

分散表現を用いた有害表現判別に基づく炎上予測

Flame prediction based on harmful expression judgement using distributed representation

三宅 剛史^{1*} 松本 和幸¹ 吉田 稔¹ 北 研二¹

Takeshi Miyake¹, Kazuyuki Matsumoto¹, Minoru Yoshida¹, Kenji Kita¹

¹ 徳島大学工学部知能情報工学科

¹ Department of Information Science and Intelligent Systems, Tokushima University

Abstract: In recent years, flaming on social media has been a problem. To avoid flaming, it is useful for the system to automatically check the sentences whether they include the expressions that are likely to trigger flaming or not before posting messages. In this research, we target two harmful expressions. There are insulting expressions and the expressions that are likely to cause a quarrel. Firstly, we constructed a harmful expression dictionary. Because a large cost requires to collect the expressions manually, we constructed the dictionary semi-automatically by using word distributed representations. The proposed method used distributed representations of the harmful expressions and general expressions as features, and constructed a classifier of harmful/general based on those features. An evaluation experiment found that the proposed method could extract harmful expressions with accuracy of approx. 70%. On the other hand, it was found that the proposed method could also extract unknown expressions, however, it wrongly extracted non-harmful expressions at a rate of approx. 40%.

1 はじめに

インターネットが発達したことにより SNS(Social Networking Service) やブログといったソーシャルメディアが広く利用されている。これらのサービスを利用することで、個人や企業、団体など誰でも自分の意見や広告などの情報を手軽に発信できるようになった。しかしその反面、発信した情報に対して多くの批判が寄せられるいわゆる「炎上」という現象が問題となっている。

ひとたび炎上が起こるとそのリスクは深刻である。個人情報の流出やそれによるメールや電話を使った問い合わせや嫌がらせ、人間関係の崩壊や全く関係のない人に影響が及ぶ場合もあり、炎上を起こしてしまった個人やその個人の所属する団体は社会的な信用を失ってしまうことになりかねない。こうしたリスクを回避するためにも、炎上を予防することが重要である。

炎上を予防するためには炎上についての理解が必要である。炎上にはさまざまな原因がある。よくあるケースとして、過激であったり、悪印象を与える内容でも SNS に対する理解が浅いため、安易に投稿してしまうことがある。例えばアルバイト先でのふざけた行動に

ついて写真付きで投稿したり、法律に触れるようなことを自慢（犯罪自慢）するような投稿である。これらの炎上を起こしてしまうユーザの多くが、なぜ自分の投稿が他の人に批判されたのか、どうして炎上してしまったのか理解できていないことが多い。そのため、第三者の視点で投稿内容を確認し、内容の指摘と訂正を促すことで炎上を未然に防ぐことができるのではないかと考えられる。

炎上事例を収集し、機械学習することにより炎上判別するという方法も考えられるが、炎上した発言そのものがすぐに削除されたりするため、データ収集が困難という問題がある。そのため、本研究では、収集しやすい悪口表現などを炎上予測の手がかりとする。

ある表現に対して炎上の可能性があるか否かの判断は難しいため、表現の単純な有害さだけで炎上を予測できるわけではない。しかし、表現が「有害である」と判断できるのであれば、その表現が原因で炎上の可能性があることを注意喚起することにより、炎上を予防できると考える。本研究では、有害表現を収集し、その分散表現を学習させ、機械学習により有害表現か否かを判別する分類器を作成することで炎上予防のためのシステムを構築することを目的とする。

*連絡先：徳島大学工学部知能情報工学科
〒 770-8506 徳島市南常三町 2 丁目 1 番地

2 関連研究

近年、炎上を工学分野において扱った研究は多く、様々なアプローチがある。山本ら [1] による、分散表現と連想情報を用いた道徳判断システムや、岩崎ら [2] による CGM における炎上の同定とその応用等がある。

本研究では、有害表現辞書の作成において、大量の有害表現を最初から人手により収集することはコスト面から不可能であると考えたため、人手により小規模な辞書を作成してから拡張する手法等を参考にする。

石坂ら [3] は電子掲示板「2ちゃんねる」を対象として悪口表現の抽出をおこなっている。この研究では、掲示板上の悪口表現を自動抽出し辞書を作成することを目的としている。2ちゃんねるから人手で表現を収集した小規模な悪口辞書を構築し、N-gram を用いた悪口表現の周辺単語モデルを作成し、このモデルを用いてさらに悪口表現を抽出することで辞書の拡張をおこなった。この手法では、新しい表現の抽出が可能であるが、悪口が書き込まれた掲示板の大規模なコーパスを用いなければ大幅な辞書の拡張は難しい。

また、既存研究では悪口表現の抽出をおこなっているが、本研究では差別表現や炎上しそうな単語も対象とする。差別表現は管理されるべき表現であり、また、炎上しそうな単語を抑制することは、炎上予防につながる。また、本研究では word2vec[4] による単語の分散意味表現を特徴として用いることで、コーパスさえ拡充すれば、類似した新しい有害表現の抽出が容易になる。

3 提案手法

3.1 提案手法の概要

図1に有害表現辞書の構築の流れを示す。まず Twitter やブログといったソーシャルメディアや書籍などのメディアから有害表現を抽出し小規模な有害表現辞書を作成する。つぎに、Twitter からランダムに収集したテキストコーパスを形態素解析器 MeCab[5] により分かち書きし、word2vec により分散表現を学習させる。有害表現辞書中の語の分散表現と類似する分散表現を持つ単語をコーパスから抽出し、有害表現辞書の自動拡張をおこなう。自動拡張する際に、意味的に類似しない語がノイズとして含まれる場合があるため、人手による選別をおこない、辞書を精練していく。

図2に分散表現に基づく有害表現判定モデル構築の流れを示す。炎上判定のための分類器を作成するために、有害表現と対を成す表現として、一般的な表現を単語辞書等から収集し登録した一般表現辞書を作成する。有害表現辞書と一般表現辞書における単語の分散表現を素性とし、有害/一般クラスに分類するモデル（有害

表現判定モデル）を Support Vector Machine(SVM) に学習させる。

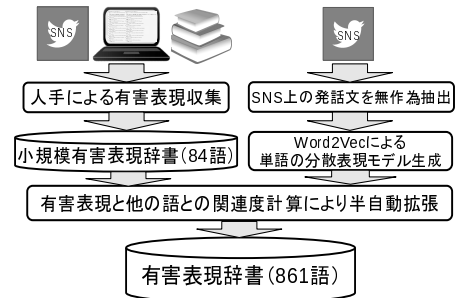


図 1: 有害表現辞書の構築

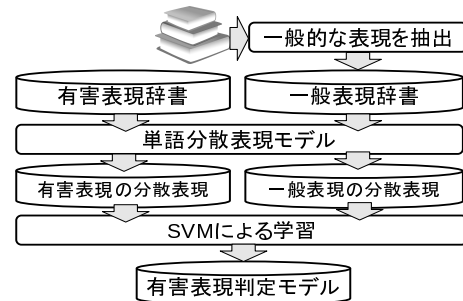


図 2: 分散表現に基づく有害表現判定モデルの構築

3.2 有害表現の抽出

3.2.1 有害表現の定義

本節では本研究での有害表現の定義について述べる。おおまかに2つの表現を対象とした。一つ目は悪口表現、差別表現等の人を罵ったり馬鹿にする際に使われる侮蔑表現、二つ目は政治や歴史問題、宗教等に関連するような意見が衝突しやすいデリケートな話題やタブーが存在するカテゴリに属する表現（以下、荒れそうな表現とする）を対象とした。表1に有害表現の一例を示す。

表 1: 有害表現の一例			
表現名称	表現例		
侮蔑表現	バカ	腰抜け	とろい
	痴呆	ロンパリ	火病
荒れそうな表現	政治	宗教団体	戦争

侮蔑表現は誰もが使いがちであるが、炎上の可能性をもった表現の1つでありなるべく管理されるべき表

現である．悪口表現は誰に対して発言してもあまり好ましくない表現とし、差別表現は特定の属性を持った人たちに対して発言してはいけない表現と定義する．差別表現には、元々は差別表現ではなかった病名や症状等を表す表現を、他者を罵るために使い始めたことにより、新たな語義や用法が追加されたものがある．そのため、本来の意味で使用しているのか差別表現として使用しているのかを判別することは、単語単位だと困難であり、文脈等を考慮した語義曖昧性解消が必要になる．本研究で構築するシステムでは、疑わしい表現であれば、有害でない用法であっても炎上の可能性がある表現として提示する．

荒れそうな表現は、炎上しやすいカテゴリの表現を判定したいため本研究で扱う対象とした．意見の衝突しやすい話題では頻繁に口論が起き、そこから団体や企業、個人などを批判、誹謗中傷したりとエスカレートして状況が悪くなり、炎上に発展するケースが多くみられる．そのため、炎上しやすい話題に関する表現も有害表現として扱うことで、炎上を抑制できると考えた．

3.2.2 有害表現の抽出について

抽出における手順を述べる．大量の表現を人手により集めることが困難なため、人手で少量の表現を集めてそれを自動で拡張することを目指す．まず、Twitter、ニュースサイト、ブログ記事、掲示板、書籍等のメディアを対象として人手で有害表現を 84 単語収集し、小規模な有害表現辞書を作成した．収集する基準は 3.2.1 節において説明した有害表現の定義に従った．最近はまだ使われない古い表現も収集対象とした．

小規模な有害表現辞書と類似度の高い表現を収集するために、word2vec の学習をおこなう．まず Twitter から無作為に約 100 万ツイートを収集し、形態素解析により分かち書きしたものをを用いて word2vec の学習をおこなった．word2vec の学習によりツイート内の単語をベクトル化することができ、ベクトルを比較することで関連度を調べることができる．word2vec の学習をおこなう際のパラメータは次元数を 200 とし、文脈の最大単語数である window 幅は 5、単語を無視する頻度である sample 値は 0.001 と初期設定のまま使用した．図 3 に、代表的な有害表現について分散表現ベクトルを t-SNE[6] を用いて次元圧縮し、二次元座標空間上にプロットしたグラフの一部を拡大したものを示す．

小規模な有害表現辞書の単語と学習した word2vec のモデルのベクトルの関連度を比較し、類似した表現をコーパスから自動抽出した．しかし、自動抽出した場合に、有害表現と意味は類似しているが有害ではない表現も収集してしまうため、抽出された語群に対し、人手により判断することで、有害な表現の選定をおこなっ

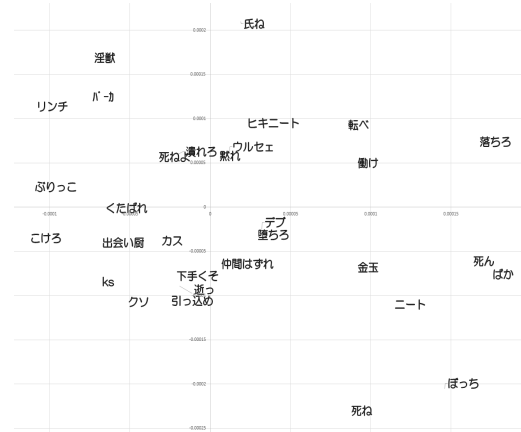


図 3: 有害表現の二次元座標空間へのマッピング

た．3000 単語ほど収集したものから 861 単語を有害な表現とした．この 861 単語を登録した辞書を有害表現辞書とする．

3.3 一般的な表現の抽出

3.3.1 一般的な表現の定義

一般的な表現とは、大衆に使われる有害でない表現とする．この表現の定義として「日本語の語彙特性 第 9 巻 [7]」を参照した．この書籍は日本語の単語に対して単語親密度という数値を表記しており、これは単語にどれだけなじみがあるかを 7 段階尺度 (1:なじみがない, 7:なじみがある) で定義したものである．表 3 に単語親密度の例を示す．

表 2: 単語親密度の例

名称	単語親密度
星空	6.176
アーケード	5.688
偏光	3.824
あおざり	1.750

なじみがあるということは大衆に知られているということである．本研究では、親密度が高い単語かつ有害でない表現を一般的な表現とした．

本研究では有害表現を SVM に学習させるため、負例として反対の意味を持った有害でない言葉が必要である．そのため一般的な表現を収集して辞書を作成する．

3.3.2 一般的な表現の抽出について

一般表現の抽出について述べる．3.2.1 節で述べたように書籍をもとに表現を抽出する．日本語の単語に対

して1から7までの7段階尺度で親密度が表記されており、親密度が高いほどなじみがあるといえる。本研究では親密度が6.3以上の表現である1384語を抽出し、word2vecの学習において作成したモデルに含まれている単語で有害な表現ではないものを選定し、そこから無作為に900語を選出した。この900語を登録したものを一般表現辞書とする。表3に抽出した表現の例を示す。

表 3: 一般的な表現の例

名称	単語親密度
アイスクリーム	6.562
正直	6.375
うるさい	6.531
会う	6.594

3.4 炎上判定システムの作成

3.4.1 分類器の作成

3.2節と3.3節で収集した有害表現辞書と一般表現辞書を用いてSVMによる学習をおこなう。SVMによる学習をおこなうためには素性ベクトルを学習させる必要がある。学習用データとして有害表現864語、一般表現900語を用いる。これらをword2vecの分散表現モデルを参照し、分散表現に変換したものを素性ベクトルとして用い、SVMによる二値（有害/一般）分類器を学習し、有害表現判定モデルを作成した。SVMの学習にはデフォルトパラメータを使用し、有害表現を正例、一般表現を負例とした。

3.4.2 有害表現判定モデルの検証

構築した有害表現判定モデルの有効性を確認するため、10分割交差検証をおこなった。交差検証の結果（精度、 $\pm 2\sigma$ レンジ）を表4に示す。 $\pm 2\sigma$ は標準偏差を2倍したもので結果のばらつきを表している。このことから分類器の精度は高く、かつ結果のばらつきは極めて少ないことがわかる。

表 4: 10 分割交差検証の結果

平均精度	$\pm 2\sigma$ レンジ
0.96	(+/- 0.03)

3.4.3 有害表現判定システムの作成

作成した分類器を用いて炎上の可能性を判定するシステム（有害表現判定システム）を作成する。入力された文章に対して形態素解析器により単語単位に分割し、各単語に対してword2vecで学習した分散表現のモデルを参照する。モデル中に単語が存在する場合、分散表現に変換して分類器に入力することで、炎上の可能性がある表現が含まれているかを判定して表示する。システムの画面例を図4に示す。

Input Sentence: 'ブロック解除しろ事故って死ね'

ブロック解除しろ事故って死ね

送信

```
[解除]: [0.935]
[事故]: [0.982]
[しろ]: [0.931]
[ブロック]: [0.897]
[死ね]: [-0.757940]
[って]: [0.970]
```

Harmfulness Score: 0.758

Word: <死ね>

は有害な表現だと考えられます。炎上の可能性があるので他の単語の使用を検討してください。

図 4: 有害表現判定システムの Web インタフェース

このシステムは文章中に含まれる単語を有害かどうか判定し、どの単語が炎上の可能性があるかを表示する。表現が有害なら炎上するとは限らないが、炎上の要因として暴言や悪口などの有害な表現は存在する。そのため訂正を促し炎上の可能性を抑制するねらいがある。

4 評価実験

4.1 実験データ

本研究で使用する実験データは、実際にインターネット上で問題となっている悪口文を用いる。サンリオのキャラクターである「シナモン」の公式ツイート [8] に対し、多数の悪口文が投稿されている。いじり感覚の悪ふざけから発言が過激化しネット上でのいじめに発展している。いじめの対象が一個人ではない実在しないキャラクタということもあり、投稿が削除されたり、アカウントが削除される事態には至っていない。この

多数の悪口文を実験対象として用いることで炎上予測性能を評価できると考えた。

実験データについて表 5 に示す。悪口文は、2015 年に投稿されたツイートから計 108 文を収集した。一部、悪口文とは言えない文もあるが、誤検出の確認のために残している。表 6 に悪口文を抜粋したものを示す。これらの悪口文を手により判断し、システムが有害と判定すべき 35 単語の表現を正解表現とした。正解表現の一覧を表 7 に示す。

表 5: 実験データ

データ名	数量
シナモンへの悪口文	108 文
正解表現	35 単語

表 6: シナモンへの悪口文の例

悪口文	正解表現
ブロック解除しろ事故って死ぬ 雲から落ちろ ええかげんぶりっこ卒業して	死ぬ 落ちろ ぶりっこ

表 7: 正解表現

正解表現一覧			
死ぬ	転べ	落ちろ	死ん
仲間はずれ	ぼっち	金玉	働け
堕ちろ	ks	黙れ	淫獣
バ ーカ	ばか	ぶりっこ	やろう
消せ	転ん	ヘイト	出会い厨
こけろ	氏ね	ウルセエ	下手くそ
引っ込め	クソ	ニート	カス
くたばれ	潰れろ	リンチ	デブ
ヒキニート	逝っ	死ねよ	

4.2 評価方法

悪口文からシステムに判定してほしい有害表現を正解表現として選定する。この選定基準は有害表現辞書を作成する際の基準と同じである。正解表現を 35 単語選定し、これらをシステムが有害と判定できればシステムは正常に判定できたとする。

正解表現が判定できているかの再現率および適合率と、誤って有害でない表現が判定されてしまった誤検出率などを比較する。また、SVM を用いずに辞書のみを用いたマッチングで有害表現がどれだけ判定できるかを実験し、SVM の有用性についても調査する。

4.3 実験結果

実験結果を表 8 に示す。

表 8: 実験結果

	提案手法	辞書のみ
判定成功	26	10
誤検出	20	0
再現率	74.3	28.6
適合率	56.6	100

この表から、SVM を用いることで有害表現辞書に存在しない未知の有害表現の判定が可能ながわかる。また、当然ながら、正解表現のうち有害表現辞書に含まれる表現については、すべて有害表現として判定できた。

実験において判定できた表現等について表 9 に示す。誤検出をした表現について表 10 に示す。

表 9: 判定に成功した表現と失敗した表現の内訳
表現一覧

成功表現		失敗表現	辞書
死ぬ	ばか	死ん	ばか
転べ	ヘイト	金玉	ヘイト
落ちろ	ぼっち	働け	ぼっち
仲間はずれ	クソ	やろう	クソ
バ ーカ	ニート	消せ	ニート
堕ちろ	カス	転ん	カス
黙れ	デブ	下手くそ	デブ
淫獣	氏ね	引っ込め	氏ね
こけろ	くたばれ	逝っ	くたばれ
潰れろ	ヒキニート		ヒキニート
ウルセエ	ぶりっこ		
ks	出会い厨		
死ねよ	リンチ		
26 単語		9 単語	10 単語

表 10: 誤検出された表現

「w」の羅列・記号	怪我・病気	命令語	その他
w	植物状態	どけよ	危
www	粉碎骨折	働けよ	非常識
wwwwww	永眠	蹴れ	
w*25	髌骨	浮け	
! *9		染めろ	

この表を見ると「w」の羅列を有害と判定しているパターンが多い。病気や怪我を表す語も含まれていることがわかる。また、命令語が含まれている。

5 考察

SVMによる正解表現の判定が26単語と、有害表現辞書による単純なマッチングよりも多く判定できている。この結果から、有害表現の判定において、分散表現を素性としたSVMによる分類が有効であることが分かる。学習させた有害表現に近い意味の単語を有害表現として判定できたと考えられる。一般的な表現20単語が有害表現と判定された。表10から、「w」の羅列が多いことが分かる。分散表現の学習段階において、「w」が有害と考えられる文章に頻出していたことが原因と考えられる。「w」はネットスラングであり、人を嘲笑ったり、小馬鹿にする際によく用いられる。

「!!!!!!!」も同じようなケースで、命令文や罵倒と一緒に用いられることが多いためだと考えられる。また、侮蔑表現に病気を基にした表現が多いことから、侮蔑表現と、怪我や病気を表す表現の分散表現が類似してしまい、有害表現として判定されてしまったと考えられる。また、同様に、命令語が、罵倒や悪口などと分散表現間の類似度が高くなる傾向にあったため、システムが有害と判定してしまった。

6 まとめ

本研究では、ソーシャルメディアや書籍等から有害表現を収集し有害表現辞書を作成した。本研究での有害表現とは悪口表現、差別表現、荒れそうな表現のことである。Twitterから100万ツイートを取得し、word2vecによる学習をおこない有害表現と類似する語をコーパスから抽出することで、有害表現辞書の拡張をおこなった。

有害表現と対を成す表現として一般的かつ有害でない表現を収集し、一般表現辞書を作成した。有害表現辞書と一般的表現辞書を学習データとして、SVMによる学習をおこない、有害表現を判定する有害表現判定モデルを作成した。このモデルを用いて、有害表現判定システムを作成し、悪口文を用いた実験をおこなった。結果として、辞書のみを用いた手法よりも多くの有害表現を抽出可能なことが分かった。誤って判定されてしまった表現について分析した結果、類似語の影響や用法に依存して有害とも解釈できてしまう語が存在することが分かった。今後の課題として、有害表現辞書の拡充と、用言や複数の語に分割されてしまう表現、文脈によって有害と判断される語への対応が考えられる。

謝辞

本研究の一部は、科学研究費補助金15K16077, 15K00425, 15K00309によりおこなわれた。

参考文献

- [1] 山本 真大, 萩原 将文: 分散表現と連想情報を用いた道徳判断システム, 日本感性工学会論文誌, Vol.15, No.4, pp. 493-501, 2016.
- [2] 岩崎 祐貴, 折原 良平, 清 雄一, 中川 博之, 田原 康之, 大須賀 昭彦: CGMにおける炎上の同定とその応用, The 27th Annual Conference of the Japanese Society for Artificial Intelligence, 2013.
- [3] 石坂 達也, 山本 和英: 2ちゃんねるを対象とした悪口表現の抽出, 言語処理学会第16回年次大会, pp.178-181, 2010-03.
- [4] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, CoRR, abs/1310.4546, 2013.
- [5] MeCab: Yet Another Part-of-Speech and Morphological Analyzer <http://taku910.github.io/mecab/>.
- [6] Maaten, L., Hinton, G.: Visualizing Data using t-SNE, Journal of Machine Learning Research, Vol.9, pp. 2579-2605, 2008.
- [7] 天野 成昭, 近藤 公久, 笠原 要, NTTコミュニケーション科学基礎研究所(監修), NTTデータベースシリーズ 日本語の語彙特性〈第9巻〉単語親密度 増補, 三省堂, 2008.
- [8] シナモン【公式】Twitter アカウント: https://twitter.com/cinnamon_sanrio?lang=ja

議論フレームワークにおいて受け入れる主張の決定方法とそれらの方法の間の関係について

Comparative Study of Methods to Determine Arguments Acceptance in Argumentation Framework

豊東 柊哉^{1*} 山口 和紀¹

Shuya BUNDO¹ Kazunori YAMAGUCHI¹

¹ 東京大学

¹ The University of Tokyo

Abstract: 議論における主張間の関係を表す議論フレームワーク (AF) において, 受け入れる主張を決定する方法として, Dung の意味論に基づく方法, 主張ごとに定めたコストに基づく方法, 主張ごとに定めた優先度に基づく方法などが提案されている. 本研究では, それらの方法の間の関係を検討し, お互いに矛盾するものだけでなく, 両立するものがあることを示す.

1 序論

議論フレームワークは, 議論を表現するためのフレームワークであり, とくに主張の説得力を評価する目的で用いられている. このために用いられる semantics (意味論) として, “説得力のある主張の集合” を定義するものである Dung の semantics があるが, この semantics では主張の受け入れやすさを評価する基準が「受け入れられる」「受け入れられない」の二値しかないため, 巨大な議論を扱う場合などには実用的でないとされている. この問題を解決するために, 多数のレベルの受け入れやすさを考えることができる semantics として, ranking semantics と呼ばれるものはいくつか提案されている. しかしながら, Dung の semantics と ranking semantics との関係についてはこれまで十分に研究されていなかった. 本論文では, Dung の semantics と ranking semantics との関係について調べ, それらが矛盾する場合だけではなく, 両立する場合もあることを示す.

2 Dung の semantics

本節では Dung の semantics を説明する.

定義 2.1 (議論フレームワーク [2]). 有限集合 A および A 上の二項関係 $\rightarrow \subseteq A \times A$ の組 $F = \langle A, \rightarrow \rangle$ を議論フレームワーク (Argumentation Framework, AF) という. $S, T \subseteq A$ に対して, ある $s \in S, t \in T$ があって

$s \rightarrow t$ であることを $S \rightarrow T$ で表す. $S = \{s\}$ のときは $s \rightarrow T, T = \{t\}$ のときは $S \rightarrow t$ のようにも書く.

定義 2.2. $F = \langle A, \rightarrow \rangle$ を AF, $S \subseteq A$ とする. $a \rightarrow b$ なる $a, b \in S$ が存在しないとき S は *conflict-free* であるという. $\text{Def}(S) = \{a \in A \mid (\forall b \rightarrow a, \exists c \in S, c \rightarrow b)\}$ と定める. S が *conflict-free* かつ $S \subseteq \text{Def}(S)$ のとき S は *admissible extension* であるという. S が *admissible extension* かつ $S = \text{Def}(S)$ であるとき S は *complete extension* であるという. F の *admissible extension* 全体, *complete extension* 全体をそれぞれ $\text{adm}(F), \text{comp}(F)$ で表す. 列 $\emptyset, \text{Def}(\emptyset), \text{Def}^2(\emptyset), \dots, \text{Def}^n(\emptyset), \dots$ は増大列であって, 十分大きな N に対し $\text{Def}^N(\emptyset) = \text{Def}^{N+1}(\emptyset)$ を満たす. この $\text{Def}^N(\emptyset)$ を *grounded extension* という. $\text{Atk}(S) = \{b \mid S \rightarrow b\}$ と定める.

3 ranking semantics の定義

本節では ranking semantics を説明する.

定義 3.1. 反射的・推移的二項関係を preorder という. $\succsim$ を A 上の preorder とし, 次のように定める:

- $a \simeq b \Leftrightarrow a \succsim b$ かつ $b \succsim a$
- $a \succ b \Leftrightarrow a \succsim b$ だが $b \not\succsim a$ ではない

定義 3.2 (ranking semantics[3]). AF F が与えられたとき $A(F)$ 上の preorder (反射的・推移的二項関係, 擬順序) $a \succsim_F b$ を返すものを *ranking semantics* という.

*連絡先: 東京大学総合文化研究科広域科学専攻
E-mail: bundo@graco.c.u-tokyo.ac.jp

4 ranking semantics の公理

ranking semantics の性質をみるために公理¹がいくつか提案されている。ここではそれらの公理を紹介するために、まずいくつか定義を導入する。実例は後半で示す。

定義 4.1. F, G : AF に対し $F \cup G := \langle A(F) \cup A(G), \rightarrow_F \cup \rightarrow_G \rangle$ と定める。 $a, b \in A(F)$ に対し、 $P_{ab} = \{b = a_0 \rightarrow a_1 \rightarrow \dots \rightarrow a_n = a \mid a_i \in A\}$ と定める。 $l = a_0 \rightarrow \dots \rightarrow a_n$ に対し $|l| = n$ と定める。 $R_n(a) = \{b \mid \exists l \in P_{ab}, |l| = n\}$ (多重集合) と定め、 $R^+(a) = \bigcup_{n \in \mathbb{Z}^+} R_{2n}(a)$, $R^-(a) = \bigcup_{n \in \mathbb{Z}^+} R_{2n-1}(a)$ と定め、前者の元を a の *defender*, 後者の元を a の *attacker* とよぶ。 $R_1(a)$ の元を a の *direct attacker* とよび、 $R_2(a) \neq \emptyset$ のとき a は *defend* されているという。

定義 4.2. *defense root* (resp. *attack root*) とは non-attacked defender (resp. attacker) のことをいう。 $BR_n(a) = \{b \in R_n(a) \mid R_1(b) = \emptyset\}$ と定めたとき $BR^+(a) = \bigcup_{n \in \mathbb{Z}^+} BR_{2n}(a)$ の元を *defense branch*, $BR^-(a) = \bigcup_{n \in \mathbb{Z}^+} BR_{2n-1}(a)$ の元を *attack branch* という。

定義 4.3. $F = \langle A, \rightarrow \rangle$ に対し次のような AF を考える: $P_n(a) = \langle \{x_0 = a, x_1, \dots, x_n\}, \{x_n \rightarrow x_{n-1}, x_{n-1} \rightarrow x_{n-2}, \dots, x_1 \rightarrow x_0\} \rangle$. ここで $i \neq 0$ に対し $x_i \notin A$.

