

Tracking the Dynamic Evolution of Participant Salience in a Discussion

Ahmed Hassan
University of Michigan
hassanam@umich.edu

Anthony Fader
University of Michigan
afader@umich.edu

Michael H. Crespin
University of Georgia
crespin@uga.edu

Kevin M. Quinn
Harvard University
kquinn@fsa.harvard.edu

Burt L. Monroe
Pennsylvania State University
burtmonroe@psu.edu

Michael Colaresi
Michigan State University
colaresi@msu.edu

Dragomir R. Radev
University of Michigan
radev@umich.edu

Abstract

We introduce a technique for analyzing the temporal evolution of the salience of participants in a discussion. Our method can dynamically track how the relative importance of speakers evolve over time using graph based techniques. Speaker salience is computed based on the eigenvector centrality in a graph representation of participants in a discussion. Two participants in a discussion are linked with an edge if they use similar rhetoric. The method is dynamic in the sense that the graph evolves over time to capture the evolution inherent to the participants salience. We used our method to track the salience of members of the US Senate using data from the US Congressional Record. Our analysis investigated how the salience of speakers changes over time. Our results show that the scores can capture speaker centrality in topics as well as events that result in change of salience or influence among different participants.

1 Introduction

There are several sources of data that record speeches or participations in debates or discussions among a group of speakers or participants. Those include parliamentary records, blogs, and news groups. This data represents a very important and unexploited source of information that contains several trends and ideas. In any debate or discussion, there are certain types of persons who

influence other people and pass information or advice to them. Those persons are often regarded as experts in the field or simply influential people and they tend to affect the ideas and rhetoric of other participants. This effect can be tracked down by tracking the similarity between different speeches. We can then imagine a debate with many people arguing about many different things as a network of speeches or participations interacting with each other. We can then try to identify the most salient or important participants by identifying the most central speeches in this network and associating them with their speakers. When we have a large dataset of debates and conversations that expand over a long period of time, the salience of participants becomes a dynamic property that changes over time. To capture this dynamic nature of the process, the graph of speeches must evolve over time such that we have a different graph at each instance of time that reflects the interaction of speeches at this instant.

We apply our method to the *US Congressional Record*. The *US Congressional Record* documents everything said and done in the US Congress House and Senate. The speeches in this data set are made by a large number of people over a long period of time. Using political speeches as test data for the proposed method adds an extra layer of meaning onto the measure of speakers salience. Speaker salience of the Congress members can reflect the importance or influence in the US legislative process. The way salience scores evolve over time can answer several interesting issues like how the influence of the speakers vary with majority status and change of party control. It can also study the dynamics of the relative distribution of attention to each topic area in different time periods.

© 2008. Licensed under the *Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported* license (<http://creativecommons.org/licenses/by-nc-sa/3.0/>). Some rights reserved.

The rest of this paper will proceed as follows. Section 2 reviews some related work. In Section 3, we describe how the data can be clustered into different topic clusters. In Section 4, we describe our method for computing the salience of different participant in a discussion, we also describe how to the network of speakers varies over time. Section 5 describes the experimental setup. Finally, we present the conclusions in Section 6.

2 Related Work

Several methods have been proposed for identifying the most central nodes in a network. Degree centrality, closeness, and betweenness (Newman, 2003) are among the most known methods for measuring centrality of nodes in a network. Eigenvector centrality is another powerful method that that has been applied to several types of networks. For example it has been used to measure centrality in hyperlinked web pages networks (Brin and Page, 1998; Kleinberg, 1998), lexical networks (Erkan and Radev, 2004; Mihalcea and Tarau, 2004; Kurland and Lee, 2005; Kurland and Lee, 2006), and semantic networks (Mihalcea et al., 2004).

The interest of applying natural language processing techniques in the area of political science has been recently increasing.

(Quinn et al., 2006) introduce a multinomial mixture model to cluster political speeches into topics or related categories. In (Porter et al., 2005), a network analysis of the members and committees of the US House of Representatives is performed. The authors prove that there are connections linking some political positions to certain committees. This suggests that there are factors affecting committee membership and that they are not determined at random. In (Thomas et al., 2006), the authors try to automatically classify speeches, from the US Congress debates, as supporting or opposing a given topic by taking advantage of the voting records of the speakers. (Fader et al., 2007) introduce MavenRank, which is a method based on lexical centrality that identifies the most influential members of the US Senate. It computes a single salience score for each speaker that is constant over time.

