人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第15回)
SIG-AM-15-10

初学者を対象としたニュース記事中のトピックの関係性に基づく可視化インタフェースの提案

Visual Interface for Novice Readers Based on Topic Relation in News Articles

西川 奈都月[1*]　盛山 将広[2]　内藤 峻[2]　松下 光範[1]
Natsuki Nishikawa[1]　Yukihiro Moriyama[1]　Shun Naito[1]　Mitsunori Matsushita[1]

[1] 関西大学総合情報学部
Faculty of Informatics, Kansai Univrsity

[2] 関西大学大学院　総合情報学研究科
Graduate School of Informatics, Kansai University

Abstract:
The aim of this research is to deepen a user's understanding of topics appeared in a series of news articles by having the user grasp the relationship between the articles about the theme they are interested in. In recent years, the number of people using Web news sites is increasing. However, news articles on the Web are diverse and enormous, it is difficult for people without expertise to grasp the relationship between articles at present. To solve the problem, We proposes a system that categorizes the articles based on the relevance of topics apeared in the news articles, reflects them in the hierarchical structure, and visualizes the relationships of the articles clearly. This paper describes the prototype of the system and verifies it.

1　はじめに

インターネットの普及によって，ニュースの受容方法が変化してきている．市場調査会社ピュー・リサーチセンター[1]の調査（2008 年 12 月）によると，インターネットを媒体としてニュースを受容する人の割合が新聞を上回った．この調査によれば，調査対象者のうち 40% の人がインターネットから国内外のニュースを得ることが多いと回答し，新聞からニュースを得ることが多い人は 35% にとどまった．また，新聞を情報源とする者の割合は 2005 年からほぼ横ばいであるが，インターネットを情報源としている人の割合は高くなっている．これらのことから，ニュースの受容方法が新聞からインターネットへと移行している傾向が伺える．

ニュース受容方法の移行の背景には，Web を介してニュースを発信する Web ニュースサイト（e.g., 朝日新聞デジタル[2]，Yahoo!ニュース[3]）の出現が挙げられる．これらの Web ニュースサイトは，誰でもニュース記事を閲覧できるというアクセスの容易性と情報更新の速さが利点である．しかし，Web ニュースサイトに掲載されている記事は多様な観点から記述され，かつその数も膨大であるため，時系列に沿った話題を整理しづらくなっている．そのため，あるテーマについての理解を深める際，そのテーマに関する事象の背景やその前後関係の把握が困難である．既存の Web ニュースサイトには，関連記事を推薦する機能を提供しているものもあるが，この機能では関連している記事を閲覧してユーザが求める新たな情報を得ることができるものの，各記事の関係性は明確に表示されない．そのため，テーマに対する予備知識や専門知識を持たない人は，複数の記事を読み比べる必要がある．

こうした背景の下，本研究では，専門知識を持たない人を対象に，あるテーマに関連している記事同士の関係性を把握しやすくすることで，テーマに対する理解を深めさせることを目指す．その方法として本稿では，(1) ニュース記事中に存在する潜在的話題（以降，トピックと記す）の関係性に基づいて複数の記事を分類して階層構造に反映する，(2) トピックの関係性をニュース記事の語句（以降，キーワードと記す）によって明確にする，という特徴を持つ可視化インタフェースを提案する．

*連絡先：関西大学総合情報学部
〒 569-1095 大阪府高槻市霊山寺町 2-1-1
E-mail: k369313@kansai-u.ac.jp
[1]http://www.people-press.org（2016/12/12 確認）
[2]http://www.asahi.com/（2016/2/17 確認）
[3]http://news.yahoo.co.jp/（2016/2/17 確認）