定義 4.4. $a \in A(F)$ の defense が *simple* であるとは、任意の a の defender は、 a の 1 つの attacker を攻撃していることをいう。また、 a の defense が *distributed* であるとは、任意の a の direct attacker b に対し、 b を攻撃しているような元は高々 1 つであることをいう。

定義 4.5. $F = \langle A, \rightarrow \rangle$ に対し F と同型な F のコピー $F' = \langle A', \rightarrow' \rangle$ と同型対応を与える写像 $\gamma: A \rightarrow A'$ を考える ($A \cap A' = \emptyset$). このとき $F' = F^\gamma$ と書く。

ranking semantics の妥当性について判断するための公理がいくつかある。本節ではそのような公理のうち、それ自体もまた ranking semantics と見なせる (具体的な方法は定義 5.1 の次の説明を参照されたい) ようなものについて見ていく。

VP $R_1(a) = \emptyset, R_1(b) \neq \emptyset$ ならば $a \succ b$

SC $a \not\succ a, b \rightarrow b$ ならば $a \succ b$

CP $|R_1(a)| < |R_1(b)|$ ならば $a \succ b$

DP $|R_1(a)| = |R_1(b)|, R_2(a) \neq \emptyset, R_2(b) = \emptyset$ ならば $a \succ b$

DDP $|R_1(a)| = |R_1(b)|, |R_2(a)| = |R_2(b)|, a$ の defense は simple で distributed, b の defense は simple だが distributed でないならば $a \succ b$

⊕DB $a \in A(F)$ とする。 $F \cup F^\gamma \cup P_{2n}(\gamma(a))$ において $\gamma(a) \succ a$

+DB $a \in A(F)$ とする。 $R_1(a) \neq \emptyset$ ならば $F \cup F^\gamma \cup P_{2n}(\gamma(a))$ において $\gamma(a) \succ a$

↑DB $a, b \in A(F)$ とする。 $b \in BR^+(a), b \notin BR^-(a)$ ならば $F \cup F^\gamma \cup P_{2n}(\gamma(b))$ において $a \succ \gamma(a)$

↑AB $a, b \in A(F)$ とする。 $b \in BR^-(a), b \notin BR^+(a)$ ならば $F \cup F^\gamma \cup P_{2n}(\gamma(b))$ において $\gamma(a) \succ a$

+AB $a \in A(F)$ とする。 $F \cup F^\gamma \cup P_{2n-1}(\gamma(a))$ において $a \succ \gamma(a)$

AvsFD AF が acyclic で $BR^-(a) = \emptyset, |R_1(b)| = 1, R_2(b) = \emptyset$ ならば $a \succ b$

5 Dung の semantics から誘導される ranking semantics

本節では Dung の semantics から誘導される ranking semantics である Dung ranking を新たに提案する。そのためにまず ranking semantics 間の関係を定義しておく。

定義 5.1. 1. ranking semantics

$\succsim$ が $\succsim'$ より強力であるとは、 $\succsim \supseteq \succsim'$ かつ $\succ \supseteq \succ'$ であることをいう。ここで、包含関係は二項関係を集合としてみたときのものである。

2. 2 つの ranking semantics $\succsim', \succsim''$ に対し、 $\succsim'$ と $\succsim''$ のいずれよりも強力な ranking semantics $\succsim$ があるとき、 $\succsim'$ と $\succsim''$ は両立するという。同様に、 $\lambda \in \Lambda$ で添字づけられた ranking semantics の集合 $\{\succsim_\lambda\}_{\lambda \in \Lambda}$ が与えられたとき、各 $\succsim_\lambda$ いずれよりも強力な ranking semantics があるなら、 $\{\succsim_\lambda\}_{\lambda \in \Lambda}$ は両立するという。

一方が他方より強力である場合は、当然両立するので、2 つの ranking semantics の間の関係は、

1. 一方が他方より強力である

¹与えられた ranking semantics がどの公理を満たすか調べるような使い方をするので、公理というよりも性質という方が適切であるが、先行研究では axiom と呼ばれているので本論文では公理と呼んでおく。

2. 1ではないが, 両立する

3. 両立しない

の3つに分類される.

VPなどの公理と他の ranking semantics は, そのままでは上記の包含関係や両立の関係を考えることができないが, VPなどの公理は「 $\bigcirc\bigcirc$ ならば $a \succ b$ 」という形で記述されているので, $a \succ b \Leftrightarrow \bigcirc\bigcirc$ または $a = b$ と定義することで, 公理を満たす最も弱い ranking semantics そのものと見ることもできる. 以下では, 公理をそのような ranking semantics と同一視し, 公理と他の ranking semantics の比較を行っていく.

各 $\succsim_\lambda$ を含むような最小の preorder は $\bigcup_{\lambda \in \Lambda} \{\succsim_\lambda\}$ の推移閉包である. ここで, $\succsim \supseteq \succsim'$ に対し, $\succsim$ が $\succsim'$ より強力であることは, 各 $a \succ' b$ に対し $b \succsim a$ でないことを意味することに注意すれば, 次の命題がいえる.

命題 5.2. $\{\succsim_\lambda\}_{\lambda \in \Lambda}$ が両立するならば, $\bigcup_{\lambda \in \Lambda} \{\succsim_\lambda\}$ の推移閉包は各 $\succsim_\lambda$ より強力である.

2つの ranking semantics の両立について考えるときには, 特に次が成り立つ.

命題 5.3. $\succsim'$ と $\succsim''$ が両立する $\Leftrightarrow$ 任意の $n \geq 0$ と, 主張の列 $a_0 \succ' b_0 \succ'' a_1 \succ' b_1 \succ'' a_2 \succ' b_2 \succ'' \dots \succ' b_{n-1} \succ'' a_n \succ' b_n \succ'' a_0$ に対し, $b_n \simeq'' a_0 \simeq' b_0$ である.

証明. $\Rightarrow$ $\succsim'$ と $\succsim''$ よりも強力な preorder $\succsim$ をとる. $a_0 \succ b_0 \succ a_1 \succ b_1 \succ \dots \succ a_n \succ b_n \succ a_0$ より, $a_0 \simeq b_0$ であるから, 両立の定義から $a_0 \succ' b_0$ は成立しない. したがって $a_0 \succsim' b_0$ である. 同様に $b_0 \succsim'' a_0$ も示される.

$\Leftarrow$ 対偶を示す. $\succsim'$ と $\succsim''$ が両立しなかったとする. このとき, $\succsim'$ と $\succsim''$ の和集合の推移閉包を $\succsim$ とおくと, $\succ' \not\subseteq \succsim$ または $\succ'' \not\subseteq \succsim$ である. どちらも同様であるから $\succ' \not\subseteq \succsim$ の場合のみ示そう. このとき $a \succ' b$ かつ $a \not\succ b$ でないような a, b がとれる. $a \succ' b$ より $a \succsim b$ であるから, $a \simeq b$ である. このとき, 推移閉包の構成から, 列 $a = a_0 \succ' b = b_0 \succ'' a_1 \succ' b_1 \succ'' \dots \succ'' a_n \succ' b_n \succ'' a_0$ がとれる. $\square$

系 5.4. $\succsim'$ が半順序であるとき, $\succsim'$ と $\succsim''$ が両立する $\Leftrightarrow$ 任意の $n \geq 0$ に対し, 列 $a_0 \succ' b_0 \succ'' a_1 \succ' b_1 \succ'' a_2 \succ' b_2 \succ'' \dots \succ' b_{n-1} \succ'' a_n \succ' b_n \succ'' a_0$ は存在しない.

証明. $\Rightarrow$ 命題 5.3 より明らか.

$\Leftarrow$ 命題 5.3 において, $\succsim'$ が半順序であれば $a \succsim' b \Leftrightarrow a \succ' b$ or $a = b$ であることと, preorder の推移律を適用すればよい. $\square$

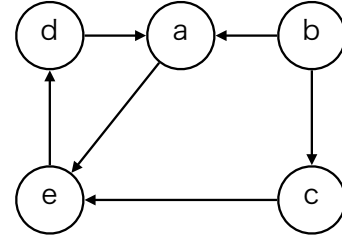


図 1: AF の例 F_1 ([2] の Figure 1)

以下では, admissible semantics や complete semantics から ranking semantics を定義するが, 特定の semantics でなくても成立することがあるので, これらの semantics を総称して semantics と呼ぶことにする. 形式的には, $\sigma: \mathcal{AF} \rightarrow 2^{2^U}$ であって各 $F = \langle A, \rightarrow \rangle \in \mathcal{AF}$ に対し $\sigma(F) \subseteq 2^A$ かつ $\sigma(F) \neq \emptyset$ なるものを semantics とよぶことになる.

σ を *adm* や *comp* などの Dung の semantics とする. 既存の $\sigma(F)$ を利用した semantics としては, 主張 $a \in A(F)$ を uni-accepted, exi-accepted, cleanly-accepted, not-accepted の4つに分類するもの [4] があるが, 以下では, これの分類よりも細かいものを考える.

以下, 本章を通じて $\sigma(F) \subseteq \text{adm}(F)$ を仮定する. このとき $E \in \sigma(F)$ に対し $\{E, \text{Atk}(E), \text{Undec}(E)\}$ は $A(F)$ の分割となる.

定義 5.5. $E \subseteq A(F)$ とする. $a, b \in E$ に対し次のような preorder を考える.

1. $a \succ_F^{E,3} b$ iff $[a] \geq [b]$ ただし $[a]$ は $A(F)$ の分割 $\{E, \text{Atk}(E), \text{Undec}(E)\}$ に付随する同値関係による同値類であり, $\geq$ は, $E > \text{Undec}(E) > \text{Atk}(E)$ と定めた $>$ に等号を加えたものである.
2. $a \succ_F^{E,2} b$ iff $[a] \geq [b]$ ただし $[a]$ は $A(F)$ の分割 $\{E, A(F) - E\}$ に付随する同値関係による同値類であって $E > A(F) - E$ とする.

$v = 2, 3$, $a, b \in A(F)$ に対し $a \succ_F^{\sigma,v} b \Leftrightarrow \forall E \in \sigma(F), a \succ_F^{E,v} b$ と定める. するとこれは preorder になる. この preorder を *Dung ranking* と呼ぶ.

例 5.6. F_1 ([2] の Figure 1, 図 1 参照) を AF の例とする. F_1 では, b が grounded のため, $E = \{b, e\}$ が唯一の complete extension となる. この E に対して, $\{E, \text{Atk}(E), \text{Undec}(E)\}$ は $\{\{b, e\}, \{a, c, d\}, \emptyset\}$, $\{E, E - \text{Atk}(E)\}$ は $\{\{b, e\}, \{a, c, d\}\}$ となり, $b, e \succ_F^{\text{comp},v} a, c, d$ となる.

次に, $\succ_F^{\sigma,v}$ と VP などの公理 (を ranking semantics) としてみたものの関係を調べる.

VP $\sigma = adm$ (admissible) のとき, $\langle \{a, b, c, d\}, \{b \rightarrow c \rightarrow d\} \rangle$ なる AF において a と d が比較不能なので VP は必ずしもいえない。だが $\succsim^{adm,v}$ と $\succsim^{VP}$ は両立する。

証明. 系 5.4 により, $a_0 \succ^{VP} b_0 \succsim^{adm,v} a_1 \succ^{VP} b_1 \succsim^{adm,v} \dots \succ^{VP} b_{n-1} \succsim^{adm,v} a_n \succ^{VP} b_n \succsim^{adm,v} a_0$ として矛盾を導けばよい。このとき $R_1(a_1) = \emptyset$ であるから $E = \{a_1\}$ は admissible であり $b_0 \succsim^{adm,v} a_1$ より $b_0 \succ^{E,v} a_1$ であるから $b_0 \in E$, すなわち $b_0 = a_1$ となる。このとき $R_1(b_0) = \emptyset$ だがこれは $a_0 \succ^{VP} b_0$ に反する。 $\square$

$\sigma = comp$ などの場合は $a \rightarrow b \rightarrow c$ において $a \simeq c$ となってしまうため両立しない。

SC $\sigma(F) \subseteq adm(F)$ なる任意の σ と, $v = 2, 3$ について両立しない。

証明. $c \rightarrow a, b \rightarrow b$ を考えれば $adm(F) = \{\emptyset, \{c\}\}$ である。 $a \not\succeq a, b \rightarrow b, b \succsim_F^{\sigma,v} a$ である。 $\square$

CP $\sigma(F) \subseteq adm(F)$ なる任意の σ と, $v = 2, 3$ について両立しない。

証明. $c_1 \rightarrow d_1, c_2 \rightarrow d_2, d_1 \rightarrow c_2, d_2 \rightarrow c_1, c_1 \rightarrow a, d_1 \rightarrow a, c_1 \rightarrow b, c_2 \rightarrow b, d_1 \rightarrow b$ とすると $adm(F) = \{\emptyset, \{c_1, c_2\}, \{d_1, d_2\}\}$ であり $|R_1(a)| = 2 < |R_1(b)| = 3, b \succsim_F^{\sigma,v} a$ である。 $\square$

DP $\succsim^{\sigma,v}$ が $v = 2$ または, $v = 3$ で $\sigma(F) \subseteq comp(F)$ のとき, 両立する。(なぜなら, $a \succ^{DP} b$ であれば, b は $\succsim^{\sigma,v}$ に関する最小元となるため。)

$\succsim^{adm,3}$ についても $\succsim^{DP}$ と両立する。

証明. 系 5.4 により, 列 $a_0 \succsim^{adm,3} b_0 \succ^{DP} a_1 \succsim^{adm,3} b_1 \succ^{DP} a_2 \succsim^{adm,3} b_2 \succ^{DP} \dots \succsim^{adm,3} a_n \succsim^{adm,3} b_n \succ^{DP} a_{n+1} = a_0$ が存在したとして矛盾を導く。 $x \in R_1(a_i)$ とする。 $R_2(a_i) = \emptyset$ より, $E = \{x\}$ とおくと $E \in adm(F)$ である。このとき $a_i \in Atk(E)$ であり, また $a_i \succsim^{adm,3} b_i$ より $a_i \succ^{E,3} b_i$ であるから $b_i \in Atk(E)$, すなわち $x \in R_1(b_i)$ となる。ゆえに $R_1(a_i) \subseteq R_1(b_i)$ となる。とくに $|R_1(a_i)| \leq |R_1(b_i)|$ である。これと $|R_1(b_i)| = |R_1(a_{i+1})|$ より, $|R_1(a_0)| \leq |R_1(b_0)| = |R_1(b_1)| \leq |R_1(a_2)| = \dots \leq |R_1(b_n)| = |R_1(a_0)|$ となり, 結局 $|R_1(a_0)| = |R_1(b_0)|$ である。 $R_1(a_0) \subseteq R_1(b_0)$ より $R_1(a_0) = R_1(b_0)$ といえる。しかしこれは $R_2(a_0) = \emptyset, R_2(b_0) \neq \emptyset$ であることに反する。 $\square$

DDP $\succsim^{\sigma,v}$ が $v = 2$ または, $v = 3$ で $\sigma(F) \subseteq comp(F)$ のとき, 両立する。(なぜなら, $a \succ^{DDP} b$ であるとする。このとき $|R_1(a)| = |R_1(b)|, |R_2(a)| = |R_2(b)|$ である。 a は distributed であるから, $|R_1(a)| \geq$

$|R_2(a)|$ であり, したがって $|R_1(b)| \geq |R_2(b)|$ である。 b の defense は simple かつ distributed でないから, ある $c \in R_1(b), R_1(c) = \emptyset$ である。したがって, $v = 2$ または, $v = 3$ で $\sigma(F) \subseteq comp(F)$ のとき, $a \succ^{DDP} b$ であれば, b は $\succsim^{\sigma,v}$ に関する最小元となる。) $\succsim^{adm,3}$ についても $\succsim^{DP}$ と両立する。

証明. 系 5.4 により, 列 $a_0 \succsim^{adm,3} b_0 \succ^{DDP} a_1 \succsim^{adm,3} b_1 \succ^{DDP} \dots \succ^{DDP} a_n \succsim^{adm,3} b_n \succ^{DDP} a_{n+1} = a_0$ が存在したとして矛盾を導けばよい。 $\varphi(a) = \{b \in R_1(a) \mid R_1(b) = \emptyset\}$ とおく。 $b_i \succ^{DDP} a_{i+1}$ ならば $|\varphi(b_i)| < |\varphi(a_{i+1})|$ である。そこで $a_i \succsim^{adm,3} b_i$ に対し $|\varphi(a_i)| \leq |\varphi(b_i)|$ が示されれば矛盾が導けて証明が終わる。いま $b \in \varphi(a_i)$ に対し, $R_1(b) = \emptyset$ だから $E = \{b\}$ とおけば $E \in adm(F)$ である。 $a_i \in Atk(E), a_i \succsim^{adm,3} b_i$ より $b_i \in Atk(E)$ である。すなわち $b \in R_1(b_i)$ となる。したがって $\varphi(a_i) \subseteq \varphi(b_i)$ であるから $|\varphi(a_i)| \leq |\varphi(b_i)|$ が成り立つ。 $\square$

⊕DB +DB が両立しないためこちらも両立しない。

+DB 任意の $\succsim^{\sigma,v}$ と両立しない。実際, $a \rightarrow a, c \rightarrow b \rightarrow \gamma(a) \rightarrow \gamma(a)$ なる AF を考えると, $\gamma(a) \succ^{+DB} a$ だが $adm(F) = \{\emptyset, \{c\}\}$ より $a \simeq^{\sigma,v} \gamma(a)$ である。

↑AB $\sigma(F) \subseteq comp(F)$ のとき, $v = 2, 3$ で両立しない。実際, $b \rightarrow a, d \rightarrow c \rightarrow \gamma(b) \rightarrow \gamma(a)$ なる AF において $\gamma(a) \succ^{+AB} a$ だが $comp(F) = \{\{b, d, \gamma(b)\}\}$ より $a \simeq^{\sigma,v} \gamma(a)$ である。また, $\sigma = adm$ のときも $v = 2$ なら同様の AF で $b \simeq^{adm,2} \gamma(b)$ となるので両立しない。 $\succsim^{adm,3}$ で両立する。

証明. +AB と全く同様なので略。

+AB $\sigma(F) \subseteq comp(F)$ のとき, $v = 2, 3$ で両立しない。実際, $b \rightarrow a, \gamma(b) \rightarrow \gamma(a), c \rightarrow \gamma(a)$ なる AF において $a \succ^{+AB} \gamma(a)$ だが $comp(F) = \{\{b, \gamma(b), c\}\}$ より $a \simeq^{\sigma,v} \gamma(a)$ である。また, $\sigma = adm$ のときも $v = 2$ なら同様の AF で $a \simeq^{adm,2} \gamma(a)$ となるので両立しない。 $\succsim^{adm,3}$ で両立する。

証明. 系 5.4 により, 列 $a_0 \succsim^{adm,3} b_0 \succ^{+AB} a_1 \succsim^{adm,3} b_1 \succ^{+AB} \dots \succ^{+AB} a_n \succsim^{adm,3} b_n \succ^{+AB} a_{n+1} = a_0$ が存在したとして矛盾を導けばよい。 $b_i \succ^{+AB} a_{i+1}$ により, 各 a_i は少なくとも 1 つの attack branch をもつ。そのようなものを 1 つとって $x_0 \rightarrow x_1 \rightarrow \dots \rightarrow x_{2k} \rightarrow a$ とおく。このとき $E = \{x_0, x_2, \dots, x_{2(k-1)}, x_{2k}\}$ とすると, E は admissible で $a \in Atk(E)$ であるから, $a_i \succsim^{adm,3} b_i$ により, $b_i \in Atk(E)$ である。したがって, a_i と b_i は同じ弱連結成分に含まれる。そこで w_i を a_i および b_i を含むような弱連結成分の元の個数とする。

このとき, $b_i \succ^{+AB} a_{i+1}$ より, $w_i < w_{i+1}$ であるが, ここから $w_0 < w_1 < \dots < w_n < w_{n+1} = w_0$ となって矛盾である. $\square$

↑DB 任意の $\succ^{\sigma,v}$ と両立しない. 実際, $c \rightarrow b \rightarrow a \rightarrow a, e \rightarrow d \rightarrow \gamma(c) \rightarrow \gamma(b) \rightarrow \gamma(a) \rightarrow \gamma(a)$ なる AF を考えると, $a \succ^{\uparrow DB} \gamma(a)$ だが $a \simeq^{\sigma,v} \gamma(a)$ である.

AvsFD $\sigma(F)$ が grounded extension を含むなら $v = 2, 3$ で成立.

証明. $E \in \text{adm}(F)$ に対し $a \notin \text{Atk}(E)$, $b \notin E$ が容易にいえるので $a \succ_F^{E,v} b$ であることはわかる. とくに ground extension は a を含み b を含まないので $a \succ_F^{\sigma,v} b$ となる. $\square$

以上をまとめると表 1 となる.

	$\succ^{adm,v}$	$\succ^{comp,v}$
VP	×	—
SC	—	—
CP	—	—
DP	×	×
DDP	×	×
⊕DB	—	—
+DB	—	—
↑DB	—	—
↑AB	$\times(v=3), -(v=2)$	—
+AB	$\times(v=3), -(v=2)$	—
AvsFD	✓	✓

表 1: ✓ は一方が他方より強力, × は強力ではないが両立, — は両立しないことを意味している.

表 1 がほとんど “—” であることから, 5 節の Dung の ranking と 4 節の公理とはやはり相性が悪いことがわかる. 序論でも述べたように, Dung の semantics では, 基準が「受け入れられる」か「受け入れられない」の二値になっているため, 要素数などの制約とは合わないことが多い. 一方で, DP や DDP のように数の制約があっても, 両立するものがあるのも驚きである. また, ↑AB や +AB のように branch があるものでも $v = 3$ の場合は両立する. さらに, AvsFD のように, Dung ranking より強力なものもある.

6 ranking semantics の具体例

次に, 具体的な ranking semantics として, [2] に紹介されている ranking semantics Cat, Dbs, Bds, M&T, SAF (それぞれ $\succ^{Cat}$ などのように表記することにす

る) と 5 節で定義した $\succ_F^{\sigma,v}$ の関係を調べる. (これらの ranking semantics の具体的な定義については M&T と SAF のみ具体的に後述する.) 実のところ, $\succ_F^{\sigma,v}$ ($\sigma = \text{adm}, \text{comp}, v = 2, 3$) は両立しない. 証明は省略するが, これらの ranking semantics では, 主張 a, b ($a \neq b$) に対し $a \succ b$ かつ $a \simeq^{adm,v} b$ や $a \sim^{comp,v} b$ となるような例が簡単に作れるためである. そこで代わりに, 次のような ranking semantics $\succeq^{\sigma,v}$ を考える.

定義 6.1. $a \succeq^{\sigma,v} b \Leftrightarrow a \succ^{\sigma,v} b$ または $a = b$.

この $\succeq^{\sigma,v}$ は半順序となる. この $\succeq^{\sigma,v}$ と ranking semantics との両立について見ていこう.

Cat, Dbs, Bds Cat, Dbs, Bds と $\succ^{adm,v}$ と $\succ^{comp,v}$ とは両立しない.

なぜならば, 図 1 の F_1 において, $d \succeq^\delta e$ ($\delta = \text{Cat}, \text{Dbs}, \text{Bds}$) が, $e \succ^{\sigma,v} d$ ($\sigma = \text{adm}, \text{comp}$) だからである.

M&T 次に M&T を説明する.

$F = \langle A, \rightarrow \rangle$ に対し, $X, Y \subseteq A$ として, $Y \leftarrow X$ を $\{(x, y) \in X \times Y, x \rightarrow y\}$ と定め, また $f(n) = n/(n+1)$ とし, $P, O \subseteq A$ に対し $\phi(P, O) = 1/2 \cdot [1 + f(|O \leftarrow P|)] - f(|P \leftarrow O|)$ と定め, さらに

$$r(P, O) = \begin{cases} 0 & (P \text{ は conflict-free でない}) \\ 1 & (P \text{ は conflict-free かつ } P \leftarrow O = \emptyset) \\ \phi(P, O) & (\text{otherwise}) \end{cases}$$

とする. ここで, $a \in A$ を 1 つ固定し, 次のような 2 人ゼロ和ゲームを考える: 自分 (proponent) と敵 (opponent) はそれぞれ戦略として (相手の選ぶ戦略を知らずに) A の部分集合 $P \subseteq A, O \subseteq A$ を選ぶ. ただし, P は $a \in P$ となるように選ばなければならない. ゲームの結果, 自分は敵から利得 $r(P, O)$ を得る. ここで, 互いに最適な混合戦略をとった場合に, 自分が得られる利得の期待値を $v(a)$ とする. $a \succeq^{M\&T} b \Leftrightarrow v(a) \geq v(b)$ と定める.

定理 6.2. M&T は $\succeq^{\sigma,v}$ ($\sigma \in \{\text{adm}, \text{comp}\}, v \in \{2, 3\}$) と両立しない.

証明. まず, $A = \{a_0, a_1, a_2, b_0, b_1, b_2, c\}$, $\rightarrow = \{a_0 \rightarrow a_1, a_1 \rightarrow a_2, a_2 \rightarrow a_0\} \cup \{a_i \rightarrow b_j \mid 0 \leq i, j \leq 2\} \cup \{b_j \rightarrow c \mid 0 \leq j \leq 2\}$ なる AF $F = \langle A, \rightarrow \rangle$ (図 2) において, $v(c) \geq 1/2$ であることを示す.

