2A1-3

時系列テキストデータ可視化手法を用いた議論解析

Analyzing discussion using a visualizing method for time series text data

平岡 美那子[*1]
Minako Hiraoka

大澤 幸生[*1]
Yukio Ohsawa

[*1] 東京大学大学院 工学系研究科システム創成学専攻
Department of Systems Innovations, School of Engineering, The University of Tokyo

To understand the flow of discussion, or discover new viewpoints, a structured discussion map is more useful than a mere record of the discussion. However, it is hard to compose a discussion map automatically because of the difficulty of extracting logics from natural language. In this paper, we propose an improved visualizing method for time series text data. We focus on the variations in the appearance frequencies of every word in discussion and utilize them to put words in a discussion map like a clock, or a railway map. Also, words are grouped into clusters and displayed together. Some words are linked if speakers said in the same remark. A set of links expresses the rough flow of discussion. We implement this method and apply it to a record of a party leaders' debate session. The result of application shows that our method is helpful to recognize the flow and viewpoints of discussion.

1. はじめに

過去に行われた議論の内容は，関係者が同じ知識を共有するために，議事録として保管されている場合が多い．議事録から議論構造が抽出できれば，過去に行われた議論の把握，新しい論点・視点の発見等，様々に活用することができる．しかし，自然言語で記述された個々の発言を自動的に厳密な議論モデルに変換することは難しい．そこで我々は，議論に出現する単語の時間的なばらつきに着目し，ラフな議論構造を抽出する手法を提案している[Hiraoka 14]．本稿では、前述の手法を基に，(1)議題の変化を考慮した，マップ上の単語のクラスタリング(2)ユーザが議論マップの粒度を自由に変えることの出来る，インタラクティブなビューワの構築，以上2点の改良を加えた手法を提案する．

2. 関連研究

既存の議論構造の可視化に対するアプローチは，大きく分けて以下の3種類にまとめられる．(1)人が自らデータを構造化する方法[Kirschner 03]．(2)テキスト中の“手がかり表現“や，文の接続情報を利用する方法[Satoh 06](3)議論に出現する単語の変遷から，議論の構造を推定する方法[Matsumura 3]．本手法は，自動的かつ後の出現頻度を利用して構造化を行っている点で，(3)のアプローチに近い．しかし，発言者の違いに着目している点，議論マップにおいて、個々の単語の座標に意味があるという点で既存手法とは異なっている．

3. 提案手法

3.1 議論マップの作成

本手法により作成される議論マップの形状を図1に示す．提案する議論マップでは，議論に出現する単語の分散・平均を基に，中央に一般的な語，外側に特徴的な語を配置しており，時計回りに議論の流れが追えるように設計されている．以下，計算手法を簡単に示す．ここで，分析対象となる議事録データには，各発言の発言時間と発言者の情報が付与されているものとする．

連絡先：平岡美那子，東京大学 工学系研究科 システム創成学専攻

(1)各発言に対し形態素解析を行い，単語文書行列を作成する．(2)各単語について，出現時間の標準偏差と平均を計算し，図2に示される計算式によりマップに配置する．ここで，出現頻度のピークが複数ある単語に対応するために，混合ガウス分布を仮定し，EM アルゴリズムによるクラスタリングを行う．(3)同一発言に含まれる単語間にリンクを結ぶ．以上の3ステップについて，より詳しい計算手法は，[Hiraoka 14]を参考されたい．

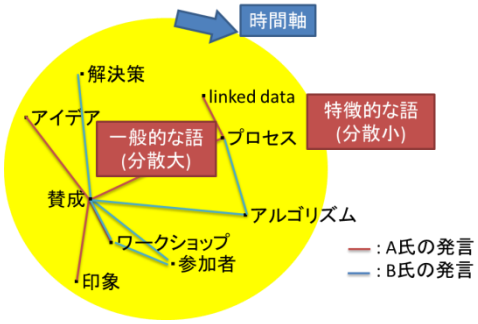


図 1 議論マップのモデル図

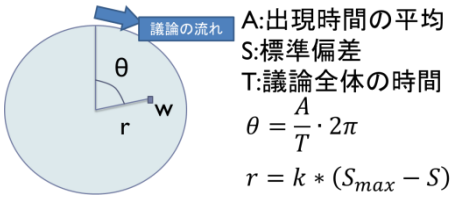


図 2 単語の配置方法

3.2 複数トピックへの対応

議題が複数含まれるような長い議事録の場合，マップが過密になってしまう．そこで，複数の議題が含まれるような長い議論データについては，議題ごとに単語をクラスタリングして表示することを考える．議題の境目においては，出現単語の組み合わせが大幅に変化すると考えられるので，クラスタリングには前節で述べた単語間のリンク情報を利用することとする．以下に，計算のアルゴリズムを示す．(1)議論全体の時間を適当な時間幅で離散化する．($T=\{T_1,T_2,\ldots\}$)(2)時間 T_i〜T_{i+1} に存在するリンク数をそれぞれ調べ，時間－リンク頻度分布を作成する．(3)

時間－リンク頻度分布に対して混合ガウス分布の推定を行い，議題の境目となる時間を決定する．(4)(3)で推定された境目の時間を基に，単語をクラスタリングする．

4. システムの実装

提案手法を，ブラウザ上で閲覧可能な Web アプリケーションとして実装した．システムの全体図を図3に示す．本システムでは，インタラクティブにマップの解像度を変更することが可能となっている．マップ表示のアルゴリズムを以下に示す．(1)ユーザが表示範囲を選択する．(2)表示範囲に含まれる単語のうち，議論全体における出現頻度の高いものから順に N 個選択する．(3)選択された単語をノードとして含むリンクを選択する．(4)(2)(3)で選択された単語(ノード)・リンクを画面に表示する．この機能により，ユーザは議論の大まかな流れを追うことも，議論の中の特定の部分の詳細を調べることも出来る．単語の複数ピークの発見及び分類，また議論全体の話題を考慮した単語分類には，統計ソフト R の Mclust パッケージを用いた．最大クラスター数を指定し，BIC 値が最も高くなる結果を採用するものとした．