2 関連研究

ニュース記事を分類してユーザに提示する方法として，(1) 時間情報に着目した提示，(2) トピックに着目した提示，の 2 種類が提案されている．

(1) を企図した手法としては，高間らや菊池らの手法が挙げられる．高間らは，地震に関するニュース記事から時間的動向情報（e.g., 地震の発生回数の時間的推移）と空間的動向情報（e.g., 地震の発生場所）を抽出し，それらを利用したニュース記事内容を可視化するシステムを提案している [5]．また，菊池らは，ニュース記事中の特徴を表すキーワードの変化によって，話題の変化の把握を可能にしている [4]．この手法では，ニュース記事数の時間的変動を折れ線グラフで示し，その上に記事中に含まれる特徴的なキーワードを日付毎に表示することで，ニュース記事に含まれる話題の変化を表現している．これらのように，ニュース記事の内容を時間情報に着目して可視化することは，視覚的に対象を捉えやすくなりニュース内容の前後関係を整理することに寄与する．しかし，時系列に提示するだけでは事前知識を持たない初学者にはニュース記事の背景を知ることは難しく，ニュースに対する理解を深める効果には限界があると考えられる．

(2) を企図した手法としては，芹澤らや薦田らの手法が挙げられる．芹澤ら [2] は，ニュース記事をトピックで分類する方法を提案している．提案手法では，LDA 法を用いて対象となるニュース記事から潜在的話題（トピック）を抽出し，そのトピックの類似度によりトピック同士を関連付け，その検証を行った．その結果，時系列情報だけでなく，ニュース記事のトピックを追跡して話題の全容を把握することで，ニュースをより理解することが可能になった．また，薦田ら [6] は，文書作成者の意図を読み手に理解させるには，複数の類似文書との関連性を意識させながら主張の特徴点を把握させ，新たな知見を適切に位置づけさせることが必要であるとして，比較対象の文書同士を階層構造で表現することによりそれらの関係性を示す手法を提案している．この研究では，文書中に存在する特徴的な単語を階層構造に反映することで包括関係が強調され，単語同士の関係性が明確になっていることが示唆されている．これらの研究から，ニュースの理解には，その話題の全容を把握することが有効であること，ならびに情報を階層構造で表現することで情報同士の関係性が明確になり，ニュースの理解につながることが示唆される．ただし，これらの手法では時間的な関連性が明示されないため，特にニュースに関する事前知識を持たない初学者にとっては，ニュースの持つ経時的関連性の把握は難しいと思われる．

そのため，本稿では (1) と (2) を組み合わせて，両手法の利点を活かした可視化手法を実現を目指す．

3 デザイン指針

本研究で対象とするユーザは，ニュースを理解する上で必要となる予備知識や専門知識を十分に持たない初学者である．本研究では，その初学者のモデルとして大学生を対象とし，以下のようなインタラクションを想定する．

> 2016 年 10 月 15 日，大学生の A さんは，災害論の授業で「日本の地震の被害」についてまとめるレポート課題が与えられた．そこで，記憶に新しい「熊本地震の被害状況」についての情報を調べることにした．
> ー Phase 1 ー
> A さんは，熊本地震について，テレビなどのニュースで聞いたことがある程度の知識しかなかった．熊本地震について理解を深めるために，Yahoo!ニュースのキーワード検索で「熊本地震　被害」と検索すると膨大な記事が該当した．
> ー Phase 2 ー
> 最新の記事である検索結果画面の 1 番上の記事を読んでみた．記事内容は，記者自身が被災者にインタビューをして，地震発生から半年たった現在の復興状況や被災者の声がまとめられた記事であった．被害状況の詳細が書かれていないものの，被災者の苦労から状況の深刻さを把握でき，この記事は参考になると判断した．
> ー Phase 3 ー
> さらに，情報を集めるため，推薦されていた関連記事を読むことにした．その記事は，阿蘇の駐在さんが仮設住宅を巡回しているエピソードであった．先ほど読んだ記事と比べて新しい情報を確認し，記事内容は参考になると判断した．
> ー Phase 4 ー
> このように関連記事を読み，知らない情報であるのか，必要な内容であるのか，確認する作業を繰り返して,「熊本地震の被害状況」に関する情報は十分に収集することができたと感じた．