In this paper, we introduce a new method for tracking the evolution of the salience of participants in a discussion over time. Our method is based on the ones described in (Erkan and Radev,

2004; Mihalcea and Tarau, 2004; Fader et al., 2007), The objective of this paper is to dynamically rank speakers or participants in a discussion. The proposed method is dynamic in the sense that the computed importance varies over time.

3 Topic Clusters

Before applying the proposed method to a data set with speeches in multiple topics, we first need to divide the speech documents into topic clusters. We used the model described in (Quinn et al., 2006) for this purpose. The model presented in this paper assumes that the probabilities of a document belonging to a certain topic varies smoothly over time and the words within a given document have exactly the same probability of being drawn from a particular topic (Quinn et al., 2006). These two properties make the model different than standard mixture models (McLachlan and Peel, 2000) and the latent Dirichlet allocation model of (Blei et al., 2003). The model of (Quinn et al., 2006) is most closely related to the model of (Blei and Lafferty, 2006), who present a generalization of the model used by (Quinn et al., 2006).

The output from the topic model is a $D \times K$ matrix \mathbf{Z} where D is the number of speeches, K is the number of topics and the element z_{dk} represents the probability of the d th speech being generated by topic k . We then assign each speech d to the k th cluster where $k = \arg \max_j z_{dj}$. If the maximum value is not unique, one of the clusters having the maximum value is arbitrary selected.

4 Speaker Centrality

In this section we describe how to build a network of speeches and use it to identify speaker centrality. We also describe how to generate different projections of the network at different times, and how to use those projection to get dynamic salience scores.

4.1 Computing Speaker Salience

The method we used is similar to the methods described in (Erkan and Radev, 2004; Mihalcea and Tarau, 2004; Kurland and Lee, 2005), which were originally used for ranking sentences and documents in extractive summarization and information retrieval systems.

A collection of speeches can be represented as a network where similar speeches are linked to each other. The proposed method is based on

the premise that important speeches tend to be lexically similar to other important speeches, and important speeches tend to belong to important speakers. Hence given a collection of speeches and a similarity measure, we can build a network and define the centrality score of a speech recursively in terms of the scores of other similar speeches. Later, we can compute the salience of a speaker as the sum of the centrality measure of all his speeches.

To measure the similarity between two speeches, we use the bag-of-words model to represent each sentence as an N -dimensional vector of tf-idf scores, where N is the number of all possible words in the target language. The similarity between two speeches is then computed using the cosine similarity between the two vectors.

A vector of term frequencies is used to represent each speech. Those term frequencies are weighted according to the relative importance of the given term in the cluster.

The vectors representing speeches contain term frequencies (or tf), which are weighted according to their inverse document frequencies to account for the relative importance of the given term in the cluster. The inverse document frequency of a term w is given by (Sparck-Jones, 1972)

$$\text{idf}(w) = \log \left(\frac{N}{n_w} \right) \quad (1)$$

where n_w is the number of speeches in the cluster containing the term w , and N is the number of documents in the cluster. We calculated idf values specific to each topic, rather than to all speeches. We preferred to use *topic-specific* idf values because the relative importance of words may vary from one topic to the other.

The tf-idf cosine similarity measure is computed as the cosine of the angle between the tf-idf vectors. It is defined as follows:

$$\frac{\sum_{w \in u, v} \text{tf}_u(w) \text{tf}_v(w) \text{idf}(w)^2}{\sqrt{\sum_{w \in u} (\text{tf}_u(w) \text{idf}(w))^2} \sqrt{\sum_{w \in v} (\text{tf}_v(w) \text{idf}(w))^2}}, \quad (2)$$

The choice of tf-idf scores to measure speech similarity is an arbitrary choice. Some other possible similarity measures are edit distance, language models (Kurland and Lee, 2005), or generation probabilities (Erkan, 2006).

