D-016

共起タグを用いた時間変化する話題の抽出手法の提案

Time Dependent Shifts of Topics based on Co-occurrence Tags

中村 浩之† 小川 祐樹† 諏訪 博彦† 太田 敏澄†
Hiroyuki Nakamura Yuki Ogawa Hirohiko Suwa Toshizumi Ohta

1. はじめに

膨大な情報の中で人々に注目される話題は日々変化し続けており、常に変化する話題を把握するためには、あらゆるニュースをチェックする必要がある.しかし、求める話題を発見するために、膨大な時間がかかるので、世の中の話題の移り変わりを簡単に知るための手段が必要である.そのため、ニュースや関連情報をまとめたインデックスのニーズが高まっている.

本研究では、話題となるニュースを把握するために、話題の抽出と分析を行う事を目的とする.ニュース情報は、社会の変化に応じて、日々変化している.そのため、単一の動画ニュースが持つ特徴を抽出するだけでは、世の中の話題を把握する事はできない.世の中にどの程度同じニュースがあるのかを把握する事が、世の中の話題を抽出する事になる.真に有益な情報を得るために、日々更新し続ける情報に対して、評価を行う事が、世の中のニュースを把握するために必要であると考えられる.

2. 先行研究

張ら(2006)は、内閣支持率の時系列データとニューストピックとの類似度を評価して、時系列データと意味的に関連するニューストピックを抽出し、ニューストピックが社会に及ぼす影響について示唆している.

また、時系列によって世の中の関心の移り変わりをニュースの時事性から着目して話題を抽出する山崎(2008)の研究では、電子番組表(EPG)を利用して類似するキーワードの出現確率を分析する事で、キーワードと番組表のテキスト情報との共起グラフより、出現確率が増加の傾向にあるキーワードを抽出し、キーワードの時事性を評価している.しかし、この時事キーワードの抽出手法は、キーワードの出現頻度や文字列長によるフィルタリングや idf のようなキーワード自体の特徴量を考慮していない.

そこで本研究は、話題となるトピック及びニュースシーンを抽出するために、共起タグを用いた時間変化する話題の分析を行う.動画ニュースに付随するメタテキストから複合語を抽出し、tf-idf による特徴量の計算を行ったのちに、語と動画ニュースシーンと特徴量による共起グラフの構築を行う.構築した共起グラフについてクラスタリングを行い話題を抽出し、その時間的変化を観察する事により、話題の「流行性」について分析と考察を行う.

3. 共起タグを用いた話題抽出

本研究では、動画ニュースをシーン単位で扱う.これまでの動画推薦システムは、番組単位で動画を推薦している.そのため一つの推薦情報のなかに、複数のニュース情報が含まれていた.ニュース情報を個別に扱う事により、類似ニュースの放送頻度を把握する事ができ、それによりどんな話題が流行しているか把握する事ができる.話題の「流行性」を分析するために、動画ニュースシーンに付随するテキストデータを以下の通り処理する.

3.1 用語の抽出

動画ニュースシーンに付随するヘッドラインなどのテキスト情報に対して、Mecab を用いた形態素解析を行う.テキスト内の名詞を対象とした単語及び複合語の出現頻度の計算を行う.

3.2 特徴量計算

文章中の特徴的な名詞を抽出するため、tf-idf を用いて、用語の特徴量の計算を行う.

3.3 類似度ネットワーク分析

シーンごとにもっている用語の特徴量からシーン同士の類似性を cos 類似度を用いて計算し、シーン to シーンの類似度ネットワークを可視化する.

3.4 クラスタリングによる話題抽出

Newman 法を用いてシーン to シーン類似度ネットワークの modularity を計算し、値が最大になるエッジについて切り分けを行う事で、ネットワークのクラスタリングを行う.

3.5 話題の流行性評価

シーン同士の類似度ネットワークより、クラスタごとのシーン数を分析し、話題の流行性(trend)の評価を行う.

$$\mathrm{trend}(i) = \frac{\text{クラスタ}i\text{のシーン数}}{\sum_{k}^{all} \frac{\text{クラスタ}k\text{のシーン数}}{\text{全クラスタ数}}}$$

4. 評価実験

4.1 データセット

2009/03/23~03/29 に TV のキー局で放送されたニュースシーンに付随するテキスト情報を扱う.表１はデータセットの一部である.

表1 ニュースシーンごとに付随するテキストデータ

ID	放送日	番組ジャンル	ヘッドライン	メモ
10019906	3/29	ニュース/報道	千葉県銚子市長リコール住民投票・失職決定	市立総合病院の診療休止をめぐり、岡野俊昭市長のリコール投票が行われた
10019898	3/29	ニュース/報道	茨城県五霞町・超軽量飛行機墜落事故	ウルトラライトプレーンが墜落.搭乗していた男性２人死亡
…	…	…	…	…

†電気通信大学 大学院情報システム学研究科 Gratuate School of Information Systems, University of Electro-Communications

4.2 シーン to シーン類似度ネットワークの推移

節3.1から3.3の分析を行い、データセットされたニュースシーンについて、シーンtoシーンの類似度ネットワークを構築する.3/23~3/29において曜日ごとのシーンtoシーン類似度ネットワークを可視化し(図1)、各曜日のネットワークに対してNewman法による自動クラスタリングを行って、各曜日のクラスタの合計数と単位クラスタあたりの平均ノード数の計算を行った.