自分は混合戦略として, $1/3$ の確率で $\{c, a_0\}, \{c, a_1\}, \{c, a_2\}$ をそれぞれ選ぶとする. このとき, いかなる $O \subseteq A$ に対しても $1/3 \cdot [\sum_{0 \leq i \leq 2} r(\{c, a_i\}, O)] \geq 1/2$ であることを示せばよく, 各 i に対し $\{c, a_i\}$ は conflict-free であること, P は conflict-free かつ $P \leftarrow O = \emptyset$ のと

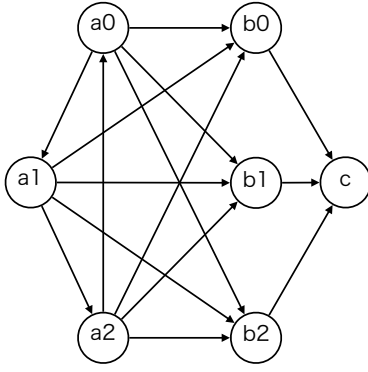


図 2: 定理 6.2 の反例の AF

きは $\phi(P, O) \leq r(P, O) = 1$ であることに注意すれば、 $1/3 \cdot [\sum_{0 \leq i \leq 2} \phi(\{c, a_i\}, O)] \geq 1/2$ であることを示せばよい。

$a_3 = a_0$ とおく。任意の $O \subseteq A$ および $0 \leq i \leq 2$ に対し $|O^{\leftarrow\{c, a_i\}}| = |\{c, a_{i+1}\}^{\leftarrow O}|$ であることを示そう。 $i = 0$ の場合で考える。 $x \in A$ に対し、 $x = a_0, a_2, c$ のとき $|\{x\}^{\leftarrow\{c, a_0\}}| = |\{c, a_1\}^{\leftarrow\{x\}}| = 0$ 、その他の場合、すなわち $x = a_1, b_0, b_1, b_2$ のとき $|\{x\}^{\leftarrow\{c, a_0\}}| = |\{c, a_1\}^{\leftarrow\{x\}}| = 1$ である。したがって、任意の $x \in A$ に対し $|\{x\}^{\leftarrow\{c, a_0\}}| = |\{c, a_1\}^{\leftarrow\{x\}}|$ であるから、 $|O^{\leftarrow\{c, a_0\}}| = \sum_{x \in O} |\{x\}^{\leftarrow\{c, a_0\}}| = \sum_{x \in O} |\{c, a_1\}^{\leftarrow\{x\}}| = |\{c, a_1\}^{\leftarrow O}|$ である。 $i = 1, 2$ のときも対称性によって同様に $|O^{\leftarrow\{c, a_i\}}| = |\{c, a_{i+1}\}^{\leftarrow O}|$ が示される。以上により、

$$\begin{aligned} & \sum_{0 \leq i \leq 2} [f(|O^{\leftarrow\{c, a_{i+1}\}}|) - f(|\{c, a_i\}^{\leftarrow O}|)] = 0 \\ \Leftrightarrow & \sum_{0 \leq i \leq 2} f(|O^{\leftarrow\{c, a_i\}}|) = \sum_{0 \leq i \leq 2} f(|\{c, a_i\}^{\leftarrow O}|) \\ \Leftrightarrow & \frac{1}{3} \left[\sum_{0 \leq i \leq 2} \phi(\{c, a_i\}, O) \right] = 1/2 \end{aligned}$$

となる。よって $v(c) \geq 1/2$ である。

次に、 $F' = \langle A', \rightarrow' \rangle$ を $A' = \{d, e, f\}$ 、 $\rightarrow' = \{d \rightarrow e, e \rightarrow d, e \rightarrow e, e \rightarrow f\}$ と定める。このとき、 f を含むような P のうち conflict-free なものは $\{f\}, \{d, f\}$ の 2 つのみであり、 $r(\{f\}, \{e\}) = 1/2 \cdot [1 + 0 - 1/2] = 1/4$ 、 $r(\{d, f\}, \{e\}) = 1/2 \cdot [1 + 1/2 - 2/3] = 5/12$ である。したがって $v(f) \leq 5/12$ である。

さて、 $F \cup F' = \langle A \cup A', \rightarrow \cup \rightarrow' \rangle$ を考えると、 [5] により、この AF においても $v(c)$ 、 $v(f)$ の値は元の AF におけるそれと変わらない。したがって $c \succ^{M\&T} f$ である。一方で、 $\text{adm}(F \cup F') = \{\emptyset, \{d\}, \{d, f\}\}$ より任意の $E \in \text{adm}(F \cup F')$ に対して $c \in \text{Undec}(E)$ 、 $f \notin \text{Atk}(E)$ であり、 $f \in \{d, f\}$ であるから、 $v \in \{2, 3\}$

に対し $f \succ^{adm, v} c$ である。したがって $\succ^{adm, v}$ と $\succ^{M\&T}$ は両立しない。 $\text{comp}(F \cup F') = \{\{d, f\}\}$ より、 $\succeq^{comp, v}$ と $\succ^{M\&T}$ も両立しない。 $\square$

SAF 次に、SAFの説明をする。SAFとは、 $F = \langle A, \rightarrow \rangle$ に対し、パラメータ $\epsilon > 0$ を 1 つ定め、 $v: A \rightarrow [0, 1]$ を、

$$v(a) = \frac{1}{1 + \epsilon} \prod_{a_i \in R_1(a)} (1 - v(a_i))$$

を満たすように定め、 $a \succ^{SAF} b \Leftrightarrow v(a) \leq v(b)$ としたもののである。このような v は存在することは不動点定理によって示されている。 v は一意に定まると予想されているが、証明はされていない。

SAF に関しては一般的な結果は得られていないため、部分的な結果のみ紹介する。ここでは、well-founded な AF (すなわち、グラフとして acyclic) について考える。このときは順番に計算していくことによって v は一意に定まるので、次が成り立つ。

定理 6.3. $F = \langle A, \rightarrow \rangle$ を well-founded な AF (グラフとして acyclic) とする。このとき、 $\succ^{\sigma, v}$ ($\sigma \in \{\text{grounded}, \text{comp}\}$ 、 $v = 2, 3$) は同じであり、 $\epsilon > 0$ を十分小さくとれば $\succ^{\sigma, v} \subseteq \succ^{SAF}$ である。

証明. [1] より、well-founded な F の grounded extension を G とすると、 $A = G \cup \text{Atk}(G)$ であるから、前半はよい。 F の他の主張から attack されていない主張全体を $G_0 = \text{Def}(\emptyset)$ とし、 F の grounded extension を G とすると、ある N があって $G = \text{Def}^N(G_0)$ である。 $M = \max_{a \in A} |R_1(a)|$ とする。 $\epsilon > 0$ を $(M + 1)^N \epsilon < 1/2$ となるように定める。 $a \in \text{Def}^n(G_0)$ に対し $v(a) \geq 1 - (M + 1)^n \epsilon$ であることを帰納法で示す。 $n = 0$ のときは $1 - \epsilon < 1/(1 + \epsilon)$ より明らか。 $n \geq 1$ に対し、 $a \in \text{Def}^n(G_0)$ とすれば、 $v(a) = 1/(1 + \epsilon) \cdot \prod_{a_i \in R_1(a)} (1 - v(a_i))$ である。各 $a_i \in R_1(a)$ に対し、ある $b \in \text{Def}^{n-1}(G_0)$ があって $b \in R_1(a_i)$ である。帰納法の仮定により $v(b) \geq 1 - (M + 1)^{n-1} \epsilon$ であるから、 $v(a_i) = 1/(1 + \epsilon) \cdot \prod_{b_j \in R_1(a_i)} (1 - v(b_j)) \leq 1 - v(b) \leq (M + 1)^{n-1} \epsilon$ である。したがって、 $|R_1(a)| \leq M$ に注意すれば

$$\begin{aligned} v(a) &= \frac{1}{1 + \epsilon} \cdot \prod_{a_i \in R_1(a)} (1 - v(a_i)) \\ &\geq (1 - \epsilon)(1 - (M + 1)^{n-1} \epsilon)^M \\ &\geq (1 - (M + 1)^{n-1} \epsilon)^{M+1} \\ &\geq 1 - (M + 1)^n \epsilon \end{aligned}$$

となる。ここで、最後の変形において不等式 $(1 - x)^n \geq 1 - nx$ を使った。

したがって、各 $a \in G = \text{Def}^N(G_0)$ に対し、 $v(a) \geq 1 - (M + 1)^N \epsilon > 1/2$ である。一方、各 $b \in \text{Atk}(G)$ に対

し, ある $a \in G$ があって $a \in R_1(b)$ であるから, $v(b) = 1/(1+\epsilon) \cdot \prod_{a_i \in R_1(b)} (1-v(a_i)) \leq 1-v(a) < 1/2$ である. よって $a \succ^{\sigma, v} b$ ($\sigma \in \{\text{grounded}, \text{comp}\}$, $v = 2, 3$) ならば $v(a) > v(b)$ であり, $a \succ^{SAF} b$ である. $\square$

以上をまとめると表 2 となる.

	$\succ^{adm, v}$	$\succ^{comp, v}$
Cat	—	—
Dbs	—	—
Bds	—	—
M&T	—	—
SAF	未 解 決 ($\checkmark$ (grounded で well- founded な場 合))	$\checkmark$ (well- founded な場 合), 未 解 決 (その他の場 合)

表 2: ranking semantics の具体例と Dung ranking の関係 ($\checkmark$ は一方が他方より強力, $\times$ は強力ではないが両立, — は両立しないことを意味している.)

表 2 は SAF 以外は “—” であるが, SAF に関しては, SAF の方が Dung ranking より強力である場合があるのが興味深い.

SAF は ϵ の値によって結果が異なるが, Dung ranking との関係でいえば, できるだけ小さい方が相性が良い. なお, $\epsilon = 0$ の場合は, complete extension の要素の v の値を 1, それ以外の要素の v の値を 0 とすると, v は不動点となる.

7 結論

本論文では Dung の semantics から自然に導かれる ranking として Dung ranking を提案した. Dung ranking はほとんど比較不能となる場合もあるが, 比較できる場合はそれと齟齬がない ranking semantics を使うことが望まれる. そこで, 齟齬がないことを「両立」として定式化し, Dung ranking と ranking semantics で一般的な公理や ranking semantics の具体例との比較を行った. その結果, 両立もしないものが多い一方で, Dung ranking と両立する公理や ranking semantics があることを明らかにした. 今後の課題としては, SAF について, より一般的な場合を解決することがあげられる. well-founded な AF については grounded と complete で SAF より Dung ranking が強力であるのは良い性質である. それを考えると, この良い性質を保ったまま, well-founded という制約をどこまで緩めることができるかが, AF の構造としてどこまで一般的な構造を考え

るが妥当であるかを考える一つの判断材料となる可能性も考えられる.

参考文献

- [1] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2), pages 321-357, 1995.
- [2] Elise Bonzon, Jérôme Delobelle, Sébastien Konieczny, and Nicolas Maudet. A comparative study of ranking-based semantics for abstract argumentation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI’16, pages 914-920, AAAI Press, 2016.
- [3] L. Amgoud and J. Ben-Naim. Ranking-based semantics for argumentation frameworks. In *Scalable Uncertainty Management*, pages 134-147, Springer, 2013.
- [4] C. Cayrol and M.-Ch. Lagasque-Schiex. Graduality in argumentation. *Journal of Artificial Intelligence Research*, 23, pages 245-297, 2005.
- [5] P.-A. Matt and F. Toni. A game-theoretic measure of argument strength for abstract argumentation. *Procs. of JELIA ’08, LNCS 5293*, pages 285-297. Springer, 2008.
- [6] J. Leite and J. Martins. Social abstract argumentation. In *Proc. 22nd Int’l Joint Conf. on Artificial Intelligence*, pages 2287-2292, 2011.

データマイニングとテキストマイニングの連携によるデータ分析支援

Data Analysis Support by Combining Data Mining and Text Mining

松本友哉 砂山渡 畑中裕司 小郷原一智

Tomoya Matsumoto Wataru Sunayama Yuji Hatanaka Kazunori Ogohara

滋賀県立大学 工学部

School of Engineering The University of Shiga Prefecture

Abstract: 近年では、アンケート分析などにデータマイニングやテキストマイニングの技術が用いられることが多くなってきた。しかし、データマイニングでは数値データ、テキストマイニングではテキストデータの分析を主とするため、その両者を含むアンケートデータ等の分析においては、用いるツールの選択が難しい状況があった。そこで本研究では、テキストマイニングのための統合環境 TETDM をベースとして、これにデータマイニングの機能を追加することで、両方向からのデータの絞込みとデータの解釈に役立てられる環境を提案する。

1 序論

近年では、アンケート分析などにデータマイニングやテキストマイニングの技術が用いられることが多くなってきた [1]。データマイニングは、大量の数値データから相関関係や法則性を発見することができるため、スーパーマーケットで一緒に購入されることが多い商品の組み合わせを調査などに利用されている。分析手法としては、頻出の組み合わせを発見するアソシエーション分析やデータ集合のまとまりを把握するためのクラスター分析などの手法が用いられている。一方、テキストマイニングは、大量のテキストデータから、文章の要約や否定的な意見の抽出ができるため、アンケート調査やコールセンターでの利用者からの意見の解析や分類に利用されている。分析手法としては、単語間のつながりを表す共起ネットワークや辞書登録をした言葉からポジティブな言葉とネガティブな言葉の判別する手法などが用いられている。そこで、データマイニングとテキストマイニングを組み合わせることにより、データマイニングでよく売れるルールを発見し、さらにテキストマイニングを使ってその消費者の意見を解析して理由まで発見するといった、両者の良い点を併せ持った分析の方法が考えられる。しかし、データマイニングでは数値データ、テキストマイニングではテキストデータの分析を主とするため、その両者を含むアンケートデータ等の分析においては、用いるツールの選択が難しい状況があった。

そこで本研究では、テキストマイニングのための統合環境統合環境 TETDM (Total Environment for Text

Data Mining) [2] をベースとして、これにデータマイニングの機能を追加することで、両方向からのデータの絞込みとデータの解釈に役立てられる環境を提案する。

2 関連研究

2.1 データ分析支援環境

データマイニングの分析支援環境としては、R[3] や Weka[4] がある。また、テキストマイニングの分析支援環境としては、「KH Coder」[5] や「UserLocal」[6] がある。これらのツールは、主に数値データまたは、テキストデータのどちらかのみを入力することを想定している。本研究では、数値データとテキストデータの2種類の入力に対応し、数値データとテキストデータ間で連携が可能な分析環境を開発する。

2.2 データマイニングとテキストマイニングの連携

「市場分析におけるテキストマイニングを活用したデータマイニングの実践」[7] という研究では、データマイニングでルールセットを発見し、テキストマイニングを用いてそれらのルールセットをフィルタリングする方法で連携がとられている。また、データマイニングから見たテキストマイニングの利用法に関する研究 [8] も存在する。これらの研究では、データマイニングまたは、テキストマイニングの片方が、もう片方を利用するという一方向の連携であるが、本研究では、

表 1: 本環境に使用する入力データの例

ID	年齢	性別	総合	外見	サイズ	テキスト
1	50	1	4	4	4	軽く、小さい、移動するのが便利です。
2	30	2	4	5	4	風量が弱い分音も静かです。
3	40	2	4	5	5	10年ほど前に買ったヒーターの替えとして購入しました。
4	40	2	5	5	5	寒がりな私にはピッタリです。満足しています。
5	50	1	3	3	4	消し忘れに注意しましょう！コンパクトサイズでも暖房力あり！満足しています。
6	50	1	5	5	5	

データマイニングとテキストマイニングの双方向において絞り込み、分析ができる環境の開発を目指す。

3 データマイニングとテキストマイニングの連携分析環境

3.1 対象データとデータ形式

本環境では、数値データとテキストデータの両方を含むデータ集合を対象としている。例として、数値による選択回答と自由記述の両方を実施しているアンケートデータがある。本分析環境の入力におけるデータの形式の例を表1に示す。

3.2 分析環境の枠組み

本研究でのデータマイニングとテキストマイニングの連携の定義について述べる。本研究において連携とは、図1における、総合得点で5と回答したデータだけを抜き出すような数値情報による絞り込み、及び、テキストの中に特定単語が出現しているデータだけを抜き出すような単語によるデータの絞り込みを行い、絞り込んだデータの数値部分をデータマイニングツールに、テキスト部分をテキストマイニングツールへの入力とすることによる数値とテキストによる絞り込みと分析を組み合わせた分析方法を指す。

図1にデータマイニングとテキストマイニングの連携分析環境の枠組みを示す。

本環境は、テキストマイニングソフトのTETDM[2]に数値情報による絞り込み機能とデータマイニングツールを統計ソフト「R」[3]を用いて追加したものである。本環境における分析の流れを次に記載する。ま



図 1: 本分析環境の枠組み

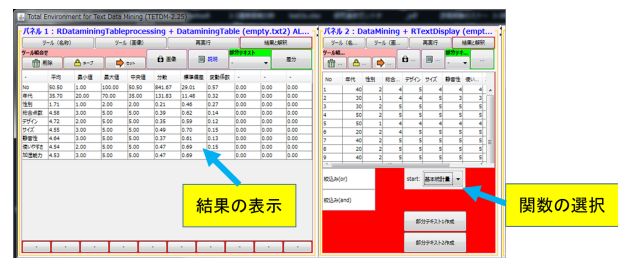


図 2: 「DataMining」及び「DataMiningTable」の操作画面

ず、図1のAにあたる部分でアンケートデータなどの数値とテキストがセットになったデータを入力する。図1のBにあたる部分では、Aで入力したデータに対する分析結果が表示され、データマイニングツール14種、テキストマイニングツール48種を組み合わせる分析ができる。図1のCにあたる部分で必要に応じて数値、または単語でデータを絞り込むことができ、再度Aへ入力し、Bにて分析を行うことができる。分析後に分析の結果から言えることを登録し、登録した知見を統合することで知見を獲得する。

3.3 入力データの形式

入力データは、属性と属性値のペアとして与えられるトランザクションデータとして与えられるものとする。そのうち、数値データ部分を抜き出したcsv形式のデータを数値データとして与え、テキスト部分はテキストファイルにまとめて入力として与える。

TETDM上でのテキストの取り扱い、以下のように分割して認識される。テキストデータ全体を「文章」、テキストデータ中の段落を「スナリバラフト」という文字列を段落間に挿入することで「セグメント」、句点で区切られた文を「文」として認識する。データを回答者などに区別する場合は、セグメントで分割する。

3.4 分析に用いられる機能

3.4.1 データマイニング用ツール

本研究では、TETDM と同じ Java に対応している点、一度作成した後も関数の追加が容易である点の2点から統計解析ソフトの「R」を用いてデータマイニング部分を TETDM に追加実装し、データマイニングとテキストマイニングの連携分析環境を構築を行った。そこで、データマイニング用ツールとして計算手法の選択に使用する先ほどの「DataMining」と計算結果を表示する「DataMiningTable」を作成した。「DataMining」及び、「DataMiningTable」の表示画面を図2に示す。

本ツールではデータマイニング機能として、基本統計量、基本統計量（フォーカス）、相関行列とアソシエーション分析 [9] を搭載している。基本統計量では、入力した数値データの列ごとの平均値、最小値、最大値、中央値、分散、標準偏差、変動係数を表示する。変動係数とは、単位の異なるデータ間のばらつきを変動係数同士で比較することで相対的に評価することができる数値である。基本統計量（フォーカス）では、入力した数値データの列全体の基本統計量に加え、着目しているデータのみを抜き出した基本統計量を追加して表示する。相関行列では、各列間の相関係数を行列形式で表示する。アソシエーション分析とは、例えば商店などで買い物が商品 A と商品 B を一緒に買う確率などを求めることができる分析である。

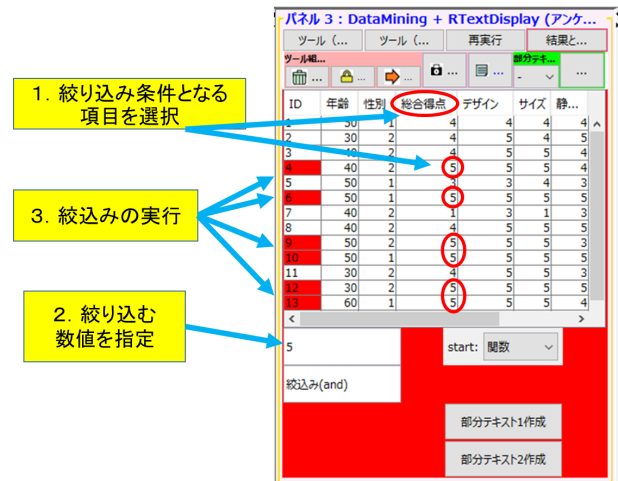


図 3: 数値情報によるデータの絞り込み用ツール

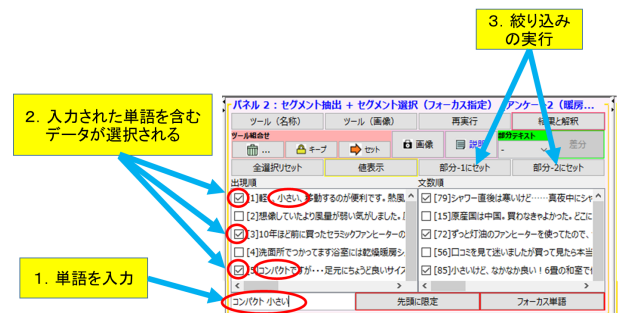


図 4: 単語情報によるデータの絞り込み用ツール

3.4.2 テキストマイニング用ツール

本分析環境の、テキストマイニングツールとして、TETDM の以下のツールを用いた。

- 「単語抽出」：指定した文字列をテキスト中でハイライト表示
- 「文章要約」：指定したキーワードに着目した文章の要約を表示
- 「失礼単語抽出」：否定的な表現をハイライト表示
- 「テキスト分類（再帰的クラスタリング）」：テキスト中の共起関係にある単語群を地図形式で表示
- 「単語情報まとめ」：単語の出現の頻度を表示

3.5 データの絞り込み機能

3.5.1 数値情報によるデータの絞り込み用ツール

数値データとテキストデータの両方を分析する上で、例えば、製品のどの部分が評価されているのかを分析するために、レビュー評価の高いテキストデータだけ

に絞り込みたい場合が想定される。そこで、TETDM 上で数値情報によるデータを絞り込むためのツールとして「DataMining」を作成した。「DataMining」の操作画面を図3に示す。

3.5.2 単語情報によるデータの絞り込み用ツール

数値データとテキストデータの両方を分析する上で、例えば、製品について「臭い」について言及している人の総合得点の平均などの数値情報や要約などを分析するために、「臭い」について言及している人だけにデータを絞り込みたい場合が想定される。そこで、TETDM 上で単語情報によるデータを絞り込むためのツールとして「セグメント抽出」がある。「セグメント抽出」の操作画面を図4に示す。

3.6 データマイニングとテキストマイニングの連携環境の構成

データマイニングとテキストマイニングの連携環境として用意した5つの組み合わせを以下で述べる。

3.6.1 「DM ↔ 原文参照」

「DataMining」の相関行列やアソシエーション分析により、レビューの総合得点と相関関係にある評価項目を発見し、「単語抽出」にてその評価項目で高評価や低評価を回答している回答者のテキストを絞り込んで原文を見ることでその理由を探ることが可能である。

3.6.2 「DM ↔ 要約」

「DataMining」の相関行列やアソシエーション分析により、レビューの総合得点と相関関係にある評価項目を発見し、「文章要約」によりその評価項目で高評価や低評価を回答している回答者のテキストを絞り込んで要約を見ることで絞り込んだデータの概要を知ることが可能である。

3.6.3 「DM ↔ 失礼単語着目」

「失礼単語地抽出」より、否定的な意見をハイライト表示させて否定的な意見や単語に着目し、「セグメント抽出」で着目した単語でデータを絞り込んで「DataMining」の「基本統計量」を見ることで、否定的な意見の男女比を見ることなどが可能となる。

3.6.4 「DM ↔ クラスタリング」

「テキスト分類（再帰的クラスタリング）」より、高評価につながる表現と共起関係のある単語に着目し、「セグメント抽出」で着目した単語でデータを絞り込んで「DataMining」の「基本統計量」を見てやることで、その単語に関係のあるの男女比を見ることなどが可能となる。

3.6.5 「DM ↔ 単語頻度」

「単語情報まとめ」より、出現頻度の高い単語など気になる単語に着目し、「セグメント抽出」で着目した単語でデータを絞り込んで「DataMining」の「基本統計量」を見てやることで、その単語に関係のあるの男女比を見ることなどが可能となる。

3.7 「DM ↔ 要約」の利用例

本節では、本環境内の「ツールセットのひとつである「DM ↔ 要約」の利用例を説明する。利用方法を説明する際の分析者の状況設定として、数値によるの評価項目と自由記述のある商品のレビューデータから、「ファンヒーター」に関するレビューの総合得点が高い回答者は、どの評価項目を重視しているのか、また、その理由を知りたい場合を想定する。以下に想定した場合の目的を達成するための「DM ↔ 要約」の具体的な分析例を示す。そのときのパネルの様子を図5に示す。

1. レビューデータの数値データ部分に該当する csv ファイルを TETDM フォルダの「csvfile」フォルダ内に格納する。
2. レビューデータのテキストデータに該当する txt ファイルを TETDM フォルダの「text」フォルダ内に格納し、メニューウィンドウ左上「テキスト入力」、「ファイル」ボタンより選択する。
3. ツール「RApplication」の「DM ↔ 要約」ボタンを押して「DM ↔ 要約」をセットする。
4. 「DataMining」のアソシエーション分析で、総合得点が5と一緒にでてくるルールセットを探す。
5. アソシエーション分析をした結果、全体の46パーセントの人が「総合得点」の評価が5点であると同時に「使いやすさ」の評価が5点で回答していることがわかる。
6. 「DataMining」の数値情報による絞り込み機能を用いて、「使いやすさ」の評価を5と回答した回答者のみデータを絞り込む。
7. 「部分テキスト作成1」ボタンを用いて、「使いやすさ」の評価を5と回答したデータのみのテキストデータに加工する。
8. 「文章要約」パネル右上の部分テキストから、「部分テキスト1」をセットし、先ほど絞り込んだテキストデータを「文章要約」の入力データとする。
9. 「文章要約」のキーワードの中から「使いやすさ」という観点から連想される「便利」、「機能」といったキーワードに着目して要約を行う。
10. 要約結果から、「余計な機能が無くわかりやすい」点が評価されていると推測できる。

