

As the process described above is designed to be fully automated it can be repeated on a regular basis. This leads to a system that is constantly fed with current information and allows the creation of a user-facing dashboard that can be used to analyse the current but also past situations.

The current landscape of Open Data in combination with the ETL and processing steps described above and the use of methods from graph analytics allows the creation of a prototype system that gives a first impression as of how in future the insight into public bodies can be significantly improved.

4   Proof-of-Concept Prototype

To demonstrate the mere technical feasibility of our approach and to allow first usability tests with focus groups, a proof of concept prototype for the presented system design was created. In this first prototype openly available data from the Austrian parliament, was used. The prototype uses politician profiles and transcripts of the sessions of the national council which are both publicly available as HTML files. With these data sources, general data of politicians (birth date, ...), their membership in political parties and their activities and absences during sessions of the national council can be derived. Furthermore, relations among politicians and parties can be calculated through meta data of the speeches held in the parliament.

An important aspect while building the prototype was extendibility, especially the *Extract* and *Transform* phase of the processing pipeline must be adaptable. The prototype was built for the national council of the Austrian parliament, but in general the system has been held modular and therefore legislative systems of other countries can be targeted as well, if the data is available in sufficient quality and of an overall similar structure.

The prototype was implemented with state of the art Java and Spring standard frameworks and consists of the following modules:

Extractor: Loads the raw HTML-Files from the Data Source (in our case the Austrian Parliament Web Site). The data-source provides an RSS feed which can be used to get up-to-date information.

Transformer: Downloaded HTML files are parsed in the transformer module. Depending on the input file different output is generated. From politician profiles the parser is able to derive a structured profile, from session transcripts the parser finds votes and debates and assigns politician profiles to the actors.

Loader: Loads the data-source independent records provided from the transformer into a relational database system.

Analyzer: Calculates basic measures and generates the relation graphs for politicians and parties. The relationship graph is built by analysing how politicians expressed sentiment towards topics discussed in the plenary. For the dataset used vast amounts of speeches and contributions from the auditorium are marked pro and contra arguments. The normalized edge weight of relationships between actors is used to express the overall pro and contra disposition between any two actors.

Community Detector: Automatically detects communities in the relation graph using a label propagation algorithm [15]. The algorithm can be configured to consider only edges in a certain weight-range such that more global or local communities can be found.

Web Visualization: Contains mainly the user interface which presents the computed metrics and provides graph visualisation.

As intended by the system design other legislative systems can be connected through replacing the extractor- and transformer-module with implementations for the respective data source. All other modules will work for other systems without the need for a change.

The real world implementation of the prototype is available as open source software. The code can be found online at Github[5]. Screenshot in figure 2a gives a first impression for the graphical representation of a legislative period. It gives rough overviews on session meta-data and highlights some of the naive metrics. Interested users can drill down for instance to politician profiles as presented in figure 2b. The profile puts the selected politician in context with other politicians in the legislative body. Graph visualisation is discussed in the next section.
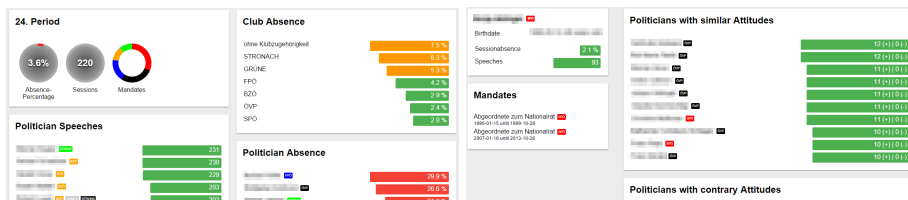
(a) Overview over a period   (b) Politician profile

Fig. 2: Prototype screenshots

5   Observations

During the course of creating the prototypical implementation for the political information system presented in this paper it became clear semi-automated and automated analysis of legislative bodies is already technically feasible. Open Data platforms provide the required data which can easily be processed and analysed with state of the art methods from data analytics, data visualisation and in this case also network analytics.

As for our showcase scenario we can also report that the data extraction process from text / HTML based transcripts works surprisingly well. Since in the showcase in-depth profiles from politicians are available actors in the transcripts can be looked up in an index which leads to a completely correct mapping of actor names to politician profiles in the observed dataset. Obviously over the years the formatting of the transcripts continuously improved such that parsing mechanisms need to adapt as well. In the observed dataset there is one major technology change. Old versions of the transcripts are actually scanned text documents instead of HTML. If these were to be analysed optical character recognition techniques would be required. For transcripts from other legislative bodies also some annotations such as the pro/contra indicators found for the Austrian parliament might be missing. In this case advanced methods from text processing such as automated sentiment analysis will be required.