The recursive definition of the score of any speech s in the speeches network is given by

$$p(s) = \sum_{t \in \text{adj}[s]} \frac{p(t)}{\text{deg}(t)} \quad (3)$$

where $\text{deg}(t)$ is the degree of node t , and $\text{adj}[s]$ is the set of all speeches adjacent to s in the network. This can be rewritten in matrix notation as:

$$\mathbf{p} = \mathbf{p}\mathbf{B} \quad (4)$$

where $\mathbf{p} = (p(s_1), p(s_2), \dots, p(s_N))$ and the matrix \mathbf{B} is the row normalized similarity matrix of the graph

$$\mathbf{B}(i, j) = \frac{\mathbf{S}(i, j)}{\sum_k \mathbf{S}(i, k)} \quad (5)$$

where $\mathbf{S}(i, j) = \text{sim}(s_i, s_j)$. Equation (4) shows that the vector of salience scores \mathbf{p} is the left eigenvector of \mathbf{B} with eigenvalue 1.

The matrix \mathbf{B} can be thought of as a stochastic matrix that acts as transition matrix of a Markov chain. An element $\mathbf{X}(i, j)$ of a stochastic matrix specifies the transition probability from state i to state j in the corresponding Markov chain. And the whole process can be seen as a Markovian random walk on the speeches graph. To help the random walker escape from periodic or disconnected components, (Brin and Page, 1998) suggests reserving a small escape probability at each node that represents a chance of jumping to any node in the graph, making the Markov chain irreducible and aperiodic, which guarantees the existence of the eigenvector.

Equation (4) can then be rewritten, assuming a uniform escape probability, as:

$$\mathbf{p} = \mathbf{p}[d\mathbf{U} + (1 - d)\mathbf{B}] \quad (6)$$

where N is the total number of nodes, \mathbf{U} is a square matrix with $\mathbf{U}(i, j) = 1/N$ for all i, j , and d is the escape probability chosen in the interval $[0.1, 0.2]$ (Brin and Page, 1998).

4.2 Dynamic Salience Scores

We use the time stamps associated with the data to compute dynamic salience scores $p_T(u)$ that identify central speakers at some time T . To do this, we create a speech graph that evolves over time.

Let T be the current date and let u and v be two speech documents that occur on days t_u and t_v . Our goal is to discount the lexical similarity of u and v based on how far apart they are. One way to do this is by defining a new similarity measure $s(u, v; T)$ as:

$$s(u, v; T) = \text{tf-idf-cosine}(u, v) \cdot f(u, v; T) \quad (7)$$

where $f(u, v; T)$ is a function taking values in $[0, 1]$.

If $f(u, v; T) = 1$ for all u, v , and T , then time is ignored when calculating similarity and $p_T(u) = p(u)$. On the other hand, suppose we let

$$f(u, v; T) = \begin{cases} 1 & \text{if } t_u = t_v = T, \\ 0 & \text{else.} \end{cases} \quad (8)$$

This removes all edges that link a speech, occurring at some time T , to all other speeches occurring at some time other than T and the ranking algorithm will be run on what is essentially the subgraph of documents restricted to time T (although the isolated speech documents will receive small non-zero scores because of the escape probability from Section 4.1). These two cases act as the extreme boundaries of possible functions f : in the first case time difference has no effect on document similarity, while in the second case two documents must occur on the same day to be similar.

We use the following time weight functions in our experiments. In each case, we assume that the speeches represented by speech documents u and v have already occurred, that is, $t_u, t_v \leq T$. We will use the convention that $f(u, v; T) = 0$ if $t_u > T$ or $t_v > T$ for all time weight functions, which captures the idea that speeches that have not yet occurred have no influence on the graph at time T . Also define

$$\text{age}(u, v; T) = T - \min\{t_u, t_v\} \quad (9)$$

which gives the age of the oldest speech document from the pair u, v at time T .

- *Exponential*: Given a parameter $a > 0$, define

$$f_{\text{exp},a}(u, v; T) = e^{-a \text{age}(u, v; T)}. \quad (10)$$

This function will decrease the impact of similarity as time increases in an exponential fashion. a is a parameter that controls how fast this happens, where a larger value of a makes earlier speeches have a small impact on current scores and a smaller value of a means that earlier speeches will have a larger impact on current scores.