これらの手順から、例えば、商品の購入者は「使いやすさ」を重視しており、この商品は、余分な機能が無くわかりやすいことが高評価につながっていると解釈できる。

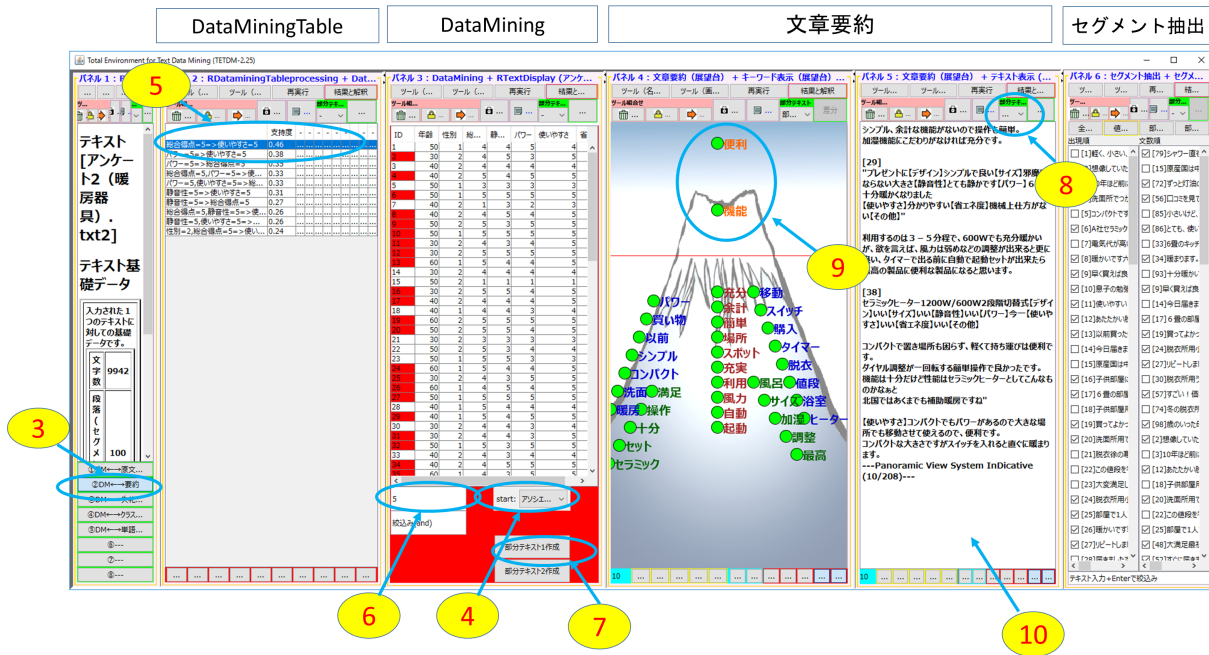


図 5: 「DM ↔ 要約」の利用例

4 評価実験

本章では、本環境の使用によるアイデアの具体性の変化を検証した実験について述べる。

4.1 実験内容

本節では、分析環境の評価実験内容について述べる。提案分析環境によって得られる知見の着目点の有無と具体性の有無を評価するために実験を行った。被験者は、大学生・大学院生の計 16 名を対象に、後述する指定の分析環境を用いて、商品のレビューデータを分析し、「利用者の満足度（レビューの総合点数）が高くなると予想される、新製品を提案」をしてもらった。提案環境で分析を行う A グループと、比較環境で分析を行う B グループの 2 グループを 8 名ずつに分けて実験を実施した。提案環境の A グループは、本研究の連携機能がある単語、数値で絞り込んだデータをデータマイニング、テキストマイニングの両方で分析できる環境で 3.6 で紹介したツールセット 5 つを使用できる。一方で比較環境の B グループは、本研究におけるデータマイニングとテキストマイニングの連携機能が無い、3.6 で紹介したツールセットをばらしたデータマイニングツール 1 種、テキストマイニングツール 5 種の計 6 つのツールを連携機能が無く個別に実施できる環境下である。今回の実験で使用するアンケートデータは、商品の機能やデザインなどを評価する 5 点満点の数値とレビューテキストがセットになった形式とし、「楽天市

場」[10] で販売されている「加湿器」と「ファンヒーター」のレビューデータを各 100 件を使用した。被験者には、商品のレビューデータより、「利用者の満足度（レビューの総合点数）が高くなると予想される、新製品の提案」を目的として、以下の手順で行ってもらった。

1. データを数値や単語によって絞り込む。
2. 分析に用いるツールをセットする。
3. 分析して発見した「結果と解釈」を登録する。
4. 手順 1 から 3 を繰り返した上で、集めた「結果と解釈」をまとめ、新製品の提案を行う。

4.2 最終提案における着目点の有無と具体性の有無の評価

本節では、最終提案における着目点の有無と具体性の有無の評価の方法と考察について述べる。データマイニングとテキストマイニングの連携の有無により、最終提案における着目点の有無と具体性の有無の評価を行った。着目点の有無とは、アイデアが性別や年齢といったどの顧客層に対して向けられたものなのか判断できる場合は着目点有とした。また、改善に関する複数の概念のうち他よりも重要度が高い旨の記述がある場合も着目点有とした。具体性の有無の評価については、例として単に「デザインを良くする」と言った提案だけでなく、デザインの方向性や程度が記載されて

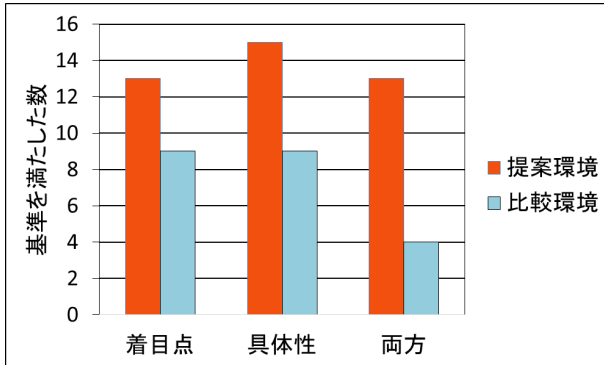


図 6: 最終提案における着目点の有無と具体性の有無の評価の集計結果

いる場合は、具体性有と評価した。提案環境と比較環境における「加湿器」と「ファンヒーター」の分析結果を評価し、集計したものを図 6 に示す。

図 6 より、提案環境の方が着目点有の数、具体性有の数、両方有の数の全てにおいて上回っている。提案環境のアイデアの一例を挙げると「購入者は女性が多く、また、女性はデザインを重視していたため、デザインを可愛らしく凝ったものにするべきである。」といった提案があった。「女性が多く」、「デザインを重視していた」という点からアプローチする客層や他の評価項目よりも重視していることがわかるため、着目点有と評価した。また、「デザインを可愛らしく凝ったもの」の点から、デザインの方向性や程度がわかることから、具体性有と評価した。これより、データマイニングツールとテキストマイニングツールを個別に使用するよりも、連携して得られた知見のほうが、着目する箇所が示され具体的な内容になったことが確認できる。

次に、提案環境と比較環境におけるデータマイニングツールとテキストマイニングツールの使用頻度の偏りと最終提案の評価点の平均を比較し、表にまとめたものを表 2 に示す。ただし、ツール使用頻度の偏りは、TM の利用率と DM の利用率との差を表し、最終提案点数は、被験者の各提案に対して着目点有で 1 点、具体性有で各 1 点と評価し、「加湿器」と「ファンヒーター」の提案の評価点を合計したものを表す。

表 2 より、提案環境のほうが、ツールの使用頻度の偏りが比較環境より 0.37 低く、最終提案の評価点の平均が比較環境より、1.2 点高いことがわかる。以上より、提案環境のほうがデータマイニングとテキストマイニングツールをバランスよく使用できるため、着目点が明確になり、かつ具体的な知見を得ることができた。

表 2: 提案環境と比較環境のツール使用頻度の偏り及び分散と最終提案の評価点の平均

	ツール使用 頻度の偏り の平均	ツール使 用頻度の 分散	最終提案 の評価点 の平均
提案環境	0.42	0.6	3.5
比較環境	0.79	0.19	2.3

5 結論

本研究では、数値データ、テキストデータの両方を含む入力に対し、データマイニングとテキストマイニングの両方向からのデータの絞り込みと分析が行える環境を構築した。本環境で分析することで、得られる知見の具体性が向上することを実験により検証した。

参考文献

- [1] 佐藤弥生, 佐々木千晶: 介護職員の「不適切ケア」の判断の拠り所-アンケートの自由記述の分析から-, 岩手県立大学社会福祉学部紀要第 18 巻, pp.11 – 22, (2016)
- [2] Total Environment for Text Data Mining : (URL)<http://tetdm.jp>
- [3] GNU R : (URL)<https://www.r-project.org/>
- [4] Weka (URL)<http://www.cs.waikato.ac.nz/ml/weka/downloading.html>
- [5] KH Coder : (URL)<http://khc.sourceforge.net/>
- [6] UserLocal : (URL)<http://textmining.userlocal.jp>
- [7] 竹内広宜, 杉山喜昭, 山口高平: 市場分析におけるテキストマイニングを活用したデータマイニングの実践-生ごみ処理機の市場分析を例として-, 知能と情報, Vol.24, No.3, pp.728 – 742, (2012)
- [8] 善田昌樹: データマイニングの視点から見たテキストマイニングの三つの利用法, 日本情報系学会誌, Vol.35, No.1, pp.4 – 18, (2014)
- [9] 荒木孝治: R と R コマンドで始める多変量解析, p.180, 日科技連, (2007)
- [10] 楽天市場 : (URL)<http://www.rakuten.co.jp/>

研究初心者のサーベイ行為を対象とした 論文整理支援システムの基礎検討

A Supporting System for Organizing Paper in Novice Students' Survey Behavior

西村 勇哉^{1*} 大杉 隆文¹ 盛山 将広² 内藤 峻² 松下 光範¹
Yuya Nishimura¹ Takafumi Osugi¹ Yukihiro Moriyama² Shun Naito² Mitsunori Matsushita¹

¹ 関西大学 総合情報学部 総合情報学科

¹ Faculty of Informatics, Kansai University

² 関西大学大学院 総合情報学研究科

² Graduate School of Informatics, Kansai University

Abstract: This goal of our study is to encourage novice students to understand the research field. We propose a thesis information visualization interface to support research reorganization for novice students. Novice Students don't have knowledge and experience of research activities. So novice beginners cannot make good use of the survey. Therefore, in this paper, we will visualize the relation of surveyed paper information. We use Euler diagram for visualization. By experiments, novice students were able to grasp information surveyed by themselves. And it was confirmed to promote understanding of the research field.

1 はじめに

現在、論文検索のための検索エンジン (e.g., Google Scholar, CiNii) や学会のポータルサイトなどを利用することで、Web 上での論文検索が可能である。年々新しい分野の研究が行われており、研究分野が細分化されている [1]。さらに、学生の卒業論文や著名な論文誌に掲載された論文など、多様な論文が存在しているため、Web 上には数多くの論文が存在する。研究者はこの膨大な論文の中から、自身の研究分野において、これまで行われてきた研究や最新の動向を把握するために、関連のある分野、取り扱う技術、研究対象とするユーザなどに着目して論文を検索し、読むことで内容を理解する。このような行為はサーベイと呼ばれ、研究活動において、自身の研究の位置付けを明確にするために欠かせない作業である。

大学研究室に所属し、研究活動を目的とした論文執筆を経験していない学生 (以下、研究初心者) は、多くの場合、研究室に所属した後に初めて研究を行うため、サーベイや研究に関する知識や経験がない場合が多い。そのため、サーベイをすべき研究分野や、サーベイを通して得られる情報の適切な整理方法などがわからず、

自身の研究分野についての理解を深めることが難しくなる。また、自身の研究分野に関係している技術や研究分野を把握していない場合、その技術や、研究分野の知見を活用することで生まれる新たな知見に気づくことができなくなる。

そこで本研究では、研究初心者を対象として、サーベイにおける検索方法やサーベイした論文についての理解を促し、研究分野への理解を深めさせるためのシステムの実現を目指す。本稿では、その端緒として、研究初心者がサーベイに対して抱える問題点に関する調査と、そこで得られた知見に基づいて作成したプロトタイプシステムについて述べる。

2 関連研究

2.1 情報の可視化

サーベイにおける論文検索方法の 1 つとして、既に読んだ論文が参照している論文をもとに次に読む論文を選択する方法がある。井上らは、ユーザが選択した論文とその参考論文との関係の可視化システムを提案している [2]。このシステムは、ユーザが選択した論文が参照しているものを文字ではなく図として提示することで、参照関係の繋がりを一目で認識させることを

*連絡先：関西大学総合情報学部総合情報学科
〒 569-1095 大阪府高槻市霊山寺町 2-1-1
E-mail: mat@res.kutc.kansai-u.ac.jp

狙いとしている。さらに、ユーザが選択した論文に対する参照理由や、実際に論文の参照箇所の情報などを表示することで、ユーザは必要となる論文を容易に取捨選択できるようにしている。実験結果から、これらの機能によりユーザの論文サーベイの検索効率が向上することが示された。一方で、論文の参照関係の可視化と、参照理由などの情報提示を別画面で行っていることが、ユーザの情報把握の負担になることが示唆された。

2.2 ユーザの意図の考慮

伊藤らは、ユーザ主導で情報の共有や創造の支援を目的としたある事柄に対しての関連情報の可視化システムを提案した [3]。提案手法では、可視化システムがユーザに対して可視化表現を変更する手段を十分に与えられていないことに着目している。そこで、多様な関連情報の表示形式を用意することでユーザ側に表現を変える手段を与えている。木構造やグラフ構造など、情報の表示形式をユーザが任意に選択することで、ユーザが注目している情報の概要を理解できたと報告されている。さらに、問題構造を可視化する機能により、新たな知見を見つけたり考えをまとめたりする発想支援に関わる効果が期待できることが示唆されている。

このことから、ユーザの情報整理の意図を考慮した支援を行うことで、研究初心者のサーベイした論文全体の関係性を把握させることにつながると考えられる。そこで本研究では、ユーザがサーベイした論文内容を考慮した情報の関係性を可視化し、論文情報の整理支援を行う。

3 論文整理を支援するインタフェース

3.1 デザイン指針

本研究では、研究初心者の学生のサーベイを支援する手段として (1) 情報を直感的に捉えることができる情報提示手法、(2) 情報比較の円滑化、この 2 点に着目しシステムの設計を行う。

伊藤らの研究 [3] では、ユーザがある事柄に対して関連する情報を付与し、その関連情報についての可視化を行うことで、ある事柄の概要把握につながると示唆されている。そこで提案インタフェースでは、関連情報として、サーベイした論文の内容を一言 (e.g., 整理支援, 情報可視化) で表す情報 (以下, テーマと記す) をユーザに入力させる。これにより、ユーザは論文の内容について考慮するようになり、論文内容を理解するきっかけを与える。論文に付与できるテーマ数を無



図 1: オイラー図を用いた整理結果の概念図

制限にすると、ユーザがサーベイした論文の内容や論文間の関係性を考慮しないことが危惧されるため、論文に付与できるテーマ数を最大 3 つという制限を設けた。また、ユーザに任意に論文内容をまとめさせるため、自由記述が可能なコメント欄を設ける。

Keller らの研究において、可視化を用いた情報提示は知識獲得の支援につながることが示唆されている [4]。そこで提案インタフェースでは、論文に付与されたテーマに基づき論文間の関係性の可視化を行う。論文間の関係性を表現するために、可視化には各集合の相互関係を表現可能なオイラー図を用いる。今回提案するインタフェースの可視化結果の概念図を図 1 に示す。提案手法ではテーマを 1 つの集合とし、円の大きさはテーマに含まれる論文数を、色の違いはテーマの違いを表す。また、情報入力時、閲覧時に、可視化画面と、ユーザが入力する情報や論文情報を同時に閲覧可能にする。これらにより論文情報を直感的に捉えることができるため、論文整理の負担を軽減できると考えられる。

3.2 実装

提案インタフェースは、HTML, CSS, JavaScript, jQuery を用いて実装した。なお、可視化部分は D3.js で実装された Venn.js を用いた。D3.js とは、データ駆動型の情報可視化ライブラリであり、データの変更に応じて動的に可視化を行うことができる [5]。

提案インタフェースの概観を図 2 に示す。提案インタフェースは、ユーザが入力する論文情報に基づいて可視化を行う可視化画面 (図 2 中 ④) と、テキストにより、可視化画面の論文情報を補足するテキスト画面 (図 2 中 ⑤) による 2 種類の画面で構成される。

可視化画面では Venn.js を用いて、1 つのテーマを 1 つの集合としたオイラー図として論文情報を可視化する。入力した論文が複数のテーマを情報として持っている場合、円は重なって描画される。円の大きさは、テーマに含まれる論文の数によって変動する。これにより、ユーザはサーベイした論文の関係性を直感的に

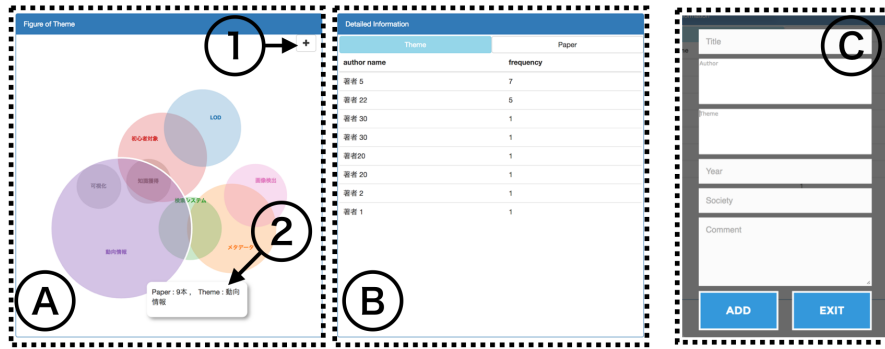


図 2: 提案インターフェース

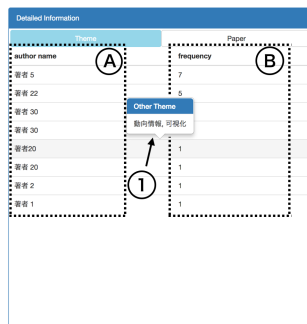


図 3: テキスト画面の例

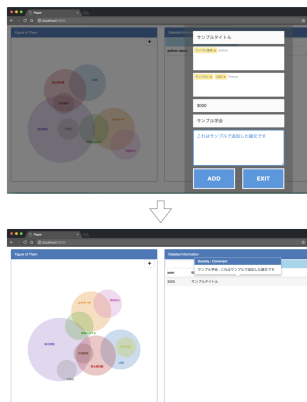


図 4: サーベイした論文の情報を入力する図

捉えることができる。また、可視化されたテーマにマウスオーバーすることでマウスカーソルの右下に現在選択しているテーマと、そのテーマに含まれる論文の数を提示する(図2中②)。この機能により、複数のテーマが混在し、ユーザが選択しているテーマが視覚的に捉えにくい場合、選択している論文情報を正確に把握できる。さらに、可視化されたテーマをクリックすることで、そのテーマに基づく情報がテキスト画面に表示される機能を実装した。これにより、テーマとテキストの情報が連動するため、ユーザは論文情報の

比較を円滑に行うことができる。テキスト画面は、全ての情報を一度に提示するとユーザの論文情報の整理の妨げになると考えられる。そのため、ユーザが選択したテーマに含まれる著者に関する情報と論文に関する情報に分けて提示している。

著者に関する情報に関して選択したテーマが付与されている論文の著者が一覧表示される(図3中④)。選択したテーマについて多くの研究をしている著者に気づかせ、どの著者がどの分野に精通しているのか把握させるために、選択したテーマについて論文を書いた数を表示し(図3中⑤)、その数が多い順に並べている。また、選択した著者が研究しているテーマをポップアップ形式で表示させる(図3中①)。選択した著者が研究しているテーマや執筆している論文数を把握することで、ユーザがサーベイの必要なテーマの把握につながると考えられる。論文に関する情報は、選択したテーマが付与された論文のタイトルと発行年が一覧で表示される。また、ユーザが任意に付与したコメントと、論文の発行元をポップアップ形式で表示する。これにより、ユーザは任意で付与した情報を振り返ることができる。

3.3 操作方法

図2中Cへの情報入力の様子を図4に示す。提案インターフェースでは、ユーザが入力フォームに論文情報を入力する。情報入力フォームは「+」ボタンを押すことで表示することができる(図2中①)。入力する情報は、タイトル、著者名、テーマ、発行年、学会名、ユーザ任意のコメント、の6種類である。入力を終え、「ADD」をクリックすることで情報が追加され、入力したテーマに基づき、論文間の関係性が可視化される。「EXIT」をクリックすると入力フォームが閉じる。可視化された円にマウスオーバーすると、その円のテーマ名と、論文の数が表示される。円をクリックするとテキスト画面に論文情報が表示される。

4 ユーザ観察

本実験の参加者は、本研究の対象となるユーザに対して提案インタフェースの機能がユーザの論文整理の作業に対して有効に働くのかを検証することを目的とする。そこで、ユーザがサーベイした論文の整理に提案インタフェースのどの機能を使っているのかに注目して観察を行った。

4.1 実験手続き

本実験の参加者は、情報学部に通う 20 代の大学生 6 人(男性:5 人, 女性:1 人)である。このうち, 3 人は研究活動を始めて 1 年未満(以下 A 群と記す), 3 人は研究活動を始めて半年未満の学生である(以下 B 群と記す)。本実験において, A 群の実験参加者は, 実験参加者が今までにサーベイした論文を電子媒体で持参させた。B 群の実験参加者は, 対象ユーザの状況と課題を設定した。設定内容は, 「コミック工学を研究する学生であり, 自身の研究分野についての先行研究を探す」というものである。

実験では, まずインタフェースの使い方と, 実験課題についての説明を行った。課題は, A 群は持ち込んだ論文を整理させ, B 群はサーベイをしながら論文を整理させた。両群, 整理が終了した段階, もしくは実験開始から 45 分後に終了し, 整理した結果をインタフェースを用いて振り返らせた。その後, 論文を 1 件サーベイさせた。実験中は画面録画を行った。

課題終了後, 自身がサーベイした論文の整理ができたか, サーベイについて理解が深まったかについてのアンケートに回答してもらい, その後, 提案インタフェースの機能について, 半構造化インタビューを行った。アンケートの質問項目を以下に示す。

1. 可視化により情報を一目で捉えられたか
2. インタフェースが整理の役に立ったか
3. インタフェースが整理後のサーベイに役立ったか
4. 情報量は適切であったか
5. サーベイの理解が深まったか
6. 研究分野への理解が深まったか

実験参加者には各質問に対して 5 段階(1: そう思わない~ 5: そう思う)で評価とその回答理由を答えさせた。回答時間に制限は設けず, 全ての設問に回答した時点で終了とした。

インタビューは, 以下の 3 つの質問に応じてもらい, その理由を述べてもらった。

1. システムに入力した情報が可視化される機能はあなたの整理の役に立ちましたか
2. テキストで表示される情報はあなたの整理の役に立ちましたか
3. 今回整理をして得られた知見において提案インタフェースは役に立ちましたか

回答時間に制限は設けず, 全ての質問に回答した段階で終了とした。また, インタビュー内容は録音を行った。

4.2 実験の結果

実験で得られた提案インタフェースの可視化画面の例を図 5 に示す。4 名の実験参加者の論文整理結果は図 5 のような大きな円 3 つを中心とした形となった。図 5 の場合, 「赤外線」を含む論文は「プロジェクション型 AR」, 「ProCam」と関係あることがわかる。「MR」と「プロジェクション型 AR」の違いを示すものとして「赤外線」が関係していることがわかる。また, 「タンジブル・ビット」は他のテーマと関連していないため, 自身の研究と関係を持っていないことがわかる。

このような可視化画面を参考にして, 5 名の実験参加者が論文に付与するテーマを考慮する様子が確認できた。特に B 群の実験参加者は 3 人全員が次に読む論文を検索したり, 選択したりする際, 論文検索画面と提案インタフェースの画面を行き来していた。この時, 提案インタフェースの可視化画面上でマウスカーソルを動かす様子が見られた。論文の内容とこれまでの自身のサーベイしてきた論文との比較や, どのテーマについてサーベイすべきかを考慮していたと考えられる。また, 「1 つしか論文を調べていなかった」, 「重なりが大きい」という発言が確認できたことから, 実験参加者は, 自身でサーベイが不足しているテーマに気づいていなかったことが確認された。可視化画面のテーマの重なりによってサーベイした論文間の関係性を把握し, 新たな知見が得られたと考えられる。

実験参加者のコメント欄の使い方は様々であった。論文内容をまとめた文章, 読んだ論文について一言を書く, 何も入力しないなど, 実験参加者によって記録する内容に違いが見られた。また, 実験参加者が自身のサーベイを振り返る際, コメントを利用している様子が確認されることが少なかった。これは, コメントに入力する内容に規定はなく自由度が高いため, 入力内容が統一されず情報が煩雑になったためと考えられる。

以上のことから, 提案インタフェースの可視化画面は, ユーザがサーベイした論文間の関係性を可視化することで, ユーザの論文情報整理の支援につながり, そこで得られた結果から新たな知見に気づくことが示唆された。しかし, 情報の補足のためのテキスト画面が

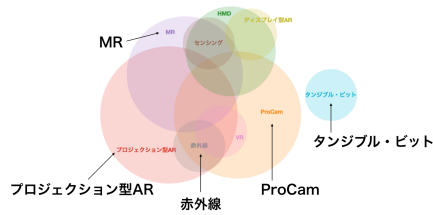


図 5: 実験結果の可視化画面の例

表 1: アンケート結果の平均値

質問番号	A 群	B 群	全体
1	4.67	4.00	4.33
2	4.33	4.67	4.50
3	3.67	3.33	3.50
4	3.33	3.67	3.50
5	2.67	4.33	3.50
6	4.00	4.33	4.17

利用されることは極めて少なかった。これは、実験時間に制限があったため、提案インタフェースに入力した論文情報の量が少なく、テーマごとの著者に関する情報の傾向が現れなかったことが原因の 1 つであると考えられる。

4.3 アンケートの結果と考察

提案インタフェースに対して行った 5 段階での評価の平均値を表 1 に示す。設問 1 番と 2 番において、全員が 4 点以上の評価をしていた。設問 3 番と 4 番においては他の質問に比べて低い点数となっている。設問 5 番において、A 群の平均点が 2 点台という結果が得られた。設問 6 番において、5 人が 4 点以上をつけていた。これらの結果から、提案インタフェースはユーザがサーベイした情報を整理をすることに対して有用性があることが示唆された。しかし、設問 3 番、5 番の結果から、サーベイした情報の整理と、サーベイで得た情報の活用につながる支援ができていない可能性がある。また、設問 4 番の回答理由から、「テキストで他の情報を提示してほしい」という回答が得られた。このことから、テキスト画面において、ユーザの論文整理に必要とされる情報を提示できていなかったことがわかる。これは現状のテキスト画面は、ユーザが入力した論文の書誌情報のみ提示しており、論文間の関係性を考慮していないためであると考えられる。

以上のことから、提案インタフェースの可視化画面は研究分野への理解を深めることにつながったことが確認できた。しかし、サーベイ行為の理解を促す有用性は示されなかった。

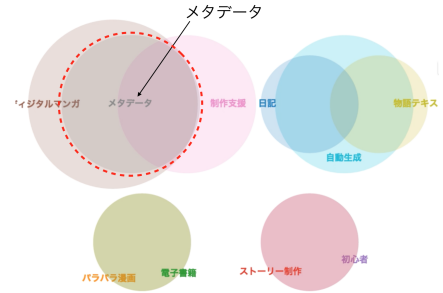


図 6: テーマが覆いかぶさる例

4.4 インタビュー結果と考察

設問 1 において、A 群、B 群全員が「役に立った」と回答し、「客観的に自身がサーベイした情報を見ることができた」という意見が得られた。また、「論文間の関係性がわかりやすかった」という意見が得られた。この回答をした実験参加者から、「抽象度の高いテーマをまとめることができた」という意見が得られた。使いにくかった点として、「他のテーマが注目したテーマに覆いかぶさる状態になった」という意見が得られた。例えば、図 6 の場合、「メタデータ」のみを選択することができない。以上のことから、ユーザが考えたテーマに基づき論文間の関係性を可視化することで、サーベイした論文のテーマの位置付けを明確にする支援ができることが示唆された。しかし、提案手法ではユーザの操作意図を反映することができていないため、インタフェースの改善が必要である。

設問 2 において、B 群 1 人から「役に立った」と回答が得られた。この実験参加者から、「気になるテーマにある著者がよく現れていることがわかった」という回答が得られた。一方、役に立たなかったと回答した 5 人の実験参加者から、「選択したテーマに含まれる著者がどの論文を書いたのかわからない」という回答が得られた。このことから、テキスト画面の情報同士の関係性を表現できていないため、論文情報の関係性を把握することに繋がらなかったと考えられる。

設問 3 において、新たに知見を得た実験参加者全員から「役に立った」という回答が得られた。「可視化された図によって気づけなかったことを見つけた」という意見が得られた。付与できるテーマ数に制限を設けたことで、「付与すべきテーマを考えたことでそのテーマについて整理ができた」という意見が得られた。このことから、自身のサーベイの全体像を把握できることが新たな知見につながったと考えられる。