Already the naive metrics presented in period overviews and politician profiles provide interesting insight. However, politician interaction network graphs as presented in figure 4 provide even deeper insight. The graphs have been automatically layouted by a force driven layout algorithm [7]. The algorithm in general tries to place nodes as far apart as possible, however the weighted edges create a opposing pull force. This leads to a layout process where politician profiles with similar attitudes get pulled close together and opposing attitudes drawn apart. In the output it is clearly shown that the network is clustered.

In figure 4 the periods 22 and 25 were chosen on purpose because these two graphs both show two clearly distinct clusters. In both visualisations the left cluster is formed by profiles in the coalition government and the right cluster contains profiles from the oppositions parties. In the 25th period we can see that in the opposition the green nodes (profiles from the Austrian Greens) are a little closer to the government than the blue nodes (Freedom Party of Austria). In general the nodes in the opposition cluster are less densely layouted than in the coalition government. This is conform with the opposing political agenda of the opposition parties. In the 22nd period however a different coalition government was formed (black, blue, and orange nodes). Again one can observe two clearly distinct clusters, however both clusters are far more dense.

The very same clusters are detected by the community detection algorithm chosen in our experiments [15]. The community detection algorithm was run exemplary on periods 20 through 25 of the dataset and the community labels assigned to the individual political profiles were compared with the official politician profiles. The algorithm in [15] describes an iterative process, in our experiments ten iterations led to stable communities. Further a threshold for edge se-

lection was used such that only edges with an absolute edge-weight above 3 were considered during community detection. This number was determined throughout multiple experiment runs and is a parameter which most likely needs to be adjusted for other datasets. Most of the Austrian representatives are organised in clubs such that it is save to assume that a politician who is a member of a governing party is part of the government. However, there are rare cases where politicians change clubs and thus move from government to opposition during a period. The chart in figure 3 shows that in worst case the community detection algorithm assigned more than 91% of the profiles to the correct group but on average (98%) it is doing far better.
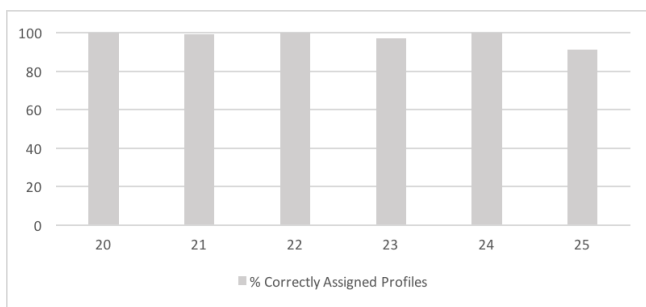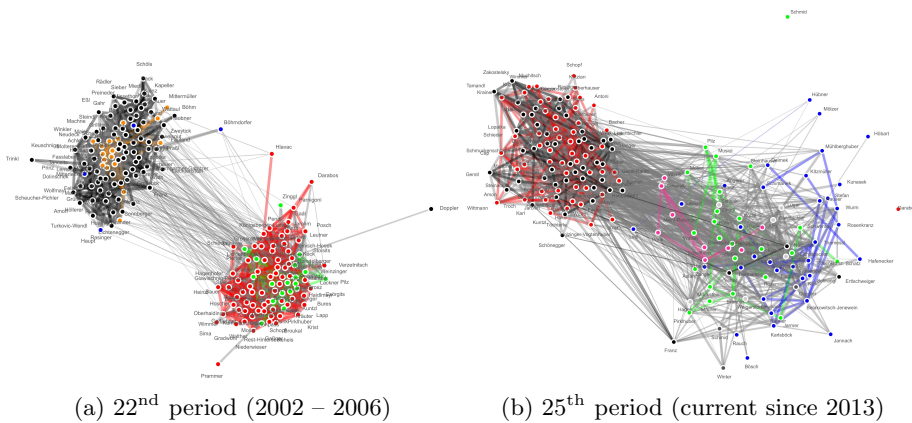

Fig. 3: % of correctly assigned profiles

(a) 22$^{\text{nd}}$ period (2002 – 2006)   (b) 25$^{\text{th}}$ period (current since 2013)

Fig. 4: Politician relation graphs

6   Conclusions

In this paper the rationale behind and the necessary steps for building an online system that allows network analysis on top of parliamentary political discourse were presented. It is highlighted how such systems may contribute to more transparent policy making in the future by allowing laymen to visualise and analyse interlinks between political figures and topics. The architecture of a computer system was presented that allows for automatic information retrieval from relevant Open Data repositories, the parsing and conversion of the data into network data, and allows the application of methods from graph theory and graph visualisation in final analysis steps. The mere technical feasibility of the architecture was demonstrated by implementing an Open Source prototype of this architecture and its practical feasibility was demonstrated by putting the system in use with data scraped from the transcripts available at the Open Data repository of the Austrian parliament.