- *Linear*: Given $b > 0$, define

$$f_{\text{lin},d}(u, v; T) = \begin{cases} 1 - \frac{1}{b} \text{age}(u, v; T) & \text{if } \text{age}(u, v; T) \leq b \\ 0 & \text{if } \text{age}(u, v; T) > b \end{cases} \quad (11)$$

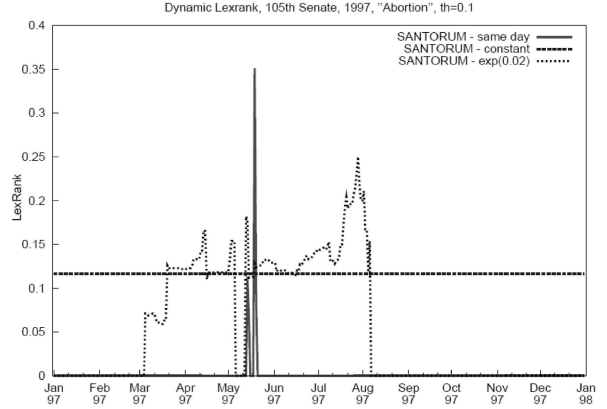


Figure 1: The Dynamic boundary cases for Senator Santorum.

This function gives speech documents that occur at time T full weight and then decreases their weight linearly towards time $T + b$, where it becomes 0.

- *Boundary*: Given $d \geq 0$, define

$$f_{\text{bnd},d}(u, v; T) = \begin{cases} 1 & \text{if } \text{age}(u, v; T) \leq d \\ 0 & \text{if } \text{age}(u, v; T) > d \end{cases} \quad (12)$$

This function gives speech documents occurring within d days of T the regular tf-idf similarity score, but sets the similarity of speech documents occurring outside of d days to 0. The case when $d = 0$ is one of the boundary cases explained above.

Figure 1 gives an example of different time weighting functions for Senator Rick Santorum (R - Pennsylvania) on topic 22 (Abortion) during 1997, the first session of the 105th Congress. The dashed line shows the case when time has no effect on similarity (his score is constant over time), while the solid line shows the case where only speeches on the current day are considered similar (his score spikes only on days where he speaks and is near zero otherwise). The dotted line shows the case when the influence of older speeches decreases exponentially, which is more dynamic than the first case but smoother than the second case.

5 Experiments and Results

5.1 Data

We used the United States Congressional Speech corpus (Monroe et al., 2006) in our experiment.

This corpus is in XML formatted version of the electronic *United States Congressional Record* from the Library of Congress¹. The *Congressional Record* is a verbatim transcript of the speeches made in the US House of Representatives and Senate and includes tens of thousands of speeches per year (Monroe et al., 2006). The data we used cover the period from January 2001 to January 2003.

5.2 Experimental Setup

We used results from (Quinn et al., 2006) to get topic clusters from the data, as described in Section 3. The total number of topics was 42. The average sized topic cluster had several hundred speech documents (Quinn et al., 2006).

We set up a pipeline using a Perl implementation of the proposed method. We ran it on the topic clusters and ranked the speakers based on the centrality scores of their speeches. The graph nodes were speech documents. A speaker's score was determined by the average of the scores of the speeches given by that speaker. After comparing the different time weighting function as shown in Figure 1, we decided to use the exponential time weight function for all the experiments discussed below. Exponential time weighting function decreases the impact of similarity as time increases in an exponential fashion. It also allows us to control the rate of decay using the parameter a .

5.3 Baseline

We compare the performance of our system to a simple baseline that calculates the salience of a speaker as a weighted count of the number of times he has spoken. The baseline gives high weight to recent speeches. The weight decreases as the speeches gets older. The salience score of a speaker is calculate as follows:

$$BS(i) = \sum_d \gamma^{d_0-d} * S_d^i \quad (13)$$

Where $BS(i)$ is the baseline score of speaker i , γ is the discounting factor, d_0 is the current date, and S_d^i is the number of speeches made by speaker i at date d . We used $\gamma = 0.9$ for all our experiments.

5.4 Results

One way to evaluate the dynamic salience scores, is to look at changes when party control of the

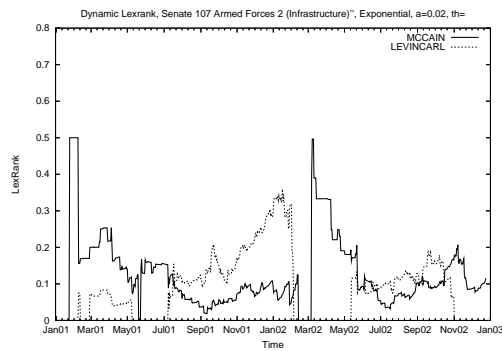
chamber switches. Similar to (Hartog and Monroe, 2004), we exploit the party switch made by Senator Jim Jeffords of Vermont and the resulting change in majority control of the Senate during the 107th Congress as a quasi-experimental design. In short, Jeffords announced his switch on May 24, 2001 from Republican to Independent status, effective June 6, 2001. Jeffords stated that he would vote with the Democrats to organize the Senate, giving the Democrats a one-seat advantage and change control of the Senate from the Republicans back to the Democrats. This change of majority status during the 107th Congress allows us to ignore many of the factors that could potentially influence dynamic salience scores at the start of a new congress.