5 議論

5.1 可視化画面の有用性と問題点

実験結果から、ユーザは自身のサーベイ内容を整理する際、可視化画面を閲覧し、自身のサーベイの全体像を把握する。そこからサーベイが足りていないテーマや、自身の研究分野に関連しているテーマなどの新たな知見につながる事が示唆された。論文を読み直すことなく新たな知見を得ることができたため、サーベイで得た論文情報の整理の負担を軽減できたと言える。

現状の可視化画面は、包括されたテーマを単体で選択できないため、1つのテーマを選択することができない。ユーザ任意のテーマについての関係性のみ注目することなど、ユーザの任意の論文情報の捉え方に対応できていない。この問題を解決するため、画面上のクリックだけでなく、ダブルクリックや、右クリックを用いてユーザ任意のテーマを選択できる機能を実装する必要があると考える。これにより、任意のテーマに基づく情報を提示できるようにする。

5.2 テキスト画面の有用性と問題点

可視化画面とテキスト画面を連動させることで画面遷移の負担がなくなり、論文情報の把握の負担軽減に寄与したと考えられる。しかし、実験結果から、テキスト画面はテーマに含まれる著者の出現頻度や傾向が表れにくい結果となった。これは、入力した情報が少なかったことが原因だと考えられる。そのため、ユーザの整理の支援につながらなかったと考えられる。テキスト画面において、著者の出現頻度や傾向を確認するために、長時間の実験を行い、論文の情報量を確保した状態でテキスト画面の有用性を検証する必要があると考える。また、テキスト画面では、これらの情報はユーザが選択したテーマに基づいて提示しているが、テキスト画面に提示される情報同士(e.g., 著者 A が執筆した論文)の関係性を表現できていない。そのため、テキスト画面の情報提示方法を検討する必要がある。

5.3 情報入力方法の有用性と問題点

インタビューの結果から、付与できるテーマ数に制限を設けることでユーザに論文間の関係性を考慮させるきっかけを与えることができたことがわかる。一方で、ユーザは論文に付与するテーマを絞りきれない問題が出ていた。論文に付与できるテーマ数を増やすことで、この問題を解決できる可能性がある。しかし、テーマを付与できる数に制限を設けない場合、1つの論文に対して多くのテーマが付与され情報が煩雑になるこ

とが危惧される。そのため、論文に付与できる適切なテーマ数を検討する必要がある。

6 おわりに

本研究では、研究初心者におけるサーベイの整理支援による研究分野への理解を目的とし、サーベイした情報における整理のための可視化インタフェースを実装し、その有用性を検証した。実験結果から、論文の関係性を把握させることで研究初心者のサーベイした論文整理の支援につながる事が示唆された。しかし、提案手法では、著者や論文の情報における関係性を考慮できていないため、論文のテーマに関する情報の可視化以外の手法はユーザの論文整理支援につながる事が少なかった。そのため、テーマだけでなくユーザ任意の情報に基づき関係性を提示する方法を検討する。

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] 井坂 徳恭, 中山 泰一: 重要論文検索システム Iask の実装と評価, 情報処理学会研究報告, Vol. 2011-CE-109, No. 10, pp. 1-8 (2011)
- [2] 井上 絢翔, 韓 東力: 参照関係の可視化による論文サーベイの効率化, 第 11 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, SIG-AM-11-09, pp. 55-61 (2015)
- [3] 伊藤 一成, 斎藤 博昭: アノテーションの概念に基づく情報可視化インタフェース, 日本データベース学会論文誌, Vol. 4, No. 1, pp. 169-172 (2005)
- [4] Keller, T., Gerjets, P., Scheiter, K., Garsoffky, B.: Information visualizations for knowledge acquisition: The impact of dimensionality and color coding, *Computers in Human Behavior*, Vol. 22, No. 1, pp. 43-65 (2006)
- [5] Michael, B., Vadim, O., Jeffrey, H.: D Data-Driven Documents, *IEEE transactions on visualization and computer graphics*, Vol. 17, No. 12, pp. 2301-2309 (2011)

密度に基づく時空間分析システムにおける 学習済み深層ネットワークを用いた画像分類

Image Classification using Pre-trained Deep Network for Density-based Spatiotemporal Analysis System

酒井 達弘^{1,2*} 田村 慶一¹ 北上 始¹
Tatsuhiro Sakai^{1,2} Keiichi Tamura¹ Hajime Kitakami¹

¹ 広島市立大学大学院情報科学研究科

¹ Graduate School of Information Sciences, Hiroshima City University

² 日本学術振興会特別研究員 DC1

² JSPS Research Fellow DC1

Abstract: Recently, during natural disasters like, earthquakes, typhoons, flood, and heavy snowfall, people actively post messages that mention situations people are facing through social media sites. Therefore, the enhancement of situation awareness in the real world using social data is one of the most attractive research topics. In our previous work, we developed a density-based spatiotemporal analysis system to identify topic-related areas in which there are a huge number of geo-tagged tweets related to a topic are posted. In this paper, we propose a novel density-based spatiotemporal analysis system with a photo image classifier using the pre-trained deep network in order to enhance situation awareness by showing accurate topic-related photos. The photo image classifier using a support vector machine (SVM) based on the pre-trained deep network is integrated into the conventional density-based spatiotemporal analysis system. To evaluate the proposed system, we used actual tweet data sets related to weather topics, “heavy rain” and “heavy snow,” in Japan. The experimental results showed that the proposed system can classify photo images related to these weather topics more sensitively compared with our previous method.

1 はじめに

近年、インターネット上のユーザはソーシャルメディアサイトを通して気象状況や発生した自然災害に関する内容を投稿するようになってきている。例えば、代表的なソーシャルメディアサイトの Twitter では、大雨、大雪、地震や台風などを目の当たりにした人がテキストメッセージや写真によってそれらの状況を伝えている [1]。通常、このような投稿には位置情報（ジオタグと呼ぶ）が付与されており、これらのジオタグ付きのツイートから気象状況や自然災害などの緊急性のあるトピックを検出し、活用することが期待されている [2]。

そこで我々は、気象状況や自然災害などモニタリングをしたいトピックを設定し、当該トピックの発生を検出し、時空間上での変化を分析するために密度に基

づく時空間分析システムを提案している [3, 4]。密度に基づく時空間分析システムでは、ナイーブベイズ分類器、 (ϵ, τ) -密度に基づく適応的な時空間クラスタリングと Bag-of-Features (BoF) の 3 つの手法を用いている。ナイーブベイズ分類器を用いてトピックに関連するツイートのみを抽出する。 (ϵ, τ) -密度に基づく適応的な時空間クラスタリングを用いることで、トピックに関連するジオタグ付きツイートが時空間上で密集している領域を時空間クラスタとして抽出し、トピックの発生や消滅を監視することができるさらに、BoF を用いて時空間クラスタに含まれる画像データから特徴ベクトルを抽出し、サポートベクターマシーン (SVM) によってトピックに関連する画像データかどうか分類し、トピックに関連する画像データのみを抽出する。また、トピックの時空間上での変化を確認するための Web インタフェースから、抽出された時空間クラスタと画像データの情報を閲覧することができる。

先行研究では、画像データの特徴ベクトル抽出器として BoF を用いていたが、一定の精度でトピックに関

*連絡先：広島市立大学大学院情報科学研究科
〒731-3194 広島市安佐南区大塚東 3-4-1
E-mail: da65003@e.hiroshima-cu.ac.jp

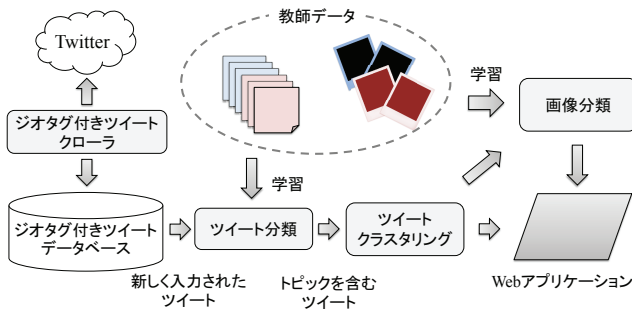


図 1: 密度に基づく時空間分析システムの概要

連する画像データの分類が可能であるものの、分類精度の向上が課題となっていた。そこで本研究では、密度に基づく時空間分析システムにおける画像分類手法として、学習済み深層ネットワークを用いた画像分類を導入する。大規模画像データを用いて学習させた深層ネットワークから抽出した特徴ベクトルは、BoFを用いて抽出した特徴ベクトルよりも汎用性が高く、画像データの特徴を十分に捉えることができ、高精度に分類することができる。

本論文の構成は以下の通りである。第2章では、密度に基づく時空間分析システムについて、簡単に説明する。第3章では、提案手法である学習済み深層ネットワークを用いた画像分類を導入したシステムを説明する。第4章では、評価実験の結果を示し、第5章で本論文をまとめる。

2 密度に基づく時空間分析システム

本章では、密度に基づく時空間分析システムについて簡単に説明する。

2.1 データ定義

Twitter 上から取得したジオタグ付きツイートを gt_i と表記し、その集合を $GTS = \{gt_1, \dots, gt_n\}$ とする。ここで、 gt_i は文書データ $text_i$ 、投稿時間 pt_i 、位置情報 pl_i と画像データ pi_i の4つから構成される。本研究では、位置情報として経度と緯度を用いる。また、モニタリングをしているトピックの内容を含むジオタグ付きツイートを関連ジオタグ付きツイート $rgt_j (= gt_{\phi(j)})$ と呼ぶ。関連ジオタグ付きツイート集合を $RGTS = \{rgt_1, \dots, rgt_m\}$ とすると、 GTS は $RGTS$ を包含しており ($RGTS \subset GTS$)、次の単射で表現される。

$$\phi(j) : RGTS \rightarrow GTS; rgt_j \mapsto gt_{\phi(j)} \quad (1)$$



図 2: Web アプリケーション画面

例えば、“大雨”についてモニタリングをしているとすると、関連ジオタグ付きツイートとは、“すごい大雨”や“強い雨が降り始めた”などの大雨が降っていることを伝えているツイートとなる。

2.2 システム概要

図1に密度に基づく時空間分析システムの概要図を示す。密度に基づく時空間分析システムでは、ツイート分類、ツイートクラスターリングと画像分類を一定時間毎に実行し、Web アプリケーション上へデータを出力する。システム全体の処理の流れを次に示す。

1. ジオタグ付きツイートクローラを用いて Twitter からジオタグ付きツイートを収集し、ジオタグ付きツイートデータベースに保存する。
2. ツイート分類部では、収集したジオタグ付きツイートに対して、ナイーブベイズ分類器を用いてモニタリングをしているトピックに関連する関連ジオタグ付きツイートとそれ以外のツイートに分類を行い、関連ジオタグ付きツイートのみを抽出する。
3. ツイートクラスターリング部では、新たに収集された関連ジオタグ付きツイートとこれまでに抽出された時空間クラスタ集合を入力として、 (ϵ, τ) -密度に基づく適応的な時空間クラスターリング [5] を用いて新しく時空間クラスタ集合を抽出する。新たに時空間クラスターリングが抽出されるとトピックの発生、時空間クラスタが消滅すると時空間ク

ラスタが存在した地域においてトピックが消滅したことを意味する。

4. 画像分類部では、時空間クラスタに含まれる関連ジオタグ付きツイートから画像データを取り出し、BoF[6]を用いて各画像データの特徴ベクトルを抽出する。次に、SVMを用いてモニタリングをしているトピックに関連している関連画像データとそれ以外の画像データに分類を行い、関連画像データのみを抽出する。SVMは事前に教師データを用いて学習済みであるとする。
5. 時空間クラスタの内容と関連画像データをWebアプリケーション上へ出力する。図2にWebアプリケーションのスクリーンショットを示す。Webアプリケーションでは、地図上に時空間クラスタとして抽出されたツイートと関連画像データが表示される。

3 提案システム

本章では、学習済み深層ネットワークを用いた画像分類を導入した新しい密度に基づく時空間分析システムを提案する。

3.1 概要

提案システムでは、先行システムにおける画像分類について、学習済み深層ネットワークを特徴ベクトル抽出器として利用する。先行システムの特徴ベクトル抽出器としてBoFを用いた画像分類は、一定の精度でトピックに関連する画像データの分類が可能であるものの、その精度の向上が課題となっていた。Twitterに投稿される画像データはバリエーションが多く、決まった種類の画像データが存在しないために、BoFでは画像データの特徴を十分に捉えることができない。提案システムでは、学習済みの畳み込みニューラルネットワーク(CNN)を画像データの特徴ベクトル抽出器として用いる。

大規模データセットを用いて学習させたCNNの中間層から抽出される特徴ベクトルは非常に汎用性が高く、様々なドメインで再利用可能であることが示されている。教師データとなる画像データ集合について学習済みの畳み込みニューラルネットワークから特徴ベクトルを抽出する。次に、SVMを用いて再学習し、分類器を構成する。

表 1: VGG-16 の構造

層名	ユニット数
入力層	150528
第 1 層 (畳み込み層)	64
第 2 層 (畳み込み層)	64
プーリング層	
第 3 層 (畳み込み層)	128
第 4 層 (畳み込み層)	128
プーリング層	
第 5 層 (畳み込み層)	256
第 6 層 (畳み込み層)	256
第 7 層 (畳み込み層)	256
プーリング層	
第 8 層 (畳み込み層)	512
第 9 層 (畳み込み層)	512
第 10 層 (畳み込み層)	512
プーリング層	
第 11 層 (畳み込み層)	512
第 12 層 (畳み込み層)	512
第 13 層 (畳み込み層)	512
プーリング層	
第 14 層 (全結合層)	4096
第 15 層 (全結合層)	4096
第 16 層 (出力層)	1000

3.2 学習済み深層ネットワークを用いた特徴ベクトル抽出

提案システムで用いるCNNは、数多く提案されている深層ネットワークの中でも、特に画像認識の分野に応用されているニューラルネットワークである。CNNは中間層に畳み込み層とプーリング層が存在し、画像データの局所的な特徴を自動的に学習することができる。

提案システムでは、特徴ベクトル抽出器としてVGG-16と呼ばれるCNNのネットワークモデルを用いる[7]。VGG-16の学習は、大規模画像認識コンペティションのILSVRC-2014で提供されたImageNetの1000分類、120万枚の画像を用いて行われている。ImageNetは一般的な内容の画像データを含んでおり、VGG-16は画像データの様々な分析に応用可能な汎用知識を学習できていると言われている。

VGG-16の具体的な構造を表1に示す。VGG-16は16層から形成される。畳み込み層では、畳み込み処理を行っており、入力に対して重みフィルタの内積を計算する。各畳み込み層は前層の入力に対して畳み込み処理を行い、次の層の入力となる特徴マップを出力する。プーリング層では、畳み込み層から出力された特徴マップを縮小する。VGG-16では、最大値プーリン

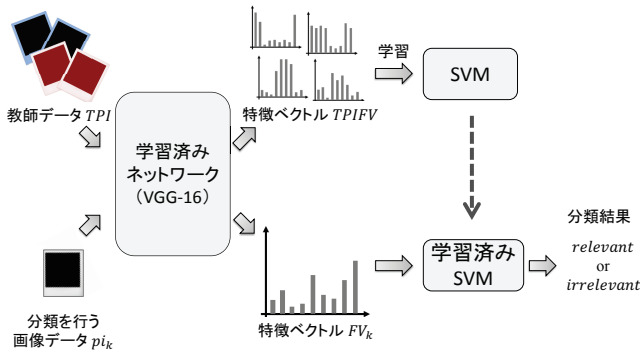


図 3: 提案システムの画像分類の処理手順

グを用いている。全結合層では、重み付き結合を計算し、活性化関数によりユニットの値を求める。VGG-16 では活性化関数として、ReLU を用いている。

提案システムでは、第 15 層の全結合層から特徴ベクトルを抽出するため、各画像データについて 4096 次元の特徴ベクトルが抽出される。画像データ i の特徴ベクトルを、 $FV_i = \{fv_{i,1}, fv_{i,2}, \dots, fv_{i,4096}\}$ とする。VGG-16 の学習に用いられている ImageNet には気象や自然災害に関する分類を含む画像データは無い。しかしながら、出力層手前の中間層には画像データの汎用的な特徴が表れるため、ソーシャルメディア上に投稿される画像データを区別する特徴ベクトルとして利用できると考える。

3.3 処理手順

提案システムは学習済み深層ネットワークを用いて抽出した特徴ベクトルと SVM を用いて画像分類を行う。教師データを $TPI = \{(tpi_1, tc_1), (tpi_2, tc_2), \dots, (tpi_l, tc_l)\}$ とし、 $tc_i = \{relevant, irrelevant\} \in CLASS$ とする。 $relevant$ はモニタリングをしているトピックに関連している画像データのクラスとし、 $irrelevant$ はそれ以外の画像データのクラスである。提案システムの画像分類の処理手順を図 3 に示す。処理手順は次の通りである。

1. 教師データ TPI の各画像データから学習済み深層ネットワークを用いて特徴ベクトル $TPI_FV = \{FV_1, FV_2, \dots, FV_n\}$ を抽出する。
2. TPI_FV を用いて SVM を学習させる。
3. 分類を行う画像データ p_k から学習済み深層ネットワークを用いて特徴ベクトル FV_k を抽出し、SVM で分類を行う。 $relevant$ クラスに分類された画像データを関連画像データとする。

4 評価実験

提案手法の有効性を確認するために、評価実験を行った。本章では、評価実験の結果を示す。

4.1 実験内容

評価実験では、モニタリングをするトピックを“大雨”と“大雪”としてそれぞれ評価を行う。比較手法としては、画像データの特徴ベクトル抽出器として、先行システムで用いられていた BoF を用いた手法 (BoF と表記する)、VGG-16 と同じ構造をした 16 層の CNN を構築し教師データを用いて学習させたネットワークを用いた手法 (CNN-16 と表記する)、提案手法である VGG-16 のネットワークを用いた手法 (VGG-16 と表記する) とで比較を行う。BoF によって抽出する特徴ベクトルの次元数は 4096 と設定した。

評価方法としては、最初に作成した教師データについて交差検定によって評価を行う。次に、実際に密度に基づく時空間分析システムによって抽出された時空間クラスタから画像データを取り出し、画像分類を行った結果を評価する。教師データ TPI は、トピック“大雨”については 2014 年 7 月に投稿された $relevant$ クラス 500 件と $irrelevant$ クラス 500 件を用いた。また、トピック“大雪”については 2014 年 1 月と 2 月に投稿された $relevant$ クラス 500 件と $irrelevant$ クラス 500 件を用いた。

4.2 交差検定

最初に交差検定による評価を行う。交差検定の分割数は 2, 4, 6, 8 と 10 分割を用いた。モニタリングをしているトピックを“大雨”としたときの交差検定の結果を図 4 に、モニタリングをしているトピックを“大雪”としたときの交差検定の結果を図 5 に示す。図 4 と図 5 には、正解率、精度と再現率をそれぞれ示している。図 4 と図 5 より、正解率、精度と再現率の全てについて、VGG-16 は BoF と CNN-16 よりも高性能であることを示した。

4.3 システム評価

次に、実際に抽出された時空間クラスタに含まれているジオタグ付きツイートに付与されている画像データを分類し、評価を行う。実験期間としては、トピック“大雨”については、日本で台風が観測され全国各地で大雨となった 2014 年 8 月 1 日から 10 日としている。トピック“大雪”については、日本全国各地で降雪のあった 2014 年 2 月 10 日から 2 月 18 日とする。

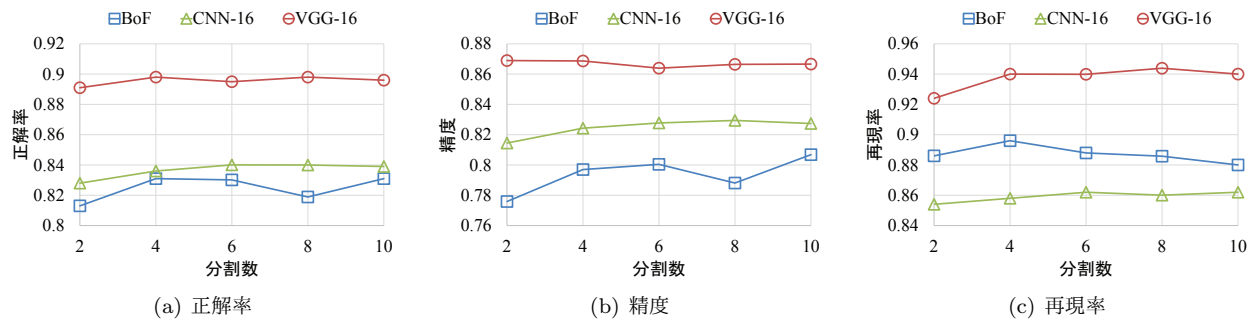


図 4: トピック“大雨”の交差検定の結果

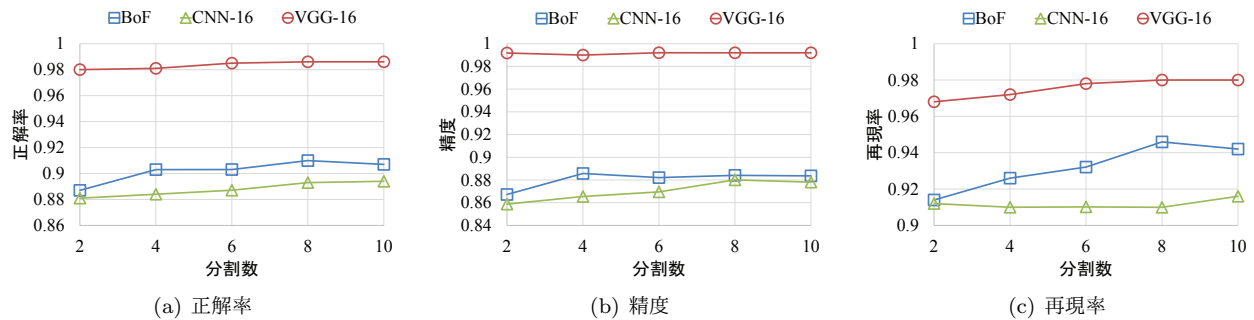


図 5: トピック“大雪”の交差検定の結果

モニタリングをしているトピックを“大雨”としたときの実験結果を表2に、モニタリングをしているトピックを“大雪”としたときの実験結果を表3に示す。表2と表3には、各日付の時空間クラスに含まれていた画像データ数、正解率、精度と再現率をそれぞれ示している。表2より、トピック“大雨”について正解率を見ると、VGG-16は10日間で最も高い正解率を示している。精度と再現率についてもVGG-16が最も良い結果を示している。また、表3より、トピック“大雪”についても同様に、VGG-16が最も良い結果が得られた。図6と図7に、BoFでは正しく *relevant* クラスに分類されなかったが、VGG-16では正しく *relevant* クラスへ分類することができた画像データの例を示す。VGG-16によって正しく *relevant* クラスへ分類された画像データには、モニタリングをしているトピック以外の人物や物体が写っているものが多いことが分かる。VGG-16を用いて抽出された特徴ベクトルは、モニタリングをしているトピック以外のものが写っていたとしても、画像データ中の当該トピックの部分の特徴を捉えることができたと言える。以上の実験結果より、大規模画像データによって学習済み深層ネットワークを特徴ベクトル抽出器として用いることの有効性を示すことができた。

5 まとめ

本研究では、画像分類として学習済み深層ネットワークを用いた画像分類を導入した新しい密度に基づく時空間分析システムを提案した。大規模画像データを用いて学習させた深層ネットワークから抽出した特徴ベクトルは、BoFを用いて抽出した特徴ベクトルよりも汎用性が高く、画像データの特徴を十分に捉えることができ、高精度に分類することができる。評価実験より、提案手法は先行研究の手法よりも高性能に画像分類ができることを示した。今後の課題としては、学習済み深層ネットワークを再学習させて新しいモデルを作成することがあげられる。学習済みの深層ネットワークを初期値とし再学習することで、汎用性があり、さらに特定の内容に最適な深層ネットワークができることを期待できる。

謝辞

本研究の一部は、JSPS 科研費 JP16J05403, JP26330139 と総務省 SCOPE (受付番号:162308002) の支援により行われた。

表 2: トピック “大雨” のシステム評価

日付	画像データ数	正解率			精度			再現率		
		BoF	CNN-16	VGG-16	BoF	CNN-16	VGG-16	BoF	CNN-16	VGG-16
8/1	130	0.68	0.73	<u>0.85</u>	0.59	0.66	<u>0.75</u>	0.76	0.76	<u>0.95</u>
8/2	217	0.76	0.71	<u>0.83</u>	0.54	0.49	<u>0.63</u>	<u>0.97</u>	0.82	0.92
8/3	203	0.75	0.78	<u>0.82</u>	0.60	0.67	<u>0.69</u>	<u>0.86</u>	0.73	<u>0.86</u>
8/4	63	<u>0.83</u>	0.67	0.81	<u>0.57</u>	0.37	0.54	<u>0.93</u>	0.71	<u>0.93</u>
8/5	63	0.67	0.70	<u>0.78</u>	0.43	0.45	<u>0.53</u>	0.94	0.88	<u>1.00</u>
8/6	87	0.61	0.63	<u>0.64</u>	0.34	0.36	<u>0.38</u>	0.84	0.89	<u>0.95</u>
8/7	53	0.58	0.58	<u>0.70</u>	0.39	0.37	<u>0.47</u>	<u>1.00</u>	0.79	<u>1.00</u>
8/8	230	0.62	0.71	<u>0.74</u>	0.34	0.41	<u>0.44</u>	0.82	0.80	<u>0.90</u>
8/9	412	0.70	0.70	<u>0.77</u>	0.39	0.38	<u>0.46</u>	<u>0.90</u>	0.82	0.87
8/10	572	0.73	0.74	<u>0.81</u>	0.56	0.58	<u>0.66</u>	0.88	0.81	<u>0.90</u>

表 3: トピック “大雪” のシステム評価

日付	画像データ数	正解率			精度			再現率		
		BoF	CNN-16	VGG-16	BoF	CNN-16	VGG-16	BoF	CNN-16	VGG-16
2/10	210	0.68	0.68	<u>0.80</u>	0.74	0.75	<u>0.83</u>	0.73	0.71	<u>0.84</u>
2/11	206	0.70	0.71	<u>0.81</u>	0.74	0.76	<u>0.81</u>	0.82	0.80	<u>0.92</u>
2/12	23	0.65	<u>0.70</u>	<u>0.70</u>	0.29	<u>0.38</u>	<u>0.38</u>	0.40	<u>0.60</u>	<u>0.60</u>
2/13	87	0.67	0.68	<u>0.77</u>	0.61	0.63	<u>0.71</u>	0.75	0.73	<u>0.85</u>
2/14	2196	0.78	0.77	<u>0.87</u>	0.82	0.84	<u>0.89</u>	0.86	0.80	<u>0.93</u>
2/15	940	0.75	0.81	<u>0.90</u>	0.86	0.90	<u>0.93</u>	0.79	0.84	<u>0.93</u>
2/16	306	0.69	0.71	<u>0.85</u>	0.72	0.76	<u>0.82</u>	0.80	0.76	<u>0.96</u>
2/17	152	0.68	0.69	<u>0.80</u>	0.66	0.71	<u>0.73</u>	0.77	0.68	<u>0.95</u>
2/18	58	0.57	<u>0.66</u>	<u>0.66</u>	0.44	<u>0.52</u>	0.51	0.71	0.62	<u>0.90</u>



図 6: VGG-16 によって正しく *relevant* クラスへ分類された画像データの例 (トピック “大雨”)



図 7: VGG-16 によって正しく *relevant* クラスへ分類された画像データの例 (トピック “大雪”)

参考文献

- [1] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the First Workshop on Social Media Analytics, SOMA '10*, pp. 71–79, 2010.
- [2] Mor Naaman. Geographic information from geo-referenced social media data. *SIGSPATIAL Special*, Vol. 3, No. 2, pp. 54–61, 2011.
- [3] Tatsuhiro Sakai, Keiichi Tamura, and Hajime Kitakami. Density-based spatiotemporal analysis system with photo image classifier using the BoF model. *Information Engineering Express*, Vol. 1, No. 4, 2015.
- [4] Tatsuhiro Sakai and Keiichi Tamura. Real-time analysis application for identifying bursty local areas related to emergency topics. *SpringerPlus*, Vol. 4, No. 162, 2015.
- [5] Tatsuhiro Sakai, Keiichi Tamura, and Hajime Kitakami. Emergency situation awareness during natural disasters using density-based adaptive spatiotemporal clustering. In *Database Systems for Advanced Applications, DASFAA 2015 International Workshops, SeCoP, BDMS, and Posters, Hanoi, Vietnam, April 20-23, 2015*, Vol. 9052, pp. 155–169, 2015.
- [6] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cedric Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 1–22, 2004.
- [7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, arXiv:1409.1556, 2014.