The presented approach, the architecture, and the resulting software system are work in progress. In future work the presented system can be extended in multiple ways. (1) With the continuous trend towards Open Data hopefully future transcripts of parliamentary discourse are already pre-annotated such that a higher data quality can be reached and errors in the loading process can be reduced to a minimum. (2) The presented software prototype and its analysis mechanisms are just the tip of the iceberg of which would be possible in the future. Users could enter the system through different analytics paths such as looking up all contributions to discourse of politicians, browsing through topics and finding relevant key players, and cross-referencing the official political discourse with material found in mass-media. Further, more metrics such as the automatic estimation cohesion and clout seem logical next steps, however, would require input and verification from other disciplines.

Although studies of various political institutions exist from the U.S. and Europe this is the first approach to build a generic framework that allows to import data from different countries. In future iterations it is believed that an application framework like the presented can be used to compare political bodies of different countries. This is also the first study that applies network analysis over data provided by the Austrian parliament.

Due to online social networks that allow direct political discourse among citizens, the trend towards Open Data, and systems like the presented that make use of the available data, future citizens have powerful tools at hand that shed clear light into the decision making process of governmental bodies.

References

1. Renzo Lucioni. Senate Voting Relationships. Private Blog, December 2013, `http://www.renzolucioni.com/senate-voting-relationships/`, last retrieved: 8th of March 2016.
2. Mike Krumboltz. The Splitting of the Senate (Now in Convenient GIF Form). Yahoo News, November 2013, `http://news.yahoo.com/`

`the-splitting-of-the-senate--now-in-convenient-gif-form-213908185.`
`html`, last retrieved: 8th of March 2016.
3. United States of Amoeba. The Economist. Print Edition 7th of December 2013.
4. GovTrack; Civic Impulse LCC. `https://www.govtrack.us/data/congress/`, last retrieved: 8th of March 2016.
5. Mason. A. Porter, Peter J. Mucha, M. E. J. Newman, and Casey M. Warmbrand. A network analysis of committees in the U.S. House of Representatives. In Proceedings of the National Academy of Sciences, vol. 102, no. 20, pp. 70577062, May 2005.
6. Alessia Amelio and Clara Pizzuti. Analyzing Voting Behavior in Italian Parliament: Group Cohesion and Evolution. In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2012, pp. 140146.
7. Mathieu Jacomy, Sebastien Heymann, Tommaso Venturini, and Mathieu Bastian. ForceAtlas2, A Graph Layout Algorithm for Handy Network Visualization. Gephi Consortium, Aug. 2011.
8. Marijn Janssen, Yannis Charalabidis, and Aanneke Zuiderwijk. Benefits, Adoption Barriers and Myths of Open Data and Open Government. Information Systems Management (ISM), vol. 29, no.4, pp. 258-268, 2012.
9. Nigel Shadbolt, Kieron O'Hara, Tim Berners-Lee, Nicholas Gibbins, Hugh Glaser, Wendy Hall, m.c. schraefel. Linked Open Government Data: Lessons from Data.gov.uk. IEEE Intelligent Systems, vol. 27, no. 3, pp. 16-24, May-June, 2012.
10. Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A. Zighed. Information Diffusion in Online Social Networks: A Survey. ACM SIGMOD Record, 2013.
11. Chien-leng Hsu, Se Jung Park, and Han Woo Park. Political Discourse Among Key Twitter Users: The Case of Sejong City in South Korea. Journal of Contemporary Eastern Asia, vol. 12, no. 1, pp. 65-79, 2013.
12. Matthew J. Kushin, and Kelin Kitchener. Getting Political on Social Network Sites: Exploring Online Political Discourse on Facebook. First Monday, vol. 14, no. 11, 2009.
13. Judith Bara, Albert Weale, and Aude Bicquelet. Analysing Parliamentary Debate with Computer Assistance. Swiss Political Science Review, vol. 13, no. 4, pp. 577-605, 2007.
14. Saatviga Sudhara, Giuseppe A Veltri, and Nello Cristianini. Automated Analysis of the US Presidental Elections using Big Data and Network Analysis. Big Data & Society, pp. 1-28, January-June, 2015.
15. Usha N. Raghavan, Rka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E, vol. 76, no. 3, p. 036106, Sep. 2007.