この例では，レポートを書くために必要なニュースの理解が深まる様子を想定している．Phase 1 では，自身が持っている知識で検索クエリを入力している．Phase 2 では，Phase 1 で得られたクエリを元に得られたニュース記事を読んで新たな知見を得ている．Phase 3 では，Phase 2 で得られたニュース記事の関連記事を読んで，初めのニュース記事と比較して新たな知見を得ている．Phase 4 では，Phase 1 から 3 の作業を繰り返している．ここで，Phase 3 でニュース記事の関連記事を読んで記事内容を比較しているのは，その記事の関連理由が明示されていないため，記事を読み

表 1: 各テーマの期間と件数

テーマ	期間	記事数
熊本地震	2016/4/15〜2016/9/23	60
東京五輪	2013/9/11〜2016/11/2	69
三菱自動車	2008/12/10〜2016/11/3	62

比べることでなぜ関連しているのかを自分自身で判断しつつテーマの理解を深めるためである．この行為を繰り返すことで，テーマへの理解が徐々に深まっていくといえる．

こうしたインタラクションを効率よく行うため，本稿ではニュース記事中のトピックを階層構造に反映し，トピックの関係性をキーワードの相違によって明確にすることでテーマの理解を促すことを企図した可視化インタフェースを提案する．

4 トピックの生成

本稿では，社会問題に関するニュース記事を扱う．ニュース記事には，Yahoo!ニュースのアーカイブ一覧に掲載されている「熊本地震」「東京五輪」「三菱自動車」の 3 つのテーマに関する内容を対象とする．収集した各テーマの期間と件数を表 1 に示す．各テーマの記事は Yahoo!ニュースを対象にスクレイピングを行い，各記事から日付，記事題目，本文を収集した．これらの収集した各ニュース記事を MeCab[4] を用いて形態素に分割し，記号，助詞，数詞を除去した．このとき，名詞連続は結合して一語とした．これらの処理によってテーマ毎に収集されたニュース記事群から，TF-IDF 法を用いて各記事を特徴づける単語を抽出した．

各テーマの記事群のある 1 つの記事から TF-IDF 法により得られた結果を表 2 に示す．なお，収集した記事は記事本文の文字数が一定ではないため，TF-IDF 法により得られた単語のうち特徴的だと判断される上位の単語のみをキーワードとして扱うのではなく，得られた全ての単語をキーワードとして扱った．

次に，hLDA 法によりトピックを抽出した．hLDA 法は LDA 法の発展的技術であり，文書内に含まれるトピックに階層関係があるという仮定に基づき，トピックの推定を行うモデルである [1]．仮定したトピックは包含関係にあるため，多くのトピックに含まれる単語は上位階層のトピックに，各トピックに特徴的な単語は下位階層のトピックに分類される傾向にある．

hLDA 法の分析を行う際には，階層数とハイパーパラメータ (α, γ, η) を予め設定する必要がある．

本稿で扱うデータは，上述した 3 テーマに関するニュース記事であるため，テーマ毎に 1 つずつ階層化

[4]https://sourceforge.net/projects/mecab/（2016/12/12 確認）

表 2: TF-IDF 法の結果の一例

熊本地震	東京五輪	三菱自動車
捜索活動	五輪商用	中継
地震死者	さまざま便乗	三菱自動車社長
熊本県警	字並び	相川
消防	開催決定	会見中継
地震警察	デザイン便乗	問題三菱自動車
熊本一連	便乗 NG	社長
揺れ観測	企業 TOKYO	会見
震度	商用便乗	燃費不正
熊本	NG 東京	–
–	東京オリンピック	–
–	東京五輪	–

されたトピックを作成した．ここで階層 1 は多くのトピックに網羅的に含まれる単語が分類され易くそのテーマ自体を表すトピックになるため，階層数を 4 に設定し，可視化に反映する階層構造は階層 2, 3, 4 とした．