知識伝達における類推による説明能力の育成

Training of Explanation Ability in Knowledge Transfer by using Analogical Thinking

石田純太¹ 砂山渡² 川本佳代¹
Junta Ishida¹ Wataru Sunayama² Kayo Kawamoto¹

¹ 広島市立大学大学院 情報科学研究科

¹ Graduate School of Information Sciences, Hiroshima City University

² 滋賀県立大学 工学部

² School of Engineering, The University of Shiga Prefecture

Abstract: Analogy can be effective means to make things easy to understand to others, as it can be seen well by comparing things do not know in known things. However, opportunities and methods of training of explanation ability by using analogical thinking do not exist generally. In this study, for the purpose of training of explanation ability in knowledge transfer by using analogical thinking, we propose a system to encourage understanding of the analogy process and the practice due to repeated. We evaluated to verify the effect of the proposed system.

1 序論

近年の変化の激しい社会を生き抜いていくために、「生きる力」を育むことが求められている。そのために、文部科学省は、「類推を用いて、説明し伝え合う活動」を重視すべき活動の1つに挙げている [1]。類推とは、以前に経験した事柄（ベースと呼ぶ）を現在直面している事柄、あるいは問題（ターゲットと呼ぶ）に当てはめることをいう [2]。

実際に、我々が新たな知識の理解を試みる際には、既知の知識の特徴を新たな知識に当てはめて理解する類推を行っている。例えば、中学校の理科の授業では、目に見えない電圧と電流、抵抗について学ぶ際、教師が電気回路を水流モデルに喩えて説明を行うことで、生徒は理解を深めることができる。このように類推は、わからない物事を何か既知のもので喩えるとよくわかるように、他人に対して物事の理解しやすい説明をするために有効な手段になり得る。しかし、類推による説明能力を育成する機会や方法は一般に存在しない。

そこで本研究では、自身が知っている知識を、その知識を知らない他人に伝える際に、類推を用いてわかりやすく説明する能力を育成するシステムを提案する。具体的には、類推のプロセスの理解を促し、その各プロセスの練習を繰り返し行うことができるシステムを構築する。このシステムにより、ユーザがコミュニケーションにおける説明能力を、独自に身につけられるようにすることを目指す。

また類推による他人への説明能力を身につけることができれば、自身が新しい知識を学ぶ際にも、類推を積極的に活用して理解を深めることや、類推を活用した新たな科学的発見が期待できる [3]。

以下本稿では、2章で関連研究について述べ、3章で

本研究における類推の定義を与える。4章で提案する類推による説明能力の育成システムの構成と詳細について、5章で構築したシステムの効果を検証する評価実験について述べ、6章で本稿を締めくくる。

2 関連研究

本章では、類推の活用と類推能力の育成、ならびに説明能力の育成に関わる研究について述べる。

2.1 類推による知識発見

解決方法がわからない問題に対して、創造的発想によって解決を目指す人の着想段階を、糸口となる知識を豊富に含む情報データベースの検索結果を類推に用いて支援した研究 [4] がある。また、ルール発想推論に、類推公理を加えることによって、ルールの生成と新述語の発見、類比の補完のすべてを統一的に実現させ、新しいスキルの発見を目指した研究 [5] がある。

これらの研究は、類推による知識発見を目指したものとなっており、類推を活用して人間を支援する点で本研究と類似している。しかし本研究では、基本的な類推プロセスの理解とその適用を促し、人が類推を行う能力を身につけられる環境の構築を目指す。

2.2 類推能力の育成

類推的思考に着目した問題解決力の育成を図った研究 [6] がある。この研究では、練習により未知の課題への類推の適用を促すことで、問題解決を支援している。

本研究とは、類推能力の育成という点で類似しているが、類推の適用による問題解決は、類推のプロセスの理解を前提とし、また問題解決のための選択肢は必ずしも多くならないため、即座に類推の適用が困難な人に対しての導入教育には、必ずしも向いていない。

本研究では、より基本的な類推のプロセスの理解を促し、より自由度が高い言葉による説明において類推を用いる支援を行う。

2.3 説明能力の育成

説明文のわかりやすさの要因は、書き手、読み手、文章表現の3つに分けられる。このうち、書き手の要因に着目した研究があり、わかりやすい説明文となるためには、読み手の知識状態に応じて説明文を書くこと必要なことが明らかになっている[7]。

すなわち、類推による説明を行う際には、説明の聞き手が知っているものをベースに選択する必要がある。その際、多くの人が知っているものをベースに選択することができれば、説明の聞き手に応じて説明を変える必要がなくなる。また様々な候補の中からベースを選択する練習を繰り返すことで、聞き手に応じた説明能力の獲得にも繋がることが期待できる。

3 知識伝達における類推

本章では、本研究で取り扱う知識伝達における類推の定義と類推の対象について述べる。

3.1 本研究における類推の定義

類推は「以前に経験した事柄（ベースと呼ぶ）を現在直面している事柄、あるいは問題（ターゲットと呼ぶ）に当てはめること」と定義され、一般的に次のプロセスで実現される[2]。

- P1. ターゲット問題の表象：問題が与えられ、それを何かの形で表象すること
- P2. ベースの検索：関連する過去の経験を長期記憶から検索すること
- P3. 写像：検索されたものの中から重要な事項を現在の問題に当てはめてみる
- P4. 正当化：写像が適切であると正当化すること
- P5. 学習：類推の結果をストックし、それを一般化すること

すなわち、P1のターゲット問題の表象で、現在直面している問題（ターゲット）の特徴を捉え、P2でその特徴に類似する過去の経験（ベース）を探し出す。P3でターゲットとベースの対応関係を見出し、P4でその対応関係の解釈を与えて、対応の正しさを確認する。最後にP5で、適用した類推を記憶にとどめ、将来の別の類推に活用できるように記憶する。

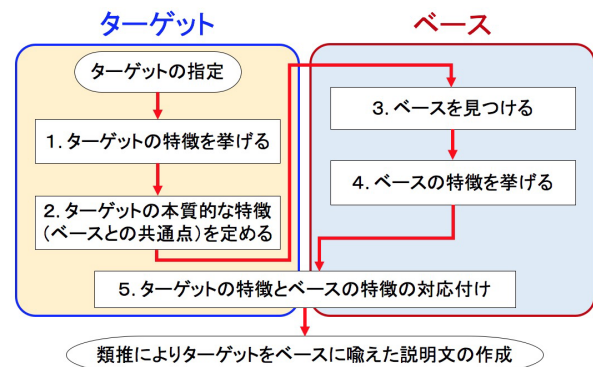


図 1: 説明文作成における類推手順

本研究においては、名詞1単語で表される自分の知っている事柄を、それを知らない他人に伝える知識伝達の際に、類推を用いて説明する練習を行う。先の類推手順に準じて用意した、説明文作成における類推手順を図1に示す。

すなわち、P1「ターゲット問題の表象」に対応して、手順1で説明したい事柄（ターゲット）の特徴を列挙する。次にP2「ベースの検索」に対応して、手順2で列挙した特徴の中から本質的な特徴を選定し、手順3で手順2で挙げた本質的な特徴を持つベースを見つける。続いてP3「写像」に対応して、手順4でベースの特徴を列挙し、手順5でターゲットの特徴との対応付けを行う。P4「正当化」は、先の手順5において、対応の適切さを確認しながら対応付けを行うとともに、この結果に従って、説明文を作成する際に再度確認する。P5「学習」は、説明文の作成によって一つの類推が完成するため、この経験が後に生かされることを期待する。

本研究においては、構築するシステムのユーザに、本節で述べた類推手順の理解を促し、各プロセスの練習を通じて、説明能力を身につけてもらう。

3.2 本研究が対象とする類推

本研究の説明文作成における類推では、他人に説明する対象となる、自分がよく知っている知識を「ターゲット」、類推の喩え先となる、自分と他人の両者がよく知っている知識を「ベース」と定義する。

このベースとターゲットの選び方によって、類推の難易度が変わってくる。本研究では、類推の考え方を初めて学ぶ人を対象に含むため、比較的類推が容易なベースとターゲットの関係を用意する。類推は、ベースとターゲットの類似性に着目して行われるため、ベースとターゲットに何らかの類似性が見出せる場合に考えやすくなる。また、ベースとターゲットの特徴に着目して対応付けを行うため、ベースやターゲットについての情報や知識が豊富なほど、対応付けが行いやすくなる。

そこで、本研究が対象とする類推における、ターゲットとベースの関係を、次の4つとした。

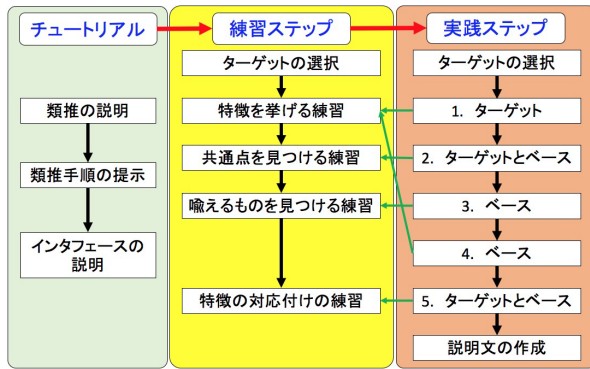


図 2: 類推による説明能力育成システムの構成と学習手順（赤と黒の矢印が学習手順を表す）

- C1. 同カテゴリ：日本語ソーラス [8] 上において，共通の親をもつ単語同士。例えば「哺乳類」を共通の親に持つ，「犬」と「猫」。
- C2. 並列カテゴリ：日本語ソーラス上において，親の親が共通となる単語同士。例えば，それぞれ「哺乳類」「爬虫類」を経て共通の親「脊椎動物」をもつ，「犬」と「蛇」。
- C3. 擬人化，逆擬人化：ものを人に喩える，または人をものに喩える。例えば，「人」を「家電製品」に喩える。あるいは「戦艦」を「人」に喩える。
- C4. 抽象カテゴリ：日本語ソーラスによらず，抽象的な言葉による共通点をもつ単語同士。例えば，「ものづくり」という共通点をもつ，「文章」と「料理」。

すなわち，C1,C2,C4は類似性が想定されるベースとターゲットの組み合わせとして，C3は，ベースあるいはターゲットを人とする事で，特徴の対応付けが行いやすい組み合わせとして用意した。

4 類推による説明能力の育成システム

本章では，提案する類推による説明能力の育成システムについて述べる。

図2に類推による説明能力育成システムの構成と学習手順を示す。システムは，チュートリアル，練習ステップ，実践ステップの3つから構成される。チュートリアルでは，類推の具体例を元に類推の手順についての説明を行う。練習ステップでは，類推手順における各ステップを，繰り返し練習可能な環境を提供する。実践ステップでは，類推手順を通して試すことができる。以下で各構成の詳細について述べる。

4.1 チュートリアル

チュートリアルは，「類推の説明」「類推手順の提示」「インタフェースの説明」の3つの要素からなる。システムを初めて利用するユーザは，最初にチュートリアルを実施して，類推の概要とシステムの使い方を学ぶ。

「類推の説明」では，類推の定義と本システムの狙い，類推の手順について説明を与える。説明は，システムの画面上にメッセージとして表示し，「OK」ボタンを押すことで次のメッセージに進むことができる。

「類推手順の提示」と「インタフェースの説明」は，類推の各手順ごとに，手順の説明とシステムの操作を繰り返しながら進める形とする。これにより，類推の進め方を視覚的に説明するとともに，システムの操作方法の習得を促す。各手順におけるシステムの詳細や操作方法は，次節で述べる。

このチュートリアルでは，電気回路を川に喩える類推をもとに説明を行う。これは，中学校の理科の授業で一般的によく使われる類推事例となっており，類推になじみのないユーザでも，類推について理解しやすいと考えたことによる。すなわち，ターゲットを電気回路，ベースを川として，どのように電気回路を川に喩えて説明するか，その一連の手順の説明を，システム上の操作を伴いながら確認できるようにする。

4.2 練習ステップ

練習ステップは，類推手順の各プロセスごとに練習するための環境とする。これは，類推における各プロセスの理解を促すためには，全ての手順を通して繰り返すだけでなく，各プロセスごとの要領を実践的に掴んでもらうことが重要と考えたことによる。またステップごとに練習できる環境は，苦手なプロセスを重点的に練習したいユーザにも対応できる。

練習ステップは，表1に示す，4つのステップと4つのレベルから構成される。各ステップは，それぞれ図1の5つの類推手順に対応する（ターゲットとベースのそれぞれの特徴を挙げる手順は，1つのステップにまとめられている）。またレベルは類推の難易度を表す指標として，3.2章で述べたC1からC4の4つのベースとターゲットの関係に基づく類推を表す。この難易度は，ベースとターゲットの間の共通点の見つけにくさを考慮して設定した。

ユーザは表1の上のステップから順に進めることができ，各ステップ内では，レベルの低いものから練習することができる。1つのレベルを5回終えることで，次のレベルに進むことができ，1つのステップ内のすべてのレベルをクリアすると，次のステップに進むことができる。

4.2.1 ターゲットの選択

システムには，ベースとターゲットの関係を表すカテゴリごとに，ターゲットとなる名詞が用意されている。用意したターゲットのカテゴリとターゲット数を表2に示す。

表 1: 練習ステップの構成

ステップ	レベル 1	レベル 2	レベル 3	レベル 4
特徴を挙げる練習	同カテゴリ	並列カテゴリ	擬人化, 逆擬人化	抽象カテゴリ
共通点を見つける練習	同カテゴリ	並列カテゴリ	擬人化, 逆擬人化	抽象カテゴリ
喩えるものを見つける練習	同カテゴリ	並列カテゴリ	擬人化, 逆擬人化	抽象カテゴリ
特徴の対応付けの練習	同カテゴリ	並列カテゴリ	擬人化, 逆擬人化	抽象カテゴリ

表 2: 用意したターゲットのカテゴリとターゲット数

カテゴリ名 (ターゲット数)	カテゴリの例
同カテゴリ (325 個)	哺乳類, 和食, 球技, 文房具, 情報機器 乗り物, 楽器, ゲームなど 30 種類
並列カテゴリ (223 個)	動物, 植物, 食べ物, 飲み物, スポーツ 施設, 家庭の道具の 7 種類
擬人化 逆擬人化 (「人」193 個, 「もの」189 個)	「人」として俳優, 偉人, スポーツ選手 歌手, アニメキャラクターなど 10 種類 「もの」として, 動物, 植物, 家電 食べ物, スポーツ, 情報機器など 8 種類
抽象カテゴリ (187 個)	モノづくり, 回る, 守ってくれる, 娯楽 一度きり, 止まったら困るなど 29 種類

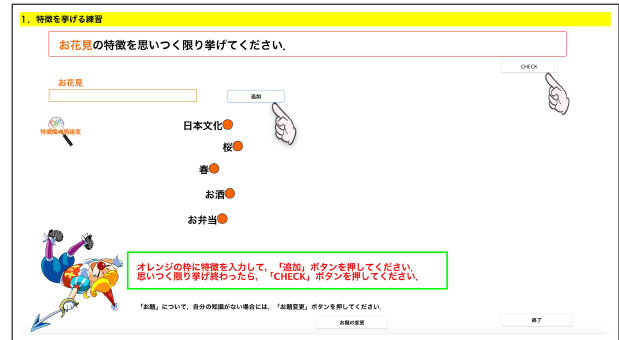


図 3: 特徴を挙げる練習ステップのインタフェース

ユーザには練習の開始前に、出題の対象となるカテゴリ（複数選択可能）をあらかじめ選択してもらうことで、練習時には、選択されているカテゴリの中からターゲットが自動的に出題される。本研究では、他人への説明を行うことを意図し、ターゲットはユーザにとって既知のものになることを想定している。そのため、あらかじめよく知らないカテゴリを除いてもらった上で、選択しているカテゴリについても、出題されたターゲットについてよく知らない時には、「お題変更」ボタンを用いてターゲットの変更を行ってもらう。

4.2.2 特徴を挙げる練習ステップ

特徴を挙げる練習ステップでは、ターゲットやベースの説明に欠かせない特徴を挙げる練習を行う。これはターゲットを説明するためには、ターゲット（説明したいもの）とベース（喩えるもの）について、説明に必要な十分な特徴を挙げるができなければ、わかりやすい説明ができないと考えられることによる。

本ステップでは、与えられたお題（ターゲットにもベースにもなり得る）に対して、思いつく限り特徴を挙げてもらう。本ステップのインタフェースを図 3 に示す。ユーザには、図 3 上部に提示されるお題の特徴を、単語としてテキストエリアに入力して特徴を挙げてもらう。

練習を終える条件は、説明に必要と考えられる特徴が 4 つ以上挙げられた状態で、右上にある「CHECK」ボタンを押すとクリアとなる。

説明に必要と考えられる特徴の判定には、入力された単語を用いた検索エンジンのヒット件数を用いた、式 (1) を用いて行う。ただし式中の F は入力された特徴を表す単語、 T はターゲット（お題）を表す単語、 $Hit(X)$

は X を検索語とした時の検索エンジンのヒット件数を表す。

$$Value(F) = \frac{Hit(F \wedge T)}{Hit(F)} + \frac{Hit(F \wedge T)}{Hit(T)} \quad (1)$$

式 (1) において、右辺の第 1 項は「ターゲットの主要な特徴となっているか」を表す特徴の十分性として、第 2 項は「ターゲットを想起可能な特徴となっているか」を表す特徴の必要性として用いている。これらの各項は、それぞれが条件付き確率を表すため、各項の値の取りうる範囲は 0 から 1 となる。

実験的に特徴になり得る単語の $Value(F)$ の値を調べた結果をもとに、第 1 項と第 2 項の両者が 0.2 未満のときには、「特徴ではない可能性がある」と判定する。またユーザの目安となるように、「特徴ではない可能性がある」以外に、 $Value(F)$ が 0.4 未満と 0.4 以上の場合について、「特徴と思われる」「特徴として欠かせない」という評価を設け、それぞれの結果を図 4 のセリフでフィードバックする。

4.2.3 共通点を見つける練習ステップ

共通点を見つける練習ステップでは、ベースを見つけるための手がかりとなる、ターゲットとベースの共通点を見つける練習を行う。これは、適当なベースを探すためには、ターゲットやベースの本質的な特徴間に類似性を見いだすことができなければ、うまくターゲットをベースで喩えることができないと考えられることによる。

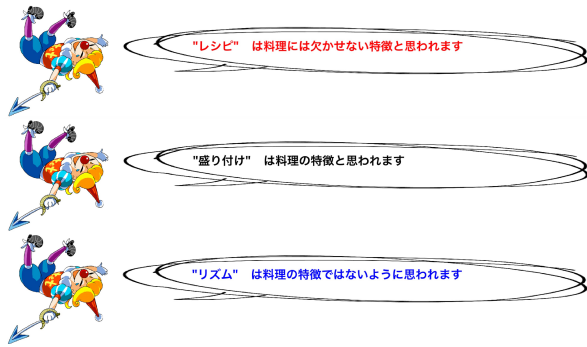


図 4: 特徴の評価のフィードバック

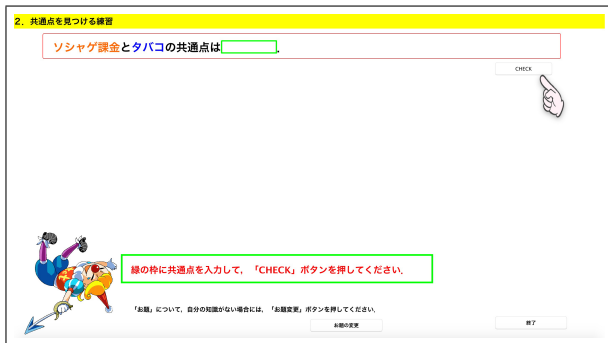


図 5: 共通点を見つける練習ステップのインターフェース

本ステップでは、提示されるターゲットとベースの共通点を答えてもらう。本ステップのインターフェースを図5に示す。ユーザは、図5の上部に示されるターゲットとベースの共通点を、単語としてテキストエリアに入力してもらう。

共通点の入力後、右上にある「CHECK」ボタンを押すと、入力された共通点の適切さの判定がなされ、以下の条件を満たして入ればクリアとなる。すなわち共通点の評価は、入力された共通点をターゲットとベースの共通の特徴として、ターゲットとベースのそれぞれについて、4.2.2 節で述べた式 (1) による特徴の適切さの評価を求め、両者ともに「特徴ではない可能性がある」と判定されなければクリアとする。

4.2.4 喩えるものを見つける練習ステップ

喩えるものを見つける練習ステップでは、ターゲットの喩え先となるベースを見つける練習を行う。これは、数多くのものの中から、ターゲットを喩えるために適切なものを探す能力が類推には必要と考えられることによる。

本ステップでは、提示されるターゲットと、喩える際の観点をもとに、ベースとターゲットとベースの共通点を答えてもらう。本ステップのインターフェースを図6に示す。ユーザは、図6の上部に示されるターゲットと観点から、ベースを単語としてテキストエリアに

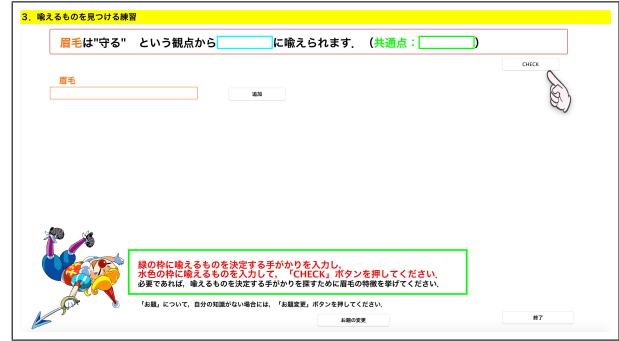


図 6: 喩えるものを見つける練習ステップのインターフェース

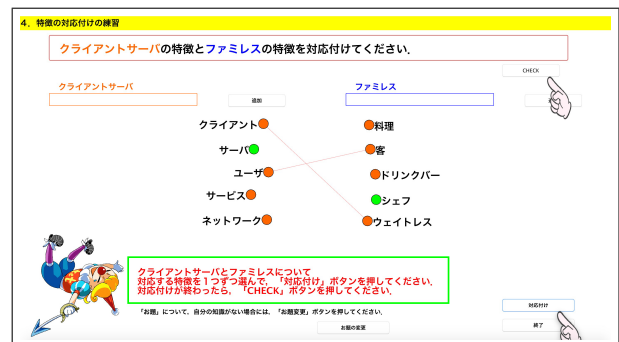


図 7: 特徴の対応付けの練習ステップのインターフェース

入力してもらう。ターゲットとベースの共通点は、観点と同一になる可能性もあるが、ベースを定めた際に、より具体的な共通点が生じる可能性を考慮して、共通点を入力するテキストエリアを設けている。

ベースと共通点の入力後、右上にある「CHECK」ボタンを押すと、入力されたベースの適切さの判定がなされ、以下の条件を満たして入ればクリアとなる。すなわち、共通点をターゲットとベースの共通の特徴として、ターゲットとベースのそれぞれについて、4.2.2 節で述べた式 (1) による特徴の適切さの評価を求め、両者ともに「特徴ではない可能性がある」と判定されなければクリアとする。

4.2.5 特徴の対応付けの練習ステップ

特徴の対応付けの練習ステップでは、ターゲットの特徴とベースの特徴がどのように対応しているかを考え、対応付けを行ってもらう。これは、ターゲットとベースとの間で、できるだけ多くの特徴間で対応付けを行うことができれば、ターゲットの多くの要素について、類推を用いて説明できると考えられることによる。

本ステップでは、ターゲットとベースの特徴間で対応する特徴同士を結びつけてもらう。本ステップのインターフェースを図7に示す。ユーザは、図7に提示されるターゲットとベース、ならびにそれぞれの5個以上の特徴について、対応があると思われる特徴のノー

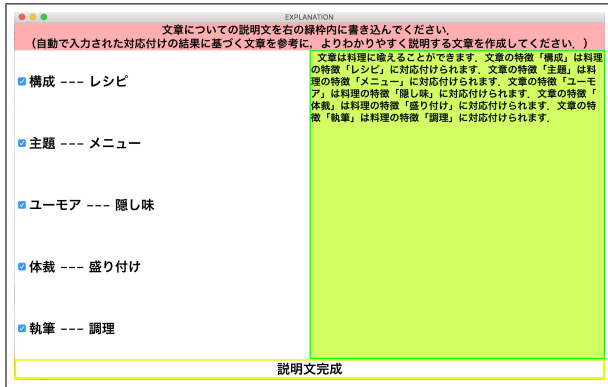


図 8: 説明文作成インタフェース

ドをマウスで選択し、右下の「対応付け」ボタンを押すことで、特徴間にリンクを貼ってもらう。

表示するターゲットとベースの特徴は、ターゲットについて書かれた Wikipedia[9] から特徴となる単語を取り出した上で、Word2Vec[10] を用いてベース側の対応する特徴の候補となる単語を取り出し、人手でターゲットに対応する特徴を選択して用意した。これら、正解となる特徴の組み合わせを 4 から 7 ペアの範囲で用意し、ダミーとなる特徴を 1 個から 2 個追加した上で表示する。

ユーザは、用意した正解となる特徴のペアのうちの 8 割以上の対応付けを正しく行った状態で、右上にある「CHECK」ボタンを押すとクリアとなる。

4.3 実践ステップ

実践ステップでは、図 2 右に示すように、すべての類推手順を通して行ってもらう、最後にターゲットの説明文を作成してもらう。これにより、類推の手順を覚えてもらうとともに、実際に類推による説明ができるようになるための練習を行う。

ターゲットとベースの対応付けが終わった後、対応付けの結果をもとに、ターゲットの説明文を作成してもらう。図 8 に説明文作成のためのインタフェースを示す。図 8 の左側には、対応付けの結果が表示され、右側には対応付けの内容が文章で書かれている。ユーザは右側のパネルの文章を編集することができ、各対応付けについて、より詳細でわかりやすい説明文を作成する。