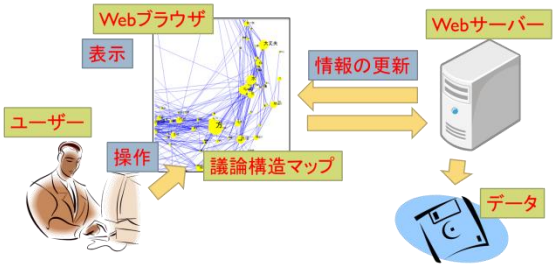


図 3 議論可視化システム

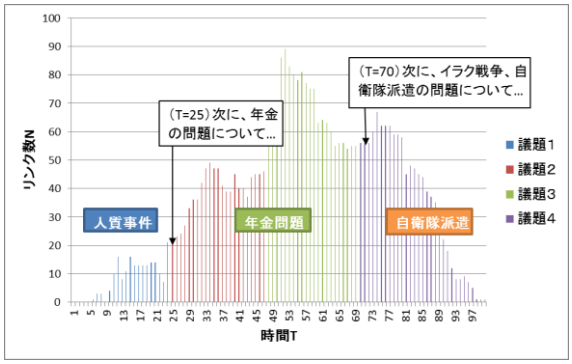
5. 適用例

実装したシステムを用いて，2004 年 11 月 10 日に行われた自民両代表の党首討論の議事録[Kokkai 14]について可視化を行った．この議事録には発言時間の情報がなかったため，議事録における改行を発言の区切りとし，発言時間を初めから順番に 1, 2…とした．表示する単語は名詞かつ総出現回数が4回以上のものとした．最大クラスター数は議題・単語どちらとも4とした．結果を表1と図 4 に示す．表1から，議題の境目がほぼ正確に検出できている事が分かる．また，発言者の使用単語やリンクの結びつきから，それぞれの主張を読み取る事が出来る．

6. おわりに

時系列テキストデータから議論マップを作成する手法を提案し，実際の議論データの可視化を行った．本手法を用いれば議事録が効果的に構造化出来ることが分かった．議論マップの解釈方法については，今後検証実験を実施したいと考えている．

表 1 リンク情報による議題のクラスタリング結果



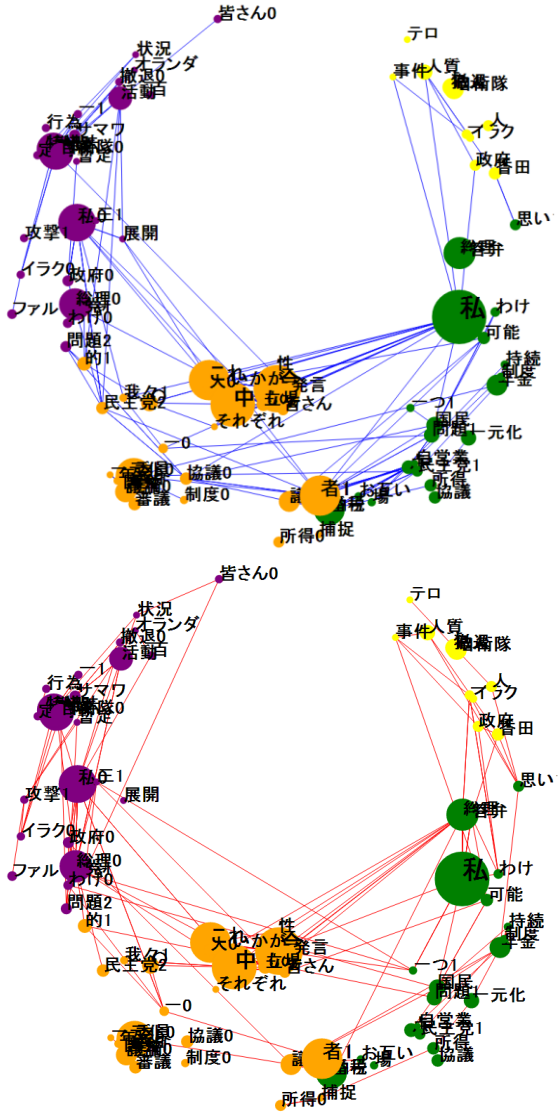


図 4 党首討論議論マップ（上：自民党 下：民主党）

参考文献

[Hiraoka 14] 平岡美那子，大澤幸生: 時系列テキストデータからの議論構造の可視化，情報処理学会第 76 回全国大会講演論文集，2014

[Kirschner 03] Paul A. Kirschner, Simon J. Buckingham Shum and Chad S. Carr (Eds.): Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making, Springer-Verlag: London, 2003.

[Satoh 06] 佐藤岳文，堀田昌英：Web マイニングを用いた因果ネットワークの自動構築手法の開発，社会技術研究論文集，2006

[Matsumura 03] 松村真宏，加藤優，大澤幸生，石塚満: 議論構造の可視化による論点の発見と理解，知識と情報：日本知能情報ファジィ学会誌，2003

[Kokkai 14]国会議事録検索システム，http://kokkai.ndl.go.jp/，(2014-1-13 アクセス)