ハイパーパラメータ (α, γ, η) については，γ が大きく η が小さい場合は，各トピックにおいて生成確率が高い単語が上位に分類されたトピックが生成され易くなることが確認されている [3]．本稿で扱うニュース記事数は 1 つのテーマに対して 60 〜 70 記事のため，トピックの話題が特徴的に現れる階層 4 に含まれる記事数が少なくなり情報量が不足する懸念がある．そのため，階層 4 に含まれる記事数が 3 つ以上になるという条件を満たすようにパラメータを選定した結果，$\alpha = 10.0$, $\gamma = 1.3$, $\eta = 0.8$ となった．これらの処理により生成した「熊本地震」についてのニュース記事におけるトピックを階層構造で表したものを表 3 に示す．

表 4 に各階層におけるトピックの数を示す．なお，今回の実装では，トピック名は階層ごとに生成された単語群から人手で選択し付与した．

5 提案システム

本稿で提案するシステムの構成を図 1 に示す．また，実装したプロトタイプのスクリーンショットを図 2 に示す．このプロトタイプは，HTML，CSS，JavaScript を用いて実装した．プロトタイプは，トピックの階層構造を表示するグラフペイン（図 2-A），記事から抽出したキーワードの一覧を表示するキーワードペイン（図 2-B），ニュース記事を表示する記事ペイン（図 2-C）から構成される．グラフペインの緑色のノード（図 2-D）はニュース記事のトピックを表している．なお，図 2 は「熊本地震」について表示しているプロトタイプシステムで，図 2-A の右側半分は「被害情報とその原因」，左側半分は「支援や復興情報」のトピックになっている．

システムは起動時にキーワードとトピックの階層構造のコンテンツを生成し，それぞれ提示する．ユーザの入力に応じて，表示するニュース記事のコンテンツ

表 3: hLDA の結果の一例 (生成確率上位 20 語)

階層	トピック名	生成単語
2	熊本地震	熊本, 熊本地震, 震度, 阿蘇, 益城, 地震, 避難, 被災, 気象庁, 西原, 熊本大分, 行方不明, 大分, 法壁, 慰霊, 補助, 料金減額, 九州高速, 復興, 訂正熊本
3	被災情報や支援復興情報	南阿蘇, 九州大教授, 犠牲東海大, 可能, 無料, 転出, 支援物資, 熊本大分, 不足深刻, 解析, 被災, 震源 m, 所有, 業者修復, 被災墓地, 同意必要, 場所全国, 前震日, 調査浮上, 春学期
4	学校や施設の復興情報	安全確認, 出席犠牲, 移住学生, ガイダンス南, 熊本キャンパス, 犠牲校舎, 地震学生, 地震店, 突然, 先駆け熊本, 小森地区, 来店客, 児童書店, 道路寸断, 建設完成, 人吉, 役場庁舎, 専門, 義手画家, 美術館再興