5 類推による説明能力の育成システムの評価実験

本章では、提案する類推による説明能力の育成システムが、ユーザの説明能力の育成に効果を発揮するかを検証した実験について述べる。

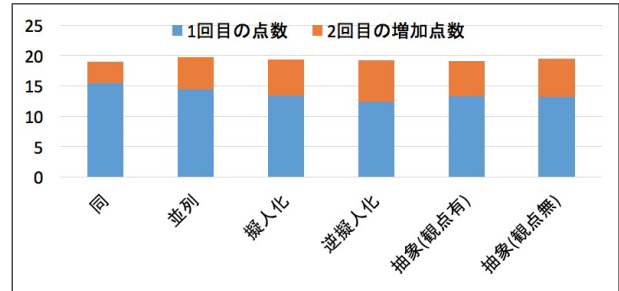


図 9: 相互評価による説明文の点数 (被験者平均)

5.1 実験手順

情報科学を専攻する大学生、大学院生 9 名に、提案システムによる類推手順の学習と説明文の作成を行ってもらう実験を行った。

実験は、1 日目に、「1 回目の説明文の作成」を行ってもらうとともに、「チュートリアル」と練習ステップの「特徴を挙げる練習」「共通点を見つける練習」を行ってもらった。説明文の作成においては、表 3 の説明文作成の課題に対して、「類推を用いてわかりやすい説明を行って下さい」と指示を与えた。2 日目に、練習ステップの「喩えるものを見つける練習」「特徴の対応付けの練習」、その後実践ステップを 5 回以上行ってもらい、1 回目と同じ課題で「2 回目の説明文の作成」を行ってもらった。3 日目に、全被験者が作成した 2 回分の説明文に対して、相互評価を行ってもらった。すなわち、説明文作成の課題ごとに存在する 18 の説明文のうち、被験者自身が作成したものを除く 16 の説明文に対して、無作為な順序で、以下の評価項目について 5 段階で評価を行ってもらった。

1. 説明能力：「説明したいもの」がわかりやすく説明されていますか？
2. ターゲットの特徴の適切性：「説明したいもの」の説明に欠かせない特徴を十分に挙げられていますか？
3. ベースの選択の適切性：「喩えるもの」の選び方は適切ですか？
4. ベースの特徴の適切性：「喩えるもの」の説明に欠かせない特徴を十分に挙げられていますか？
5. 特徴の対応付けの適切性：喩えを用いて、うまく対応付けて説明されていましたか？

5.2 実験結果と考察

図 9 に、2 回の説明文の相互評価の点数 (前節の 5 項目の合計で 25 点満点) を示す。すべての課題において、システムでの説明文の評価が高くなり、各項目ごとの評価値も、1 回目に比べて 2 回目の点数が高い結果となっていた。このことから、提案システムによる練習

表 3: 説明文作成の課題

ターゲットとベースの関係	説明課題
1. 同カテゴリ	好きな「球技」を「野球またはサッカー」に喩えて説明してください。
2. 並列カテゴリ	好きな「和食」を「何かの洋食」に喩えて説明してください。
3. 擬人化	好きな「乗り物」を「芸能人の誰か」に喩えて説明してください。
4. 逆擬人化	好きな「芸能人」を「何かの家電」に喩えて説明してください。
5. 抽象カテゴリ (観点有)	「コンビニ」を、「便利」という観点で何かに喩えて説明してください。
6. 抽象カテゴリ (観点無)	「消費税」を何かに喩えて説明してください。

表 4: 2 回目の説明文作成時に挙げられた特徴の数と対応付けが行われた特徴の数 (被験者平均)

課題	ターゲット	ベース	対応付け
同	5.4	5.3	5.2
並列	5.1	4.9	4.8
擬人化	5.2	4.8	4.6
逆擬人化	6.9	5.0	5.2
抽象 (観点有)	4.8	4.4	4.1
抽象 (観点無)	4.9	3.8	4.2

表 5: 事後アンケートの結果 (5 段階評価)

質問	5,4(肯定)	2,1(否定)
類推の手順を理解できましたか?	9	0
類推ができるようになりましたか?	8	1
説明に類推が有効だと思いますか?	9	0

は、よりわかりやすい説明文の作成に効果があったと言える。

表 4 に、2 回目の説明文作成時に、被験者によって挙げられた特徴の数、および対応付けが行われた特徴の数を示す。ターゲットとベースのそれぞれについて、5 個前後の特徴が挙げられ、そのほとんどの特徴について対応付けが行われていたことがわかる。このことから、被験者は類推の手順を理解し、特徴を挙げた上でそれらの対応付けを行い、それをもとに説明文を作成したことがわかる。表 5 に示す事後アンケートの結果においても、被験者全員が類推の手順を理解したと回答しており、またほとんどの被験者が類推ができるようになったと回答した。これらのことから、被験者はシステムを利用することによって、類推の手順を理解することができたと考えられる。

6 結論

本稿では、自身が知っている知識を、その知識を知らない他人に伝える際に、類推を用いてわかりやすく説明する能力を育成するシステムを提案した。評価実験により、システムを用いた被験者が、わかりやすい説明を行える能力を身につけられることを確認した。

今後は、類推を用いた知識伝達だけでなく、新しい物事を類推によって理解するための枠組み、また類推を用いて新しい考え方を発見できる枠組みに拡張していきたいと考えている。

参考文献

- [1] 文部科学省：言語活動の充実に関する指導事例集～思考力、判断力、表現力等の育成に向けて～【中学校版】，(2011).
- [2] 鈴木宏昭：説明と類推による学習，波多野誼余夫 (編)『認知心理学 5 学習と発達』，pp.149-179，東京大学出版会，(1996).
- [3] 植田一博：科学者の類推による発見，人工知能学会論文誌，Vol.15，No.4，pp.608-617，(2000).
- [4] 田中一男：アナロジーを用いた着想支援方式，情報処理学会研究報告知能と複雑系 (ICS)，Vol.1992，No.70，pp.41-50，(1992).
- [5] 金城敬太，尾崎知伸，古川康一，原口誠：アナロジーを組み込んだルール発想推論によるスキル獲得支援，人工知能学会論文誌，Vol.29，No.1，pp.188-193，(2014).
- [6] 佐藤治樹，松田稔樹：類推的思考に着目した問題解決力育成のための技術教育教材の開発，日本教育工学会研究報告集，JSET10-1，pp.129-136，(2010).
- [7] 岸学，綿井雅康：手続き的知識の説明文を書く技能の様相について，日本教育工学雑誌，Vol.21，No.2，pp.119-128，(1997).
- [8] Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko Kanzaki : Enhancing the Japanese Word-Net, The 7th Workshop on Asian Language Resources, in Conjunction with ACL-IJCNLP, (2009).
- [9] Wikipedia : (URL) <https://ja.wikipedia.org/>
- [10] Word2Vec : (URL) <https://code.google.com/archive/p/word2vec/>

キャラクターを用いたテキストマイニングスキルの獲得支援

Assistance for Acquiring Text Mining Skills Using Characters

後藤 賢悟¹ 砂山 渡² 川本 佳代¹
Kengo Goto¹ Wataru Sunayama² Kayo Kawamoto¹

¹ 広島市立大学大学院情報科学研究科

¹ Graduate School of Information Sciences, Hiroshima City University

² 滋賀県立大学 工学部

² School of Engineering, The University of Shiga Prefecture

Abstract: 近年、膨大なテキストデータから有用な情報を取り出すテキストマイニングの技術が注目され、多数のマイニングツールが開発されている。しかし、データの分析を行うことができる人は少なく、マイニングの初心者がスムーズに分析を行える環境が求められている。そこで本研究ではテキストマイニングツール TETDM の機能として、初心者ユーザを対象に、テキストマイニングスキルの獲得を目的としたチュートリアルシステムを提案する。

1 はじめに

近年、ネットワークや記憶装置の発達により、世の中には膨大な量のテキストデータが蓄積されている。そこで、それらの大量のテキストデータから有用な情報を取り出すテキストマイニングの技術が注目を集めており、現在までに多数のテキストマイニングツールやソフトウェアが開発されている。しかし、現状ではデータの分析を行うことができる人間は少ない。そこで、これまでテキストマイニングを行ってこなかった人が、スムーズにツールを使って分析を行える環境を構築することが重要となってくる。

テキストマイニングのソフトウェアの一つ、TETDM (Total Environment for Text Data Mining: テキストマイニングの統合環境) [1] では、利用者のテキストマイニングの実力に応じて「モード」を変更できることから、テキストマイニングの知識がない初心者にとっての入門ツールとして気軽に利用できることが期待される。しかしながら、初心者ユーザは自力でテキストマイニングのスキルを獲得することは非常に難しい。

そこで本研究では TETDM の機能として、初心者ユーザを対象とした、アシストキャラクターを用いたチュートリアルシステムを構築することで、ユーザのテキストマイニングスキルの獲得と向上を目指す。

2 関連研究

キャラクターを使用して、ユーザーの支援を行う研究として、益子らの「キャラクターを用いた学習継続

支援ツールの開発」[2] や、渡辺らの「セルフラーニング型授業におけるエージェントキャラクターによる学習支援」[3] がある。これらのアシストキャラクターの存在は、ユーザのシステムに対する利用意欲を高めると共に、ユーザの学習スキル向上に大いに貢献していることがわかる。実際、ここで取り上げた以外にも、多くの学習システムにはこのようなアシストキャラクターによる支援システムが搭載されている。

しかしながら、テキストマイニングの分野においては、このようなアシストキャラクターを使ったユーザのテキストマイニングスキル向上支援を行う研究は行われていない。

数あるテキストマイニングツールの中で、ユーザのテキストマイニングスキルの獲得を目的としたチュートリアルを搭載したツールはほとんど存在しない。データマイニングツールではあるが、チュートリアルに近い機能を搭載している KEEL [4] というツールがある。KEEL はデータマイニングのために、回帰、分類、クラスタリング、パターンマイニングのようなアルゴリズムを実装したツールである。KEEL では、実際にデータマイニングを行うときに利用する一連の機能をまとめた実験を設計する機能が提供されている。初めてデータマイニングを行うとき、ユーザはアルゴリズムがどのように動作するかを学ぶ必要がある。そのための教育のための機能が搭載されており、ユーザはアルゴリズムのリアルタイムビューに着目しながら短期間で実験を実施することによって、アルゴリズムの違いを学ぶことができる。本研究でも同様に、テキストマイニングのスキルを獲得を目的とした練習を短時間で繰り返し練習を行うことができるチュートリアルを実装する。

3 TETDM と利用モード

TETDM(Total Environment for Text Data Mining)は、テキストデータマイニングのための統合環境として作成された。TETDMでは豊富なツールが提供されており、それらのツールを使用することで利用者は文章のより詳細な分析を行うことができ、新たな発想が得られる可能性がある。現在、最新版としてTETDMサイト内 [5] で公開されている。

TETDMには、利用者のテキストマイニングの実力に応じて、以下の4つの起動モードを切り替えることができる。初めてTETDMを利用する人は、最初はスーパーライトモードから利用し、慣れたらライトモード、通常モードと順に利用モードを切り替えていくことを推奨している。

- スーパーライトモード
初心者が最初に利用出来るモード。あらかじめ用意された8つのツールセットを利用出来る。
- ライトモード
少しだけツールを操作してみたい人向けのモード。
- 通常モード
いろいろなツールを操作してみたい人向けのモード。より汎用性がある実用的な分析ができる処理ツールと可視化ツールが選択できるようになる。
- 拡張モード
より詳細な分析をしたい人向けのモード。専門的な内容を含む利用を想定した処理ツールと可視化ツールを選択できるようになる。

4 テキストマイニングのプロセスとスキル

本研究では、テキストマイニングを行う際に、ユーザが最終的に知識を獲得するまでの理想的なテキストマイニングプロセスを図1のように定義する。以下で、テキストマイニングの各プロセスの内容と各プロセスで必要とされるスキルについて述べる。

4.1 テキスト収集&テキスト整形

テキスト収集は、ユーザが実際に分析したいテキストを準備するプロセスである。テキスト整形は、分析したい内容に関するテキストを準備し、そのテキストをシステムに入力可能な状態にするプロセスである。例としては、テキストの段落分けや、句点の挿入処理、キー

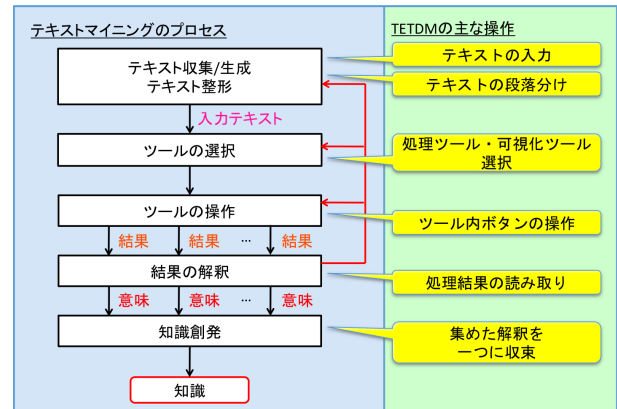


図 1: テキストマイニングのプロセス

ワード設定などが挙げられる。TETDMではテキスト収集はテキストファイルを用意するか、TETDM上でコピー&ペーストするかのどちらかで達成できる。テキスト整形は段落の区切りとなる場所を空行としておけば、ボタン一つで整形可能となっている。

テキスト収集プロセスでは、テキストを集める時に、分析の目的に応じてどのようにテキストを集めてくればいいのかをユーザが把握するスキルが必要となる。テキスト整形のプロセスについては、集めたテキストの量や、分析の目的に応じた段落分けを行うスキルが必要となる。また、テキスト整形では扱う品詞の設定ができるスキルも必要となる。文章中に出てくるものや人物などを知りたい場合は名詞、文章中の行動や動作を示した単語を特定したいなら動詞、大小や高低などの状態を知りたい場合は形容詞をそれぞれキーワードとして設定することが必要だ。

4.2 ツール選択

ツール選択は分析の目的に応じたテキストマイニングツールを選択するプロセスである。TETDMでは処理ツール(入力されたテキストに対し、テキストマイニング処理を行うツール)と可視化ツール(処理ツールによるテキストマイニング処理の結果を画面へ出力するツール)がある。それらを組み合わせることでツールのセットを行うことで達成できる。現在のTETDM(Ver2.21)では処理ツールが46種類、可視化ツールが35種類あるため、様々な組み合わせを試してツールの役割を覚えることが重要となる。

ツールの選択プロセスでは、分析の目的に適したツールを選択できるスキルが必要となる。その為には、各ツールを選択することでどのような処理結果を得ることができるかを把握する必要がある。

4.3 ツール操作

ツール操作は有意義な結果を得るため、ツールを操作するプロセスである。TETDM では、ツールの選択後に処理ツール内や可視化ツール内に配置されているボタンを押すなどの操作が当てはまる。

ツールの選択プロセスでは、操作の必要があるツールを選択した場合、分析の目的に適したツールの操作ができるスキルが必要となる。その為には、各ツールを操作することでどのような処理結果を得ることができるかを把握する必要がある。

4.4 結果の解釈

結果の解釈プロセスでは、ツールの出力結果（評価値）と、それが表す意味の解釈（どのような点が良いのか・どのように修正するのか）を行う。解釈は一度だけ行えば終わりではなく、場合によっては他のツールを選んだり、先ほどとは異なる操作を行うなど、別の視点から新たな解釈を行い、複数の解釈を集める必要がある。TETDM では各パネルに結果と解釈を登録できるフォームがあり、このフォームに結果と解釈を重要度（結果と解釈が自分にとってどのくらい重要なのか）と共に記述することで達成できる。

結果の解釈プロセスでは、処理結果のどこに着目するか、どのように結果を読み解いていくかについてのスキルが必要となる。また、次の知識創発プロセスで知識創発を行うためには、多くの解釈を集める必要があるため、場合によっては再度キーワード設定やツールの選択・操作を行う必要がある。そのため、次にどのような解釈を集めるべきか、そのためにはどのように行動すればいいかを考えるスキルも必要となるだろう。

4.5 知識創発

知識創発は得られた解釈を一つにまとめることによって大局的な知識を得るためのスキルとする。信頼性が高い知識を得るためにはより多くの解釈をまとめることが重要となる。TETDM では登録した解釈を一つにまとめることができる知識創発ウィンドウを利用することで達成できる。

知識創発プロセスでは、集めた解釈をまとめて、より多角的な視点から見た説得性の高い解釈へとまとめて上げることができるスキルが必要となる。そのためには、集めた解釈をどのように結合していくのかをどの順番で、どのように筋道を立てて解釈をまとめていくかが重要となる。

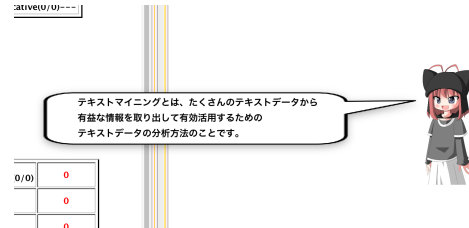


図 2: キャラクターの吹き出しによる説明文の表示

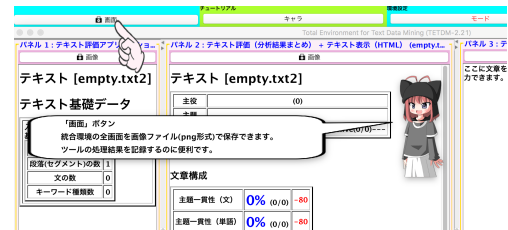


図 3: 指アイコンでのボタン提示のアシスト

5 キャラアシストチュートリアル

本章では、初心者のテキストマイニングスキルの獲得をアシストする、チュートリアルシステムについて述べる。

5.1 チュートリアルの枠組み

本チュートリアルは、「TETDM の操作方法の説明」と「テキストマイニングの流れ」の2つの課題群から構成されている。このうち、テキストマイニングのスキルに関わるのは後者となり、スキル獲得のための前提となる操作説明を別途設ける。

テキストマイニングスキル獲得を目指したチュートリアルを行う前に、まずはTETDMの基本的な操作方法をユーザに習得してもらう必要がある。「TETDMの操作方法の説明」のチュートリアル課題群では、TETDMの環境の操作説明チュートリアルである「操作の説明」と、TETDMで選択して利用出来るテキストマイニングツールの利用方法のチュートリアルである「ツールの説明」の2つの課題群に課題を分類した。環境の説明とツールの説明の課題を分けた理由は、ツールの説明に関してはツールの利用方法がわからない時に確認したいツールの説明のチュートリアルだけを確認できるようにするためである。また、TETDMは多くの開発者がツールの開発を行っており、仮にツールがこれから増えた際、チュートリアルもそれに応じて拡張しなければならないが、ツールの説明のチュートリアルだけが独立していることで、課題の拡張が容易であるというメリットもある。

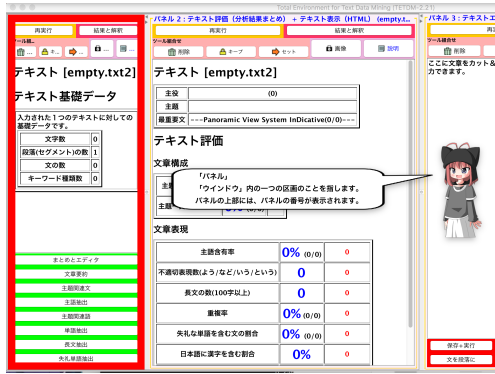


図 4: 四角でウィンドウを囲むアシスト

TETDM の利用方法のチュートリアル課題を終え、環境やツールの使い方を習得したユーザを対象に、テキストマイニングスキルの獲得を目指したチュートリアル「テキストマイニングの流れ」を行なってもらふ。本チュートリアルの課題は、テキストマイニングのプロセスの流れに則り、実践的な練習を行うことで、ユーザにテキストマイニングの流れと、各プロセスにおけるマイニングスキルの獲得を促す。

5.2 キャラクターによるアシスト内容

5.2.1 吹き出しを用いた説明文の表示

本チュートリアルにおける説明文は、キャラクターの台詞という形でキャラクターの左側に表示(図2)される。説明文は、最大で1行30文字×3行まで表示され、ユーザは吹き出しをクリックすると次の説明文が表示される。吹き出しを右クリックした場合は、一つ前の説明文に戻るようになっている。また、ユーザによっては吹き出しを連打して説明文を読まずにチュートリアルを進める恐れもあったため、新たに表示された説明文は説明文1行につき0.4秒間はクリックしても次の説明文が表示されない仕様になっている。また、キャラクターと吹き出しはドラッグ&ドロップで好きな位置に移動が可能のため、仮にボタン上にキャラクターや吹き出しが現れた場合は簡単に移動できる。キャラクターを移動したら、移動後に表示される吹き出しは移動したキャラクターに追従して表示されるようになる。仮にキャラクターを画面左に移動した場合、吹き出しは画面外に出ず、キャラクターの右側に表示されるようになっている。

表 1: 「TETDM の操作方法 (環境)」スーパーライトモード用課題

課題番号	課題名
1	はじめに
2	TETDM の概要
3	スーパーライトモードでできること
4	TETDM の各部の説明
5	TETDM のボタンの説明
6	TETDM の用語解説
7	ゲームモードの説明
8	テキストマイニングの流れ
9	テキストの入力方法
10	テキストの段落分け
11	ツールの選択方法

表 2: 「TETDM の操作方法 (ツール)」スーパーライトモード用課題

課題番号	課題名
1	まとめとエディタ
2	文章要約
3	主題関連文
4	主語抽出
5	主題関連語
6	単語抽出
7	長文抽出
8	失礼単語抽出

5.2.2 指アイコンおよび枠囲みによるボタン等のポイント表示

キャラクターの台詞による説明では、文章ではわかりにくく、説明が難しい内容も多い。例えば、ボタンの位置や各種マイニングツールの操作方法などの説明を行う場合がそれに当たる。そこで、ボタン位置等を指アイコンを用いてユーザに提示するアシストの導入(図3)を行う。このようなアシストは、TV ゲームのチュートリアル画面などではよく見られるが、ツールやアプリケーションの使い方の説明にはあまり用いられていない手法である。また、ウィンドウ全体を示すなど、指アイコンでは説明の限界がある場合は、ウィンドウを四角で囲む(図4)アシストを行う。

5.3 TETDM の操作方法についてのチュートリアル課題

本節では TETDM の操作方法についてのチュートリアル課題について、各起動モードで利用出来る機能とともに述べる。

5.3.1 スーパーライトモード

スーパーライトモードでは、テキストエディタにテキストをコピー&ペーストすることでのテキストの入

表 3: 「TETDM の操作方法（環境）」ライトモード用課題

課題番号	課題名
1	ライトモードでできること
2	TETDM のボタンの説明 2
3	TETDM のボタンの説明 3
4	キーワード設定について
5	句点の挿入
6	キーワード設定 (1)
7	キーワード設定 (2)
8	キーワード設定 (3)
9	アプリボタンによるツールの選択

表 4: 「TETDM の操作方法（ツール）」ライトモード用課題

課題番号	課題名
1	2 ちゃんまとめ
2	タイピング

力や、テキストの段落分けの機能など、簡単な機能のみが扱える。また、選択できるツールは、ツールは処理ツールと可視化ツールがあらかじめ組み合わせられたテキスト評価アプリケーションの 8 つのツールセットを利用出来る。チュートリアルも、この内容に沿って作成を行う。

スーパーライトモードでのチュートリアル課題は、TETDM の環境の操作説明の課題については表 1 に示す 11 題、ツールの使い方の課題については表 2 に示す 8 題を用意している。チュートリアルでは、サンプルテキストを用いて、環境の基本的な操作方法やツールの利用方法を学んでいく。

5.3.2 ライトモード

ライトモードでは、新たにキーワード設定の機能が追加される。また、選択できるツールセットも、新たに 2 種類増える。ライトモードでのチュートリアル課題は、TETDM の環境の操作説明の課題については表 3 に示す 9 題、ツールの使い方の課題については表 4 に示す 2 題を用意している。

5.3.3 通常モード

通常モードでは、テキストファイルを選択することでのテキストの入力や、処理ツールと可視化ツールを自由に組み合わせるツールの選択方法が新たに追加されており、ツールの種類数も処理ツール 27 種類＋可視化ツール 18 種類と、一気に増加している。また、新たに処理結果と解釈の登録機能と、知識創発の機能が追加された。通常モードの操作方法のチュートリアルも、

表 5: 「TETDM の操作方法（環境）」通常モード用課題

課題番号	課題名
1	通常モードでできること
2	TETDM のボタンの説明 4
3	テキストの入力方法 2
4	ツールの選択方法 2
5	ツールの選択方法 3
6	結果と解釈の登録
7	知識創発 1
8	知識創発 2

表 6: 「TETDM の操作方法（ツール）」通常モード用課題

課題番号	課題名
1	単語情報まとめ
2	文情報まとめ
3	セグメント情報まとめ
4	国語辞書
5	単語間関連速度
6	単語冗長文抽出
7	類似文抽出
8	再帰的クラスタリング
9	漢字の割合
10	タグクラウド

この内容に沿って展開され、サンプルテキストを利用した練習が可能である。

通常モードでのチュートリアル課題は、TETDM の環境の操作説明の課題については表 5 に示す 8 題、ツールの使い方の課題については表 6 に示す 10 題を用意している。なお、通常モードで行われるツールの説明の課題は、複数のツールを組み合わせる利用することが前提のツールについては一つの課題で組み合わせられる全てのツールについての説明を行っている。

5.3.4 拡張モード

拡張モードでは、複数のテキストファイルを同時に入力したり、テキストの一部分を部分テキストに設定して部分的にテキストマイニングできる機能などが追加された。ツールについては、処理ツール 46 種類＋可視化ツール 35 種類が選択可能で、分析の用途ごとに用意されたツールセットを搭載したアプリケーション 4 種類が新たに選択可能になっている。

拡張モードでのチュートリアル課題は、TETDM の環境の操作説明の課題については 6 題、ツールの使い方の課題については 17 題用意している。なお、拡張モードで行われるツールの説明の課題は、アプリケーションで選択できるものには課題名に通し番号が入力されている。

表 7: 「TETDM の操作方法（環境）」拡張モード用課題

課題番号	課題名
1	拡張モードでできること
2	TETDM のボタンの説明 5
3	テキストの入力方法 3
4	ツールの選択方法 4
5	ツールの選択方法 5
6	部分テキストの設定方法

表 8: 「TETDM の操作方法（ツール）」拡張モード用課題

課題番号	課題名
1	1-1: テキスト集合評価アプリケーション
2	1-2: テキスト分類
3	1-3: テキスト間類似度 A
4	1-4: テキスト間類似度 B
5	1-5: レポート評価
6	2-1: 部分テキスト作成アプリケーション
7	3-1: 文章推敲アプリケーション
8	3-2: 主題語含有率
9	3-3: 段落間類似度表示
10	3-4: 段落間木構造
11	4-1: プロフィールチェックアプリケーション
12	形態素解析
13	意見文抽出
14	単語置換
15	辞書再構築
16	単語抽出
17	段落の並び替え

表 9: 「テキストマイニングの流れ」スーパーライトモード用課題

課題番号	課題名
1	はじめに
2	テキストマイニングの流れ
3	テキストを集めよう
4	テキストを整形してみよう
5	ツールを選択して結果を読み取ろう (1)
6	ツールを選択して結果を読み取ろう (2)
7	ツールを選択して結果を読み取ろう (3)
8	ツールを選択して結果を読み取ろう (4)
9	ツールを選択して結果を読み取ろう (5)
10	ツールを選択して結果を読み取ろう (6)
11	集めた解釈をまとめてみよう
12	処理結果を画像処理しよう