On average, we expect committee chairs or a member of the majority party to be the most important speaker on each topic followed by ranking members or a member of the minority party. If our measure is capturing dynamics in the centrality of Senators, we expect Republicans to be more central before the Jeffords switch and Democrats becoming central soon afterwards, assuming the topic is being discussed on the Senate floor. We show that the proposed technique captures several interesting events in the data and also show that the baseline explained above fails to capture the same set of events.

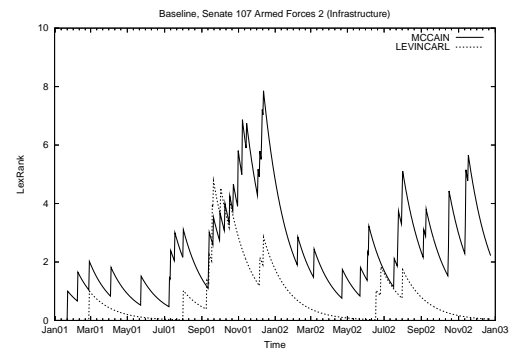
Figure 2(a) shows the dynamic salience scores over time for Senator John McCain (R - Arizona) and Senator Carl Levin (D - Michigan) on topic 5 (Armed Forces 2) for the 107th Senate. McCain was the most salient speaker for this topic until June 2001. Soon after the change in majority status a switch happened and Levin, the new chair of Senate Armed Services, replaced McCain as the most salient speaker. On the other hand, Figure 2(b) shows the baseline scores for the same topic and same speakers. We notice here that the baseline failed to capture the switch of salience near June 2001.

We can also observe similar behavior in Figure 3(a). This figure shows how Senate Majority Leader Trent Lott (R - Mississippi) was the most salient speaker on topic 35 (Procedural Legislation) until July 2001. Topic 35 does not map to a specific committee but rather is related to maneuvering bills through the legislative process on the floor, a job generally delegated to members in the Senate leadership. Just after his party gained

¹<http://thomas.loc.gov>

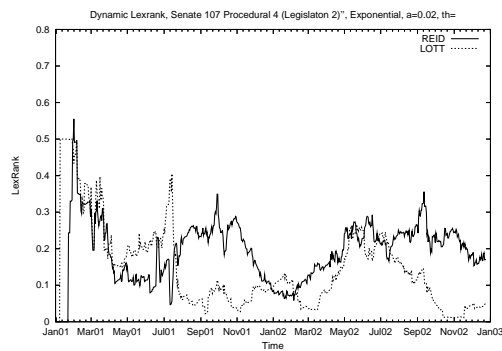


(a) Dynamic Lexrank

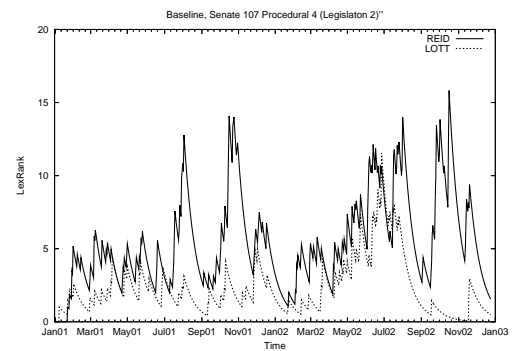


(b) Baseline

Figure 2: The Switch of Speakers Salience near Jun 2001 for Topic 5(Armed Forces 2).



(a) Dynamic Lexrank



(b) Baseline

Figure 3: The Switch of Speakers Salience near Jun 2001 for Topic 35(Procedural Legislation).