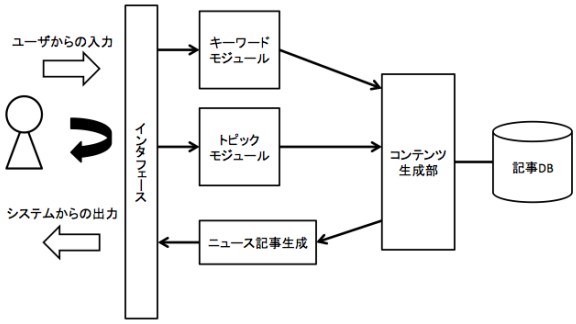

図 1: システム構成図

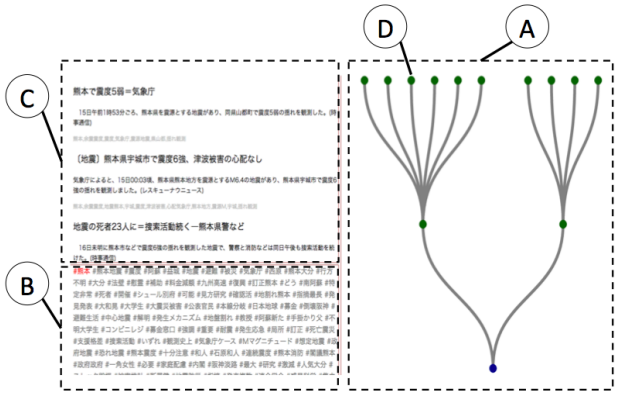

図 2: プロトタイプのスクリーンショット

の生成を行う．コンテンツ生成は (a) トピックに関するニュース記事を生成する方法，(b) キーワードに関するニュース記事を生成する方法のふたつである．

トピックに関するニュース記事を生成する方法は，図 2-A のトピックを表すノードの情報を元に行われる．このノードはマウスオーバーすることによって，トピック名を確認できる．また，トピックのノードをクリックすると，トピックのコンテンツ生成部が入力された情報をもとに，記事のデータベースからそのトピックに属する記事を図 2-C へ出力する．そのクリックと同時に．ノードがハイライトされ，テーマ全体から見たトピックの位置を強調する．

キーワードに関するニュース記事を生成する方法は，

表 4: テーマごとのトピック数

	熊本地震	東京五輪	三菱自動車
階層 2	1	1	1
階層 3	2	2	3
階層 4	10	13	10

図 2-B のキーワードの情報を元に行われる．このキーワードをクリックすると，キーワードのコンテンツ生成部が入力された情報をもとに，記事のデータベースからキーワードに属する記事を検索し，図 2-C へ出力する．そのクリックと同時にキーワードがハイライトされ，クリックしたキーワードが強調される．また，クリックしたキーワードが属するトピックのノードがハイライトされ，キーワードの違いによってトピックの関係性が明確になる．

6 実験

提案システムを用いて，トピックの階層構造を表示することでテーマに対する理解に及ぼす影響を調べるために，提案システムを用いた情報整理と，紙媒体を用いた情報整理の比較を行う実験を行った．

6.1 実験手続き

本実験は，提案システムおよび紙媒体の情報を実験協力者に与え，テーマ中に存在するトピックをまとめてもらう形式で実施した．実験協力者は関西大学総合情報学部に在籍する学生 12 名（男性 5 名，女性 7 名）であった．用いたニュース記事のテーマは，「熊本地震」「東京五輪問題」「三菱自動車の燃費偽装」の 3 種類である．提案システムではこれらのトピックを階層構造により可視化したものを，紙媒体では時系列順にニュース記事を並べたものを実験協力者に提示した．なお実験の際には，提案システムと紙媒体との試行順，テーマの種類はランダムとした．

実験では，実験協力者に対して (1) 実験の目的，(2) 実験の流れ，(3) 実験課題，について説明したのち，提案システムの操作説明を行った．実験は，提案システ

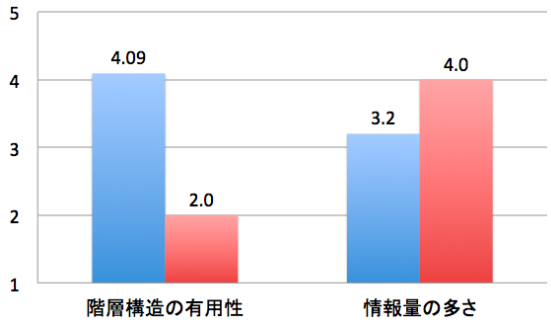

図 3: 5 段階評価の平均点

ムを用いてテーマ中に存在するトピックをまとめる形式で行わせた．

与えた情報を元にテーマ中に存在するトピックをまとめてもらう際には，時間制限は設けず，実験協力者が「トピックをまとめ終えた」と自ら判断した時点で終了とした．実験終了後,「質問項目 1: トピックの整理における提案システムと紙媒体の比較について」「質問項目 2: 提案システムの階層構造がニュースの理解に役に立ったか」「質問項目 3:提案システムの情報量は適切か」の 3 つの質問項目からなるアンケートに答えてもらった．