表 10: 「テキストマイニングの流れ」ライトモード用課題

課題番号	課題名
1	はじめに
2	テキストマイニングの流れ
3	キーワード設定をしてみよう
4	簡単なツールの操作をしてみよう (1)
5	簡単なツールの操作をしてみよう (2)
6	解釈のまとめ

発の機能は通常モードから利用出来る機能なので、解釈を集めるところまでは行わない。

5.4 テキストマイニングのスキル獲得に向けたチュートリアル課題

本節ではテキストマイニングの流れのチュートリアル課題について、各起動モード毎に述べる。

5.4.1 スーパーライトモード

スーパーライトモードでのスキル獲得を目指したテキストマイニングの流れのチュートリアル課題は、表 9 に示す 12 題を用意している。スーパーライトモードでは、最初にテキストマイニングの全体的な流れの説明を行った後、テキスト収集&テキスト整形プロセスから結果の解釈プロセスまでの流れを練習できる。

「テキスト収集&テキスト整形」プロセスでは、テキストの効果的な集め方や目的に応じた段落分けの方法が学べる。ツール選択プロセスは、スーパーライトモードで扱える 8 つのツールセットを、どのような場面で選択すれば良いかを学べる。ツールの操作プロセスは、スーパーライトモードはツールの操作を基本的には行わないようにしたので、ここでの説明は省いている。結果の解釈プロセスでは、ツールの処理結果の読み取り方を学べる。結果と解釈の登録機能と知識創

5.4.2 ライトモード

ライトモードでのスキル獲得を目指したテキストマイニングの流れのチュートリアル課題は、表 10 に示す 6 題を用意している。ライトモードでは、新たに追加されたキーワード設定の機能を軸に、スーパーライトモードで学んだテキスト収集&テキスト整形プロセスから結果の解釈プロセスまでの流れを補足する形で学んでいく。また、ライトモードでは、スーパーライトモードでは行わなかったツールの操作を行い、処理結果がどのように変化するかを学ぶことができる。

5.4.3 通常モード

通常モードでのスキル獲得を目指したテキストマイニングの流れのチュートリアル課題は、表 11 と表 12 に示す 18 題（9 題 2 パターン）を用意している。通常モードでは、結果と解釈の登録機能と、知識創発インタフェースが新たに利用可能になったことから、本格的にテキストマイニングの全プロセスを通した分析の練習が可能である。

表 11: 「テキストマイニングの流れ」通常モード用課題

課題番号	課題名
1	はじめに
2	テキストマイニングの流れ
3	テキストファイルからのテキスト入力
4	テキストの大筋を見ていこう
5	背景語を見てみよう
6	単語の出現頻度を見てみよう
7	単語の出現頻度を見てみよう (2)
8	再帰的クラスタリング
9	集めた解釈から知識創発をしてみよう

表 12: 「テキストマイニングの流れ」通常モード用課題 (別パターン)

課題番号	課題名
1	はじめに
2	テキストファイルからのテキスト入力
3	テキストの大筋を見ていこう
4	主題の一貫性を見てみよう
5	主題の一貫性を見てみよう (2)
6	不適切表現の有無を見てみよう
7	単語冗長文の有無を見てみよう
8	集めた解釈から知識創発をしてみよう
9	知識創発で原因と結果を把握しよう

6 キャラアシストチュートリアル の有効性検証実験

6.1 実験内容

本節では実験内容について述べる。本実験の目的は、TETDM キャラアシストチュートリアル利用者のテキストマイニングスキル獲得・向上の検証を行うことである。被験者は、TETDM の利用経験がなく、なおかつテキストマイニングの経験がない大学生、大学院生の男女計 16 名とした。16 名の被験者のうち、8 名は今回作成したキャラアシストチュートリアルのうち、「TETDM の操作方法」「テキストマイニングの流れ」の両方の課題に取り組む提案群とした。残りの 8 名を比較群とし、「TETDM の操作方法」のチュートリアル課題のみに取り組み、テキストマイニングの流れに関しては、テキストマイニングの各プロセスについての操作方法やアドバイスを記入した紙媒体を元に、実際に TETDM を操作しながらテキストマイニングの流れについて練習するという形をとった。

実験期間は 3 日間で、最初の 2 日間で課題を行ってもらい、最終日には事後テストを行った。事後テストは、提案群・比較群同様に同じテキストを TETDM で分析してもらった。事後テストに用いたのテキストは、ポータルサイト Amazon で販売される nikon 製カメラのレビューを収集したものである。事後テストの内容は、被験者は nikon とは別の会社の社員であると仮定して、ライバル会社 nikon に勝つためのカメラのアイデ

表 13: 根拠に基づいたカメラのアイデア数 (被験者平均)

提案群	比較群
3.1	1.5

表 14: キーワード設定回数 (被験者平均)

提案群	比較群
2.3	0.6

アを nikon のカメラのレビューを TETDM を用いて分析することで考えるというものである。

6.2 実験結果と考察

本節では実験結果と考察について述べる。

最初に、事後テストで被験者に考えてもらったカメラのアイデアの中で、根拠に基づいたアイデアの個数についての比較を行った結果を表 13 に示す。t 検定の結果、両軍のアイデア数には有意な差があると確認できた。このことから、提案群のほうが分析を行った根拠に基づいたアイデアが多い傾向にあり、テキストマイニングのスキル獲得は提案群のほうが優れていたことがわかる。

また、各被験者の分析手順の差を比較するために、事後テスト後に回収した TETDM の操作履歴のログデータから、テキストマイニングの各プロセスにおける、提案群と比較群のテキストマイニングの手順の差を比較した。

テキスト収集&テキスト収集プロセスでの両群のキーワード設定を行った回数の差を表 14 に示す。キーワード設定回数は、提案群のほうが比較群よりはるかに多い結果となった。

ツールの選択プロセスでの、ツールの選択回数を表 15 に、ツールの操作プロセスの、ツールの操作回数を表 16 にそれぞれ示す。この結果から、提案群のほうがより多くツールの選択や操作を行ったことが確認出来る。

提案群のキーワード設定やツールの選択、操作の回数が増えた要因には、テキストマイニングプロセスの実践に関して紙で説明されるよりも、チュートリアルによって実際に試す機会を与えることで各プロセスでの分析や操作をより多く行っていたからだと考えられる。これらの操作回数増加は、より良い分析結果を得るための試行錯誤の回数が増加したと考えることができる。

結果の解釈プロセスでの、TETDM のツールの処理結果と解釈の登録機能を用いて解釈をまとめた個数の

表 15: ツールの選択回数 (被験者平均)

提案群	比較群
50.4	30.1

表 16: ツールの操作回数 (被験者平均)

提案群	比較群
32.5	11.5

データを表 17 に示す。この結果から、提案群のほうがより多く結果と解釈を登録したことがわかる。

このような結果となった理由は 2 つ考えられる。まず一つ目は、提案群はキーワード設定やツールの選択・操作でより良い処理結果を集めようと試行錯誤を何度も繰り返した結果、理想的な処理結果を集めることができたために、解釈の登録数も増加したと考えられる。また、二つ目の理由としては、提案群はテキストマイニングスキル獲得のためのチュートリアル課題で何度も実践練習を行ったことで、処理結果を読み取って自分なりにまとめる力が比較群よりも向上していることも考えられる。従って、結果の解釈プロセスでは提案群のほうがテキストマイニングスキルを獲得できたと言える。

知識創発プロセスでの、提案群と比較群で各ユーザが知識創発を行ったかについてを、表 18 に示す。この結果から、提案群のほうが知識創発の機能を用いて集めた解釈をまとめようとしていたことがわかる。このような結果になった理由として、提案群は多くの解釈が集まり、自分の頭だけでは解釈をまとめきれなくなった結果、積極的に TETDM の知識創発機能を用いて解釈をまとめるように意識することができたからだと考えられる。従って、知識創発プロセスでは、提案群のほうがテキストマイニングスキルを獲得できたと言える。

7 結論

本研究では、TETDM の機能の一つとして、TETDM 利用者のテキストマイニングスキルの向上させるためのキャラクターを用いたチュートリアルシステムの開発を行った。その後、評価実験により本システムが TETDM を始めて利用するユーザのテキストマイニングスキルの獲得・向上に有効であることを検証した。

今回はどんな分析の目的のテキストマイニングにも通用する基本的なスキルを獲得することを目的としたチュートリアルの構成を行ったので、今後は分析の目的に応じた、より専門的なスキルを獲得できるチュートリアル課題の拡張なども行っていきたい。

表 17: 結果と解釈の登録数 (被験者平均)

提案群	比較群
8.1	3.5

表 18: 知識創発を行った被験者数

	提案群	比較群
1 つの結論にまとめた	3	1
複数の結論にまとめた	3	2
知識創発を行っていない	2	5

参考文献

- [1] 砂山渡・高間康史・西原陽子・梶並知記・串間宗夫・徳永秀和：統合環境 TETDM を用いたマイニングツールの開発と利用の実践，人工知能学会論文誌，Vol.29, No.1, pp.100-112, (2014)
- [2] 益子博貴，松村敦，宇陀則彦：キャラクターを用いた学習継続支援ツールの開発，ARG 第 1 回 Web インテリジェンスとインタラクション研究会 (ARG SIG-WI2)，(2012)
- [3] 渡辺博芳，水柿恵：セルフラーニング型授業におけるエージェントキャラクタによる学習支援，情報科学技術フォーラム一般講演論文集 (N. 教育・人文科学)，Vol.3, No.4, pp.335-337, (2004)
- [4] J.Derrac,J.Luengo,Alcal'a-Fdez,A.Fern'andez,S.Garc'la:Using KEEL-Software as a Educational Tool:A Case of study Teaching Data Mining, Next Generation Web Services Practices(NWeSP),2011 7th International Conference on, Salamanca,pp.464-469,(2011)
- [5] TETDM:(URL) <http://tetdm.jp/>
- [6] 中垣内李菜，川本佳代，砂山渡：統合環境 TETDM を用いたテキストマイニングにおける初心者のためのツール選択支援，第 27 回人工知能学会全国大会，3B3-NFC-01a-1, (2013)
- [7] キャラクターなんか機:(URL) <http://khamix.sakura.ne.jp/download.shtml>

Twitterを用いた感染症発生動向の視覚化

Visualization of the occurrence trend of infectious diseases using Twitter

松本 流星^{1*} 吉田 稔^{1,2} 松本 和幸^{1,2} 北 研二^{1,2}
Matsumoto Ryusei¹ Yoshida Minoru^{1,2} Matumoto Kazuyuki^{1,2} Kita Kenji^{1,2}

¹ 徳島大学工学部知能情報工学科

¹ Faculty of Engineering, Tokushima University

² 徳島大学大学院理工学研究部

² Institute of technology and science, Tokushima University

Abstract: The purpose of this research is to extract comments on infectious diseases from Twitter and to create an infection map. As a method, we check the facts of the extracted utterance, extract the past tweets of the infected person from the result, and estimate the place where the person lives. We used SVMs both for the fact check and place estimation.

1 はじめに

本研究では、Twitter から感染症に関する発言を抽出し、感染症マップを作成することを目標とする。現在、一般ユーザが感染症などの病気に関する情報を入手する場合、テレビのニュース、新聞、インターネットなど限られた情報源しか存在しない。また、世の中に飛び交っている多種多様な情報の中でも病気や感染症といった人間の生命に関する情報は、多くの人々にとって重要であると考えられる。これらを解決する方法として、SNS の中でも特にリアルタイム性が高く、且つ一度に大量の情報を入手できる Twitter を利用することで、感染症に関する情報を入手し、現在どこでどのような感染症がどのくらい流行しているのかということがわかる感染症マップのようなものがあれば、一般ユーザにとって有益な情報となり得る。

本研究では、Twitter 上に投稿されているツイートから感染症の名称が含まれているツイートを抽出する。抽出したツイートの文章の内容から投稿した人物が感染症に感染しているかどうかを判定する事実性解析を行う。事実性解析の結果、感染症に感染していると判定された場合、そのツイートの投稿者の過去のツイートを検出し、そのツイートの内容からその人物の住んでいる位置を推定する。位置推定した結果を用いて感染症マップを作成する。

2 既存研究

Twitter と病気を組み合わせた研究はこれまでにいくつか行われている。荒牧ら [1] の研究では、Twitter 上に投稿されているインフルエンザに関連する発言を抽出し、抽出した発言に対して SVM を用いた事実性解析を行い、その結果よりインフルエンザの流行検出を行うという手法を提案している。また、荒牧が開発した Twitter からインフルエンザに関連する発言を抽出し、日本全国の地図と重ねることでインフルエンザの流行地図を提供するサービスの「インフルくん」[6] がある。しかし、これらはインフルエンザのみを対象としており、また感染している投稿者のツイートに GPS により付与されたジオタグ (位置ラベル) やプロフィールに記載されている現在地を元に感染者の位置を示すため、ジオタグが付与されていないツイートやプロフィールに現在地が記載されていないユーザに対して対応ができないといった問題点がある。また、既存研究ではインフルエンザの流行検出を最終目標としているが、本研究ではインフルエンザ以外の複数の感染症を対象とし、感染症に感染している人物の住んでいる場所を推定し、地図上に示すことで感染症マップを作成することを最終目標とする。

2.1 事実性解析

事実性解析とは、文章中のある事象が実際に起こったことなのか、あるいは起こる可能性を述べただけなのかを判定する手法である。本研究では、以下の二項

*連絡先：徳島大学工学部知能情報工学科
〒 770-8506 徳島県徳島市南常三島町 2 丁目 1 番地
E-mail: c501306048@tokushima-u.ac.jp

目について事実性解析による判定を行う。

① ツイートの内容からその投稿者が感染症に感染しているかどうか

② 位置情報を表す語を含むツイートの内容からその投稿者がその場所に住んでいるかどうか

事実性解析の手法としてSVMを用いた教師あり機械学習を用いて行う。SVMについては2.2項で詳しく述べる。

2.2 SVM

SVM(サポートベクトルマシン)[5]とは、機械学習における教師あり学習のパターン認識のアルゴリズムの一種である。SVMは、線形入力素子を利用して2クラスのパターン識別器を構成する手法である。

線形分離超平面上に正例と負例の2つのクラスが存在し、この2クラスを分類する識別面を求めるとする。識別面が正例と負例それぞれの最も近い事象との距離であるマージンが最大になるように識別面を設定する。この識別面を用いることで未知の事象に対してのクラスを求めることができる。

本研究では、感染症に関しては感染症に感染しているツイートを正例、感染症に感染していないツイートを負例とし、位置推定に関しては位置情報を表す語を含むツイートの中で投稿者が自分の住んでいる場所でツイートしているものを正例、それ以外のものを負例として識別面を求め、クラスが未知の新たなツイートに対して分類を行う。

3 提案手法

3.1 提案手法の概要

本研究で提案する手法は、Twitter上に投稿されている感染症に関するツイートの文章を対象に、クラス分類を行い、その投稿者が感染症に感染しているかどうかを判定するものである。その結果から、感染症に感染していると判定された人物の過去のツイートの文章より、その人物が住んでいる場所を推定し、それらの結果を元に視覚化を行う。以下に手順を示す。

- Twitter上に投稿されている感染症の名称が含まれているツイートを取得する
- 取得したツイートに対して事実性解析を行い、投稿者が感染症に感染しているか否かを判定する
- 感染症に感染していると判定されたユーザの過去のツイートを取得する

- 取得したツイートに対して位置情報データベースを用いて位置情報を表す語を検出し、位置推定を行う
- 位置推定結果より、地図上にプロットした感染症マップを作成する

この提案手法を数式化すると以下ようになる。

- ① ユーザ U から得られた感染症に関する1ツイートを u_i とする
- ② 感染症に関するツイート u_i に対して事実性解析を行い、感染していると判定したツイートを u_{it} とする
- ③ 感染していると判定したツイート u_{it} を投稿したユーザ U の過去のツイートを u_p とする
- ④ 過去のツイート u_p に対して位置情報データベースを用いてマッチングを行い、位置情報を表す語 $x(u_p)$ を検出する
- ⑤ 語 $x(u_p)$ を含んでいる過去のツイート u'_p の前後でつぶやかれているツイート u'_q を検出し、 u'_p との関係性を元にその人物の住んでいる位置を推定する
- ⑥ ツイート u'_p の中から位置を推定したツイート u'_t に対して事実性解析を行う

3.1.1 感染症に関するツイートの事実性解析

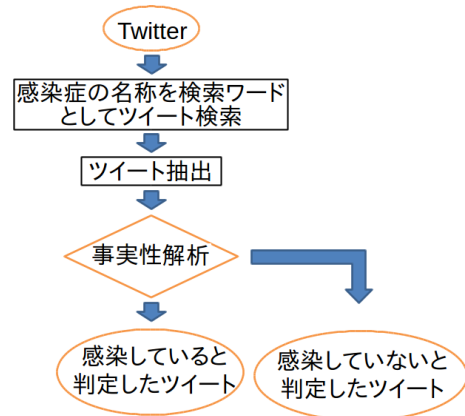


図 1: 感染症に関するツイートの事実性解析のフローチャート

感染症に関するツイートの事実性解析について、簡単な流れを図1に示す。

- ① Twitter上に投稿されているツイートから、感染症の名称を検索ワードとして検索を行い、その名称が含まれているツイートを取得する。
- ② 取得したツイートに対してSVMを用いた事実性解析を行い、その投稿者が感染症に感染しているか否かを判定する。その結果から、感染症に感染していると

判定したツイートと感染していないツイートとに分類する。

3.1.2 位置情報の推定

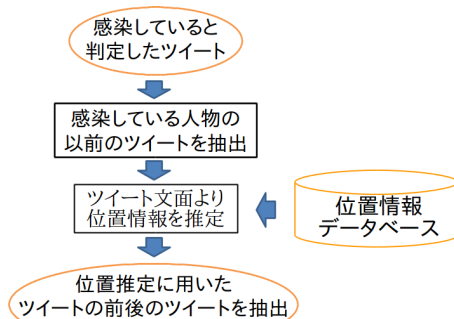


図 2: 位置情報の推定のフローチャート

位置情報の推定について、簡単な流れを図2に示す。

- ① 事実性解析の結果、感染症に感染していると判定されたユーザの過去のツイートを取得する。
- ② 取得した過去のツイートの中から地名やスポット名などを含んでいるものを、位置情報データベースを用いて検出する。
- ③ 検出した語を含んでいるツイートの前後でつぶやかれているツイートを抽出し、そのツイートとの関連性を見る。その結果から、位置推定を行う。

3.1.3 位置推定に関するツイートの事実性解析

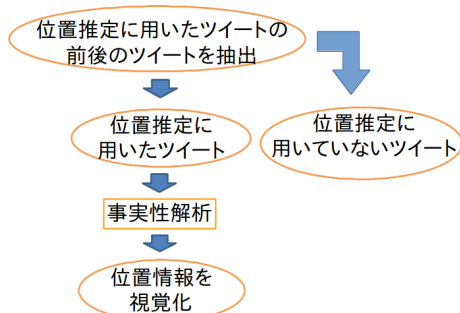


図 3: 位置推定に関するツイートの事実性解析のフローチャート

位置推定に関するツイートの事実性解析について、簡単な流れを図3に示す。

- ① 位置推定に用いたツイートと用いていないツイートとに分類する。
- ② 位置推定の精度をより向上させるために、位置推定に用いたツイートに対して感染症に関するツイートと

同様の事実性解析を行う。

- ③ 事実性解析を行った後に、位置推定に用いたツイートを元に地図上にプロットし、視覚化を行う。

3.2 提案手法のアルゴリズム

3.2.1 感染症に関するツイートの取得

ここでは、Twitter からツイートを取得する方法を説明する。本研究ではツイートを取得するために TwitterAPI を使用する。検索対象として夏に流行している7種類の感染症の名称もしくは同義語を用いる。対象とする感染症については、4節で詳しく述べる。同義語を用いる理由として、本研究で対象にした感染症の多くはその名称よりも同義語がツイート文に含まれていることが多かったためである。例として「流行性耳下腺炎」という感染症を挙げると、この感染症の別名は「おたふく風邪」であり、一般的には後者の方が使われていることが殆どである。このように正式名称よりも通称で用いられることが多い感染症が多いため、同義語を用いる。

ここで例として、「ヘルパンギーナ」という感染症の名称を用いて検索を行う実験方法を説明する。

図3.4は取得したツイート例であるが、これを各感染症毎に400ツイート取得し、事実性解析を行う。

病院で検査したら、ヘルパンギーナだった
ヘルパンギーナ疑惑浮上
やっぱ流行ってるよね、ヘルパンギーナ
@XXX 今流行してるヘルパンギーナじゃないですか？

図 4: ヘルパンギーナのツイート例

3.2.2 感染症に関する事実性解析

ここでは、取得したツイートに対して事実性解析を行う。事実性解析を行う手法としてSVMを用いる。SVMを用いる判定には、学習用データとテストデータが必要となる。ここで使用する学習用データは、予め対象となる各感染症の名称で取得した1000ツイートとする。また、1000ツイートの内訳として正例と負例それぞれが500ツイートとなるようにツイートを取得する。テストデータは、3.1.1項で各感染症毎に取得した400ツイートとする。学習用データに対して感染症に感染しているか、感染していないかを目視で判定し、手動でラベル付け(1, -1)を行う。図5が学習用データの例である。ラベルの意味として、「1」は感染しているもの、「-1」は感染していないものとなっている。

この学習用データを用いて事実性解析を行い、取得したツイートに対してラベル付けを行う。

病院で調べたらヘルパンギーナだったわ 1
今ヘルパンギーナでダウン中 1
ヘルパンギーナかもしれない -1
今周りでヘルパンギーナが流行ってる -1
ヘルパンギーナ怖いなあ・・・ -1

図 5: 感染症に関する学習用データの例

明日から研修旅行だぜ!
↓
1日目終了乙
↓
2日目オワタ
↓
東京ばなな買い忘れた。買おう。(「東京」を検出)
↓
長いようで短いような研修旅行でしたー!

図 7: 旅行もしくは観光を表しているツイート例

3.2.3 位置情報の推定

前項の感染症に関する事実性解析で感染していると判定されたツイートから、そのツイートを行っているユーザのユーザ ID を用いて、そのユーザの過去につぶやかれているツイートを抽出する。また、過去のツイートの基本的な取得条件として、そのユーザの最新のツイートから 200 ツイートまでを取得する。200 ツイートで取得する理由として、ツイート数が少ないユーザから位置情報を推定するのは非常に難しくなるからである。そのため、取得可能なツイート数が 200 以下のユーザは除外する。また、推定を行う人数は各感染症毎に感染していると判定された中から無作為に選んだ 100 人とする。

過去のツイートを抽出したら、そのツイートに対して位置情報データベースを用いてマッチングを行う。位置情報データベースとは、県名や市名、駅名などのスポット名といったような位置情報を表す語を元に作成したデータベースである。県名と市名に関しては、国土交通省国土政策局国土情報課が提供している位置情報ダウンロードサービス [7] の「大字・町丁目レベル位置参照情報」というものを参照している。図 6 はその情報の一部である。この位置情報データベースと抽出した

1	国道路線名	市区町村名	大字・町丁目名	市区町村番号	道路番号	道路番号	経度	緯度	経度
2	北海道	札幌市中央区	南十二条西五丁目	2	12	106529.3	-75.6611	43.037701	141.321212
3	北海道	札幌市中央区	南十二条西六丁目	2	12	106532.4	-75.6614	43.037501	141.320699
4	北海道	札幌市中央区	南十二条西七丁目	2	12	107388.8	-75.6113	43.029995	141.323831
5	北海道	札幌市中央区	南十二条西八丁目	1	12	107379.7	-75.6485	43.030403	141.323831
6	北海道	札幌市中央区	南十二条西九丁目	1	12	106984.4	-75.6391	43.034403	141.323831
7	北海道	札幌市中央区	南十二条西十丁目	1	12	106945.9	-75.6378	43.033805	141.323208

図 6: 位置情報参照データ

ツイートをマッチングを行うことで、ツイートの文章中に含まれている位置情報を表す語を検出する。位置情報を表す語を検出できたら、その語を含んでいるツイートの前後でつぶやかれているツイートを抽出する。このツイートを利用することで、位置情報を表す語を含むツイートとの関係性を見ることができる。例として旅行もしくは観光中であることを示しているツイートがある場合、その前後でつぶやかれているツイートはその旅行先あるいは観光先に関係しているツイートである傾向がある。取得したツイートに対して旅行あるいは観光でその場所を訪れているかを手動で判別する。そのツイートの前後でつぶやかれているツイートを抽出し、それらのツイートに一貫性があれば旅行あるいは観光中であると判定する。図 7 が例である。

3.2.4 位置推定に関する事実性解析

前項の位置情報の推定精度をより向上させるために、位置情報に関しても感染症に関するツイートと同様に事実性解析を行う。ここで使用する学習用データは、位置情報を表す語が含まれている 1000 ツイートとする。また、1000 ツイートの内訳として正例と負例それぞれ 500 ツイートとなるようにツイートを取得する。ここでの正例は、位置情報を表す語の場所に住んでいると判定できるツイートを意味し、負例はその場所に住んでいない判定できるツイートを意味している。テストデータは、前項の位置推定でユーザの住んでいる位置を推定できるようなツイートとする。学習用データに対して正例あるいは負例であるかどうかを目視で判定し、手動でラベル付け (1, -1) を行う。図 8 が学習用データの例である。ラベルの意味として、「1」は投稿者が自分の住んでいる場所でツイートしているもの、「-1」はそれ以外のものとなっている。

思ったより早く東京に帰れそうだ 1
1ヶ月の出張やっと終わったー!!!ただいま名古屋!!! 1
@XXX 昨日はお疲れ様ー!!また大阪で待ってる! 1
週末は一人で東京に行く -1
買ってから明日学校行こうかな 博多で一泊降るべき? -1

図 8: 位置情報に関する学習用データの例

4 実験

本研究では、以下の 7 つの感染症を対象として、事実性解析の精度評価の結果を示す。

- ① ヘルパンギーナ
- ② おたふく風邪
- ③ 感染性胃腸炎
- ④ 流行性結膜炎
- ⑤ マイコプラズマ
- ⑥ プール熱
- ⑦ 手足口病

本研究では、夏に流行する感染症を対象としているため、7 月～8 月の間で収集したツイートを実験データとして用いる。位置推定に関しては、新たに取得したツイッターアカウントを対象に位置推定を行った結果を示す。

また、本研究では、SVM を用いた事実性解析を行うために Scikit-learn[4] を用いる。Scikit-learn とは、Python のオープンソース機械学習ライブラリで、SVMをはじめとし、Random Forest、Gradient Boosting、k 近傍法などを含む様々な分類、回帰、クラスタリングアルゴリズムを備えている。本研究では、SVM を用いた教師あり機械学習を行う。

4.1 評価方法

感染症に関する評価方法に関しては、4.2.1 項の式 (4.1) を用いた精度評価を行う。位置情報の推定に関する評価方法に関しては、新たに評価用アカウントからデータを取得する。評価用データとして、新たに Twitter から 100 個のユーザアカウントを取得する。取得条件は以下のようになっている。

- ① プロフィールに位置情報が記載されている
- ② Twitter クライアント名、アカウント名、表示名、プロフィールのうち、1 つ以上に「BOT」「