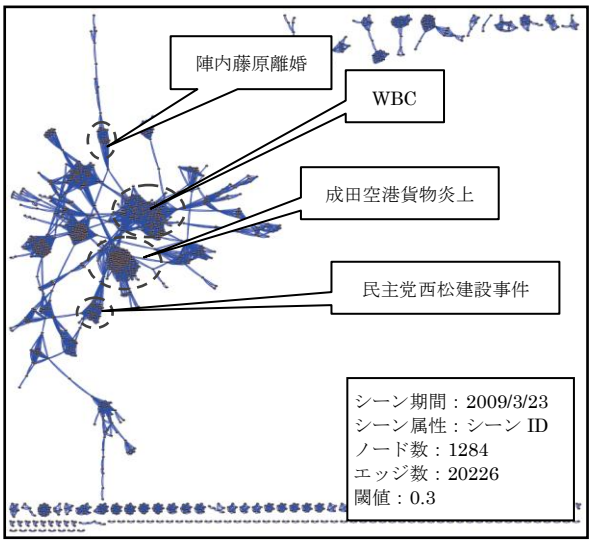

図 1. 月曜日のシーン to シーン類似度ネットワーク

尚、各曜日の類似度ネットワークの閾値は暫定的に 0.3 と設定している.これらのクラスタリングの結果からノード数の多いクラスタがどういった話題をもっているかを分析するため、クラスタ内のシーンで最も多く頻出するタグを算出している.最頻出タグの一致具合をにより話題の一致性を判断し、一週間を通して、流行性の高い話題の推移を分析し図 2 に示した.

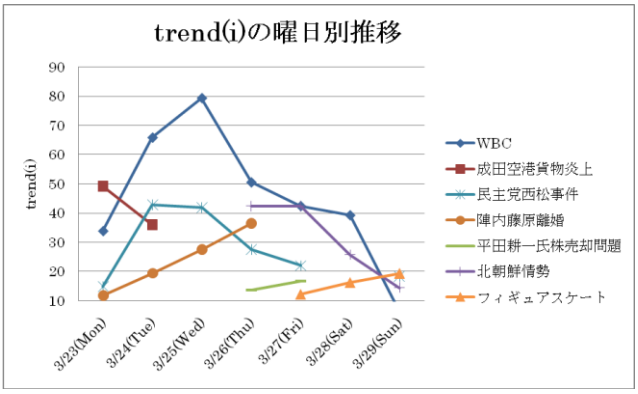

図 2 流行性(trend)の曜日別推移

5. 考察

図 2 より、流行性の推移は、急に上昇して、その後下降するタイプの話題と下降はせずに、徐々に上昇する話題がある事がわかった.基本的には前者が多く見られる結果であったが、WBC や陣内氏藤原氏の離婚に関する話題のように前日と違う新しい情報が更新される場合、流行性が上昇する傾向にあるものと考えられる.例えば WBC の場合、月曜日に準決勝で米国と試合をして勝利している.火曜日に韓国と決勝戦を行い勝利し、水曜日は WBC2 連覇を祝福したニュースが多く報道されている.木曜日に凱旋帰国を果たしており、一連の流れの中で「WBC で 2 度目の優勝をした」という段階が最も高い流行性を示している事は、妥当な推移を示していると考えられる.

一方、成田空港の貨物機炎上や北朝鮮情勢に関するニュースは、その事件が起こった日が流行性が最も高くなり、徐々に減少をしている.貨物機炎上のような通常では起こり得ないニュースは、その性質上、話題性が高いニュースとして、唐突に出現する事も妥当であると考える.

今後、流行性の推移のパターンを分析する事により、唐突に出現するニュースを新規性の視点で、捉える事も考えられる.今後の課題としたい.

また、土日にほとんど全ての話題について、流行性が下がっている事が確認されている.これは、土日には、一週間をまとめたニュース番組など要約性の高いニュースが多く報道されているものと考えられる.

より価値のある話題を抽出するために流行性以外にも着目する必要があり、話題の内容を一つのシーンで網羅した「要約性」や、ニュースの話題や内容の新しさを表した「新規性」を新たな指標として用いる事で、ニュースの価値をさらに多面的に評価していく.

これらの指標を推薦に用いる事により、ユーザーの嗜好だけでなく、動画ニュースの社会的特徴を捉えた推薦が可能となると考える.

6. 結論

我々は、話題性のあるトピックの抽出手法として、共起タグの特徴量よりニュースシーンの類似度ネットワークを形成し、ネットワークのクラスタ情報から、「流行性」の時系列変化を分析している.

上位に頻出するタグが共通する場合に同じ話題として扱う事で、時系列で連続的に変化する動画ニュースの流行性を評価する事ができている.今後は、流行性以外の話題の評価基準についても考慮し、動画ニュースがもつ話題の多面的価値を評価できる指標を提案していく.

参考文献

[1] 張 一萌，何 書勉，小山 聡,田島 敬史，“時系列データに意味的に関連するニューストピックの発見”, DEWS, (2006)
[2] 山崎 智弘, “強連結成分分解を利用した電子番組表からの話題抽出” ,Journal of the DBSJ Vol.7 , No.1 (2008)
[3] 松尾 豊，石松 満，“語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム” , Techical Papers, (2002)
[4] Hideo Fujii and W. Bruce Croft. “A comparison of indexing techniques for Japanese text retrieval., Proceedings of SIGIR ’93, pp.237-246, (1993)
[5] 大島 裕明，小山 聡，田中克己，“文書群をクエリとした似て非なる文書の検索” , DBSJ Lerrers , Vol.5, No.1 (2006)
[6] K.T. Franzi and S. Ananiadou, “Extracting nested collocations”, COLINGpp.41-46 , (1996)
[7] Hiroshi Nakagawa, “Automatic Term Recognition based on Statistics of Compound Nouns” Terminology Vol.6,No2, (2000)
[8] 形態素解析システム Mecab http://mecab.sourceforge.net/
[9] M.E.J. Newman, ”Fast algorithm for detecting community structure in network”, (2003)