6.2 提案システムと紙媒体の比較

質問項目 1 では，提案システムと紙媒体のどちらがテーマ中に存在するトピックをまとめやすかったか選択してもらった後，提案システムと紙媒体を比べてそれぞれの利点を記入してもらった（表 5 参照）．

まとめやすさについては，12 名の実験協力者のうち 10 名が提案システムを，2 名が紙媒体を各々選好した．提案システムを選好した実験協力者からは,「ニュース記事をキーワードから絞る方法とトピックから絞る方法の 2 つの方法のうち，どちらかに力を入れてニュースを探すことができたから」「クリック 1 つで記事を絞ることができ，紙媒体と比べて時間の短縮になると感じたから」などの意見が得られた．一方，紙媒体を選好した実験協力者からは,「同時に大量の情報を処理しやすいから」,「時系列順で探しやすかったから」という意見が得られた．

提案システムの利点については,「トピックから記事を絞ることで，期間が離れていても関連するものが見られることができた」,「階層構造を用いることで，トピックのつながりを意識することができた」など，階層構造で提示されるトピックに関する点が指摘された．また，階層構造全体に着目した意見として,「ニュース記事を絞る際に，トピックの名称を見ることで知りたい情報以外でも，そのテーマについての大まかな問題について知ることができた」,「いろんな角度から目的の事柄を見ることができるので，理解を深められるのでは」などが得られた．一方，紙媒体の利点については「ニュース記事の前後を比べて推移していく様子や，記事内容が元々は関係のない分野へと広がっていくのを感じることができて社会への影響が大きくなっていく様子を捉えることができた」「最新の情報を見るだけで重要なことがわかる」など，時系列性に関わる意見が多かった．

6.3 階層構造によるニュースの理解度向上

質問項目 2 では，提案システムの階層構造がニュースの理解に役に立つのかを調べるために，階層構造の有効性について 5 段階（1: 有効でない — 5:有効である）で評価してもらい，その詳細について自由記述形式で回答を収集した．その結果，12 名による評価の合計平均は 3.8 であった．高い評価を示した実験協力者からは,「キーワードをクリックすることで，どの階層のどの位置にある記事を見ているのかを理解できた」「トピック同士が繋がっていたので，脱線が防げる」という意見が得られた．一方で低い評価を示した実験協力者からは,「階層構造から調べたい情報について必要な情報が得られなかった」という意見が得られた．

6.4 提示される情報量の多寡

質問項目 3 では，提案システムの情報量は適切かについて．5 段階（1: 多い — 5:少ない）で評価してもらい，その理由を答えてもらった．12 名による評価の合計平均は 3.3 であった．情報量が多いと判断した実験協力者からは,「キーワードの量が多くて探すのに苦労した」という意見が得られた．一方で情報量が少ないと判断した実験協力者からは,「キーワードをクリックした際に表示される記事の量が少なかった」という意見も得られた．

7 考察と議論

7.1 アンケート結果から得られた知見

6.2 節の質問項目 1 の選好媒体毎に実験参加者を分けた場合の，質問項目 2 と 3 の各々の平均を図 3 に示す．図 3 中の青色のデータは「提案システムの方が整理しやすかった」と回答した実験協力者の 5 段階平均，赤色のデータは「紙媒体の方が整理しやすかった」と回答した実験協力者の 5 段階平均を各々示している．

