人工知能学会合同研究会 2015 に参加して

澤 井 裕 介
Yusuke SAWAI
電子情報学専攻修士課程　2015 年度修了

1. はじめに

2015 年 11 月 12 日（木）～13 日（金）に慶應義塾大学日吉キャンパスで行われた人工知能学会合同研究会 2015 に参加し，インタラクティブ情報アクセスと可視化マイニング第 11 回研究会において『文書ストリームにおけるトピックダイナミクスの階層化ビジュアリゼーション』という題で 30 分間の発表を行った．

2. 発表内容

2.1 発表概要

日々刻々と変化する世界の情勢を把握することは，社会を生きる人々の重要な関心の対象といえる．これまで，世界の情勢は，新聞，ラジオ，テレビなどの主要メディアに携わる人々によって，限られた紙面，時間制限の中で情報の取捨選択せざるを得ない状況があった．しかし，近年，WEB 世界の発展により，世界情勢を伝える発信源は無尽蔵に増え続け，主要メディアで取り上げられなかった情報はもとより，ソーシャルメディアの発達により，これまで主要メディアの情報を知る側であった人々が意見を述べ，情報の発信源としても成長していることから WEB 空間に文書ストリームが無数に出現しており，またそこにはダイナミクスが存在し得ると予想される．そのようなトピックのダイナミクスを捉えるためには，観測できる対象として，複数の文書ストリーム間の関係を捉え，それらをわかりやすく整理，分析できる環境を構築することが望ましいと思われる．

そこで本研究では，生活や文化，政治や経済などの主要なトピックが含まれる新聞データを用いて，トピックを階層的に捉え，年間主要トピックとデイリートピックとの関係を可視化しつつ，デイリートピック間の関係や活性度の変化を時間軸に沿って視覚的に分析できる TimeLine を用いた階層的可視化ビジュアリゼーション法を提案する．本稿では，本研究の第一歩として，主要メディアのトピックを階層的に Timeline 上で捉えた際の視覚的分析に関する可能性を探るため，毎日新聞データセットを用いた実データによる実験で，ダイナミクスの一面として，トピックの分離・融合や活性度を可視化し，さらに年間主要トピックとデイリートピックとの関係度を可視化することで，文書ストリームのダイナミクスにおいて，複数の面から変動を分析することができるトピックダイナミクスの階層化ビジュアリゼーション法を提案する．

2.2 提案法

文書群の時系列データ（文書ストリーム）として 1 年間の新聞記事群を考え，そのトピックダイナミクスの可視化法を提案する．文書ストリーム D における年間レベルでの主要トピックの出現と消滅のダイナミクスを調べるために，文書群データについて多重トピックを考慮した文書の確率的生成モデルである HDP-LDA（Hierachical Dirichlet Process-Latent Dirichlet Allocation）によりモデル化する．各 t に対して，第 t 日の記事群 Dt を，それに属するすべての記事を単純につなぎ合わせて一つの文書と考え，単語頻度ベクトルを構成する．そして HDP-LDA モデルに基づいて，観測データに対する，潜在トピック集合および，各潜在トピックの下での単語生成ベクトルを，それぞれ推定する．我々は，ここでの各潜在トピックを文書ストリーム D の年間主要トピックと呼び，それら年間主要トピックスを抽出し分析する．また第 t 日にどのくらい活発であったかを，事後確率を用いて，活性度と呼ぶ．文書ストリームにおける日レベルでのトピックについて調べるために，各 t に対して，第 t 日の文書群における多重トピックを考慮した文書の確率的生成

モデルの HDP-LDA によりモデル化する．潜在トピック集合および，各潜在トピックの下での単語生成ベクトルを，それぞれ推定する．ここでの各潜在トピックを文書ストリーム D における第 t 日のデイリートピックと呼び，それらデイリートピックスを抽出し分析する．またデイリートピックがどのくらい活発であったかを，事後確率の和を用いて，活性度と呼ぶ．また，各年間主要トピックと各デイリートピック，隣接するデイリートピックの関係度を各単語出現ベクトルのコサイン類似度で分析する．

2.3 実験結果

毎日新聞データベースより，2013 年 1 月 1 日から 2013 年 12 月 31 日の新聞記事の中で 1 面，2 面，3 面，経済面，社会面の記事を文書ストリームとして使用した．また，これらの記事において，1 日に 1 回以上出現している単語という条件のもと BOW 表現に変換した．その語彙の総数は，1,333 となった．2013 年の実験データに対し，潜在トピック推定法を適用したところ，抽出された年間主要トピックの総数は 18 個となった．また，各 t 日において，デイリートピックを推定したところ，各 t 日のデイリートピック数は，図 1 に示すような結果となった．図 1 の横軸は t となっており，縦軸は各 t 日のデイリートピック数となっている．すべての日において，デイリートピックが一定数以上生成されていることや，日によって数が変動していることが確認できる．

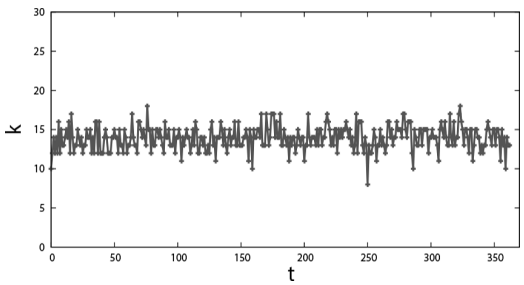

図 1　デイリートピックの変動

7 月 21 日と 22 日のデイリートピックの可視化結果を図 2 に示す．21 日は，参院選挙の開票日であり，22 日は参院選挙の開票日直後の日となる．22 日のデイリートピック 4 は，21 日のデイリートピック 1 と 3 が融合したものである．21 日のデイリートピック 1，3 および 22 日のデイリートピック 4 に対して，関連の深い新聞記事を調べたところ，21 日のデイリートピック 1 は，エジプトでの武装勢力の攻撃に関する記事や，中国の影響力を懸念する記事，中国でのテロ行為の記事であり，21 日のデイリートピック 3 は，投開票日の話題や，与党と野党の攻防の記事であった．22 日のデイリートピック 4 は，自民圧勝，ねじれ解消，アベノミクスが支持されたという記事であった．このように提案可視化法により，デイリートピックの分離・融合など詳細なトピックダイナミクスの分析ができる．

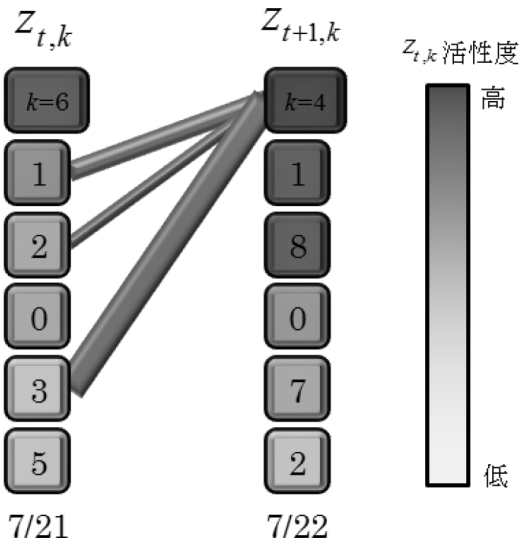

図 2　7 月 21 日，22 日の可視化結果

3. おわりに

発表を行うにあたり，ご指導いただいた木村昌弘教授，熊野雅仁実験講師，ご意見いただいた研究室の皆様に深く感謝致します．