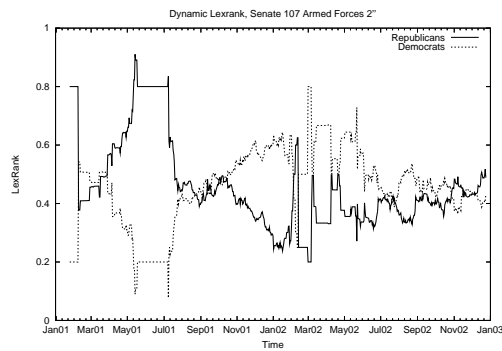
majority status, Senator Harry Reid (D - Nevada) became the most salient speaker for this topic. This is consistent with Reid's switch from Assistant minority Leader to Assistant majority Leader. Again the baseline scores for the same topic and speakers in Figure 3(b) fails to capture the switch.

An even more interesting test would be to check whether the Democrats in general become more central than Republicans after the Jeffords switch. Figure 4(a) shows the normalized sum of the scores of all the Democrats and all the Republicans on topic 5 (Armed Forces 2) for the 107th Senate. The figure shows how the Republicans were most salient until soon after the Jeffords switch when the Democrats regained the majority and became more salient. We even discovered similar behavior when we studied how the average salience of Democrats and Republicans change across all topics. This is shown in Figure 5(a) where we can see that the

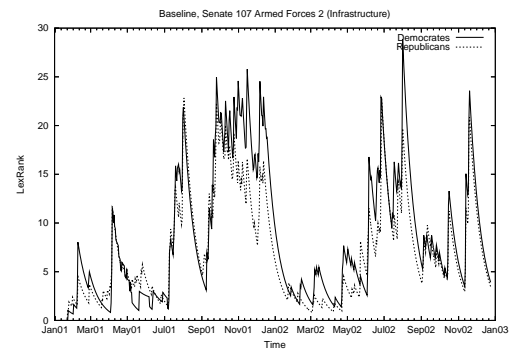
Republicans were more salient on average for all topics until June 2001. Soon after the change in majority status, Democrats became more central. Figures 4(b) and 5(b) show the same results using the baseline system. We notice that the number of speeches made by the Democrats and the Republicans is very similar in most of the times. Even when one of the parties has more speeches than the other, it does not quite reflect the salience of the speakers or the parties in general.

An alternative approach to evaluate the dynamic scores is to exploit the cyclical nature of the legislative process as some bills are re-authorized on a fairly regular time schedule. For example, the farm bill comes due about every five years. As a new topic is coming up for debate, we expect the saliency scores for relevant legislators to increase.

Figure 6 shows the dynamic scores of Senator Thomas Harkin (D - Iowa), and Senator Richard

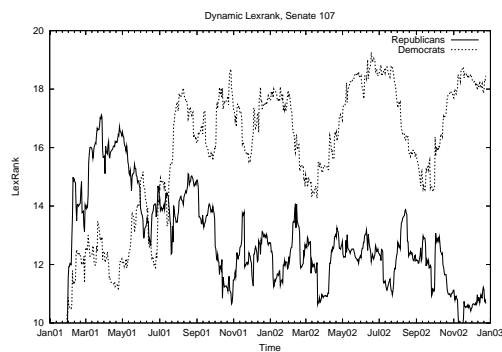


(a) Dynamic Lexrank

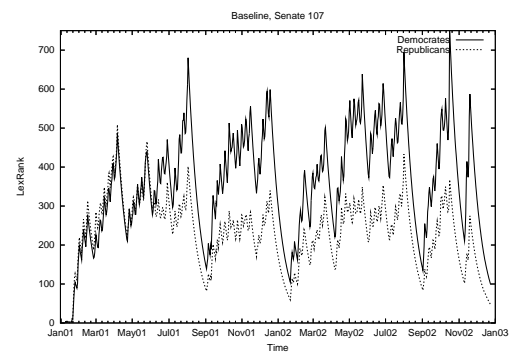


(b) Baseline

Figure 4: The Switch of Speakers Salience near Jun 2001 for Topic 5(Armed Forces 2), Republicans vs Democrats.



(a) Dynamic Lexrank



(b) Baseline

Figure 5: The Switch of Speakers Salience near Jun 2001 for All Topics, Republicans vs Democrats.

Lugar (R - Indiana) during the 107th senate on topic 24 (Agriculture). The two senators were identified, by the proposed method, as the most salient speakers for this topic, as expected, since they both served as chairmen of the Senate Committee on Agriculture, Nutrition, and Forestry when their party was in the majority during the 107th Senate. This committee was in charge of shepherding the Farm Bill through the Senate. The scores of both senators on the agriculture topic significantly increased starting late 2001 until June 2002. The debate began on the bill starting in September of 2001 and it was not passed until May 2002.