図 3 の「階層構造の有効性」から，紙媒体に比べて提案システムの方が使いやすかったと考えている人は

表 5: 情報の得やすさ

被験者	選好	選んだ理由
A	紙媒体	同時に大量の情報を処理しやすい．読みやすかった．タイトルが分かりやすかった．
B	紙媒体	時系列順に書いてあったから探しやすかった．
C	システム	レポートに関連するキーワードを見つけやすいから．
D	システム	すばやく情報収集ができる．
E	システム	キーワードをクリックすることでそれに関する記事が見れるため．
F	システム	ただの時系列ではなく，調べたいトピックが必要な情報だけ見れたから．
G	システム	調べたい内容に関するトピックから記事を見つけられるから．
H	システム	誰を選んでも必要な記事をよんでいくことができたから．冊子の方でも関係ありそうな言葉を探して読んでいたので，その作業が減らせた．
I	システム	いらない情報は見なくてすむ．
J	システム	紙媒体の扱いにくさ，クリック一つで記事が表示されるため，時間の短縮になる．
K	システム	ほしい情報がピンポイントでわかる．
L	システム	自分が知りたいトピック，キーワードを重点的に検索できる．

「提示される階層構造が有効である」と判断していることが確認された（$W(2,10) = 1.0, p < 0.05$）．この結果から，探したい情報が含まれるトピックだけでなく，トピックを広げて大まかな情報を確かめたり，トピックを絞って細かな情報を得たりすることで，トピックの関係性を明確にできたと言える．

一方，図 3 の「情報量の多さ」から，提示される情報量の多寡については実験協力者の媒体の選好にかかわらず，有意差は見られなかった（$W(2,10) = 6.0, n.s.$）．これは 4 章で述べたように，収集した記事の文章量は統一性を持たないことから，TF-IDF 法を用いてキーワードを抽出する際，各記事に対して抽出するキーワードの量を制限していないことが原因として挙げられる．

7.2 実験協力者の観察から得られた知見

実験協力者の操作方法を観察したところ，ニュース記事をキーワードから絞る方法と，トピックから絞る方法のふたつを交互を使い分けていた実験協力者がいる一方，片方の機能のみ使用する実験協力者がおり，機能の習熟度にばらつきが見られた．また，実験協力者の中にはシステムの使用方法や機能を忘れてしまう実験協力者もいた．これは，提案インタフェースが実験協力者にとって必ずしも操作しやすいものではなかった可能性を示唆している．この問題を解決するために，これらふたつの機能をインタフェース上で関連づけて提示するなどの改善が必要だと考えている．

8 おわりに

本研究の目的は，専門知識を持たない人を対象に，あるテーマに関連している記事同士の関係性を把握しやすくすることで，テーマに対する理解を深めさせることである．ニュースの理解には，そのテーマに関する事象とその前後関係を把握することが必要であり，Web ニュースサイトでは困難である．そこで本稿では，ニュース記事中のトピックの関係性を明確にするような可視化インタフェースの提案をして検証を行った．

謝辞

本研究の遂行にあたり，文部科学省科学研究費（課題番号: 15H02780）の助成を受けた．記して謝意を表す．

参考文献

[1] Blei, D. M., Griffiths, T. L., Jordan, M. I. and Tenenbaum, J. B.: Hierarchical Topic Models and the Nested Chinese Restaurant Process, *Advances in Neural Information Processing Systems*, Vol. 16, pp. 106–114 (2003).

[2] Serizawa, M. and Kobayashi, I.: Topic Tracking Based on Identifying Proper Number of the Latent Topics in Documents, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 16, No. 5, pp. 611–618 (2012).

[3] Yamashita, R., Okamoto, K. and Matsushita, M.: Exploratory Search System Based on Comic Content Information Using a Hierarchical Topic Classification, *Asian Conference on Information Systems* (2016).

[4] 菊池匡晃, 岡本昌之, 山崎智弘: 階層的クラスタリングを用いた時系列テキスト集合からの話題推移抽出, 第 6 回日本データベース学会年次大会 (2008).

[5] 高間康史, 山田隆志: タグ付きコーパスを用いた地震記事からの動向情報抽出可視化システム, 知能と情報, Vol. 18, No. 5, pp. 711–720 (2006).

[6] 薦田和弘, 大澤幸生: 複数文書の相対的特徴可視化による理解支援, 第 5 回インタラクティブ 情報アクセスと可視化マイニング研究会, pp. 41–46 (2013).