6 Conclusion

We presented a graph based method for analyzing the temporal evolution of the salience of par-

ticipants in a discussion. We used this method to track the evolution of salience of speakers in the *US Congressional Record*. We showed that the way salience scores evolve over time can answer several interesting issues. We tracked how the influence of the speakers vary with majority status and change of party control. We also show how a baseline system that depends on the number of speeches fails to capture the interesting events captured by the proposed system. We also studied the dynamics of the relative distribution of attention to each topic area in different time periods and captured the cyclical nature of the legislative process as some bills are re-authorized on a fairly regular time schedule.

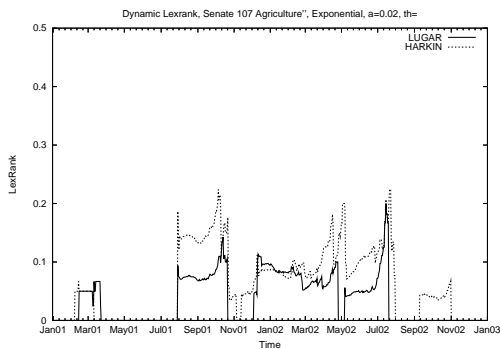


Figure 6: The Farm Bill Discussions on the Relative Distribution of Attention to Topic 24 (Agriculture).

Acknowledgments

This paper is based upon work supported by the National Science Foundation under Grant No. 0527513, "DHB: The dynamics of Political Representation and Political Rhetoric". Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

- Blei, David and John Lafferty. 2006. Dynamic topic models. In *ICML 2006*.
- Blei, David, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Brin, Sergey and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *CNIS*, 30(1–7):107–117.
- Erkan, Güneş and Dragomir Radev. 2004. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*.
- Erkan, Gunes. 2006. Language model-based document clustering using random walks. In *HLT/NAACL 2006*, pages 479–486. Association for Computational Linguistics.
- Fader, Anthony, Dragomir Radev, Michael Crespín, Burt Monroe, Kevin Quinn, and Michael Colaresi. 2007. Mavenrank: Identifying influential members of the us senate using lexical centrality. In *EMNLP 2007*.
- Hartog, Chris Den and Nathan Monroe. 2004. The value of majority status: The effect of jeffords's switch on asset prices of republican and democratic firms. *Legislative Studies Quarterly*, 33:63–84.
- Kleinberg, Jon. 1998. Authoritative sources in a hyperlinked environment. In *the ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677.
- Kurland, Oren and Lillian Lee. 2005. PageRank without hyperlinks: Structural re-ranking using links induced by language models. In *SIGIR 2005*, pages 306–313.
- Kurland, Oren and Lillian Lee. 2006. Respect my authority! HITS without hyperlinks, utilizing cluster-based language models. In *SIGIR 2006*, pages 83–90.
- McLachlan, Geoffrey and David Peel. 2000. *Finite Mixture Models*. New York: Wiley.
- Mihalcea, Rada and Paul Tarau. 2004. TextRank: Bringing order into texts. In *EMNLP 2004*.
- Mihalcea, Rada, Paul Tarau, and Elizabeth Figa. 2004. Pagerank on semantic networks, with application to word sense disambiguation. In *COLING 2004*, pages 1126–1132.
- Monroe, Burt, Cheryl Monroe, Kevin Quinn, Dragomir Radev, Michael Crespín, Michael Colaresi, Anthony Fader, Jacob Balazer, and Steven Abney. 2006. United states congressional speech corpus. Department of Political Science, The Pennsylvania State University.
- Newman, Mark. 2003. A measure of betweenness centrality based on random walks. Technical Report cond-mat/0309045, Arxiv.org.
- Porter, Mason, Peter Mucha, Miark Newman, and Casey Warmbrand. 2005. A network analysis of committees in the U.S. House of Representatives. *PNAS*, 102(20).
- Quinn, Kevin, Burt Monroe, Michael Colaresi, Michael Crespín, and Dragomir Radev. 2006. An automated method of topic-coding legislative speech over time with application to the 105th–108th U.S. senate. In *Midwest Political Science Association Meeting*.
- Sparck-Jones, Karen. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–20.
- Thomas, Matt, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *EMNLP 2006*, pages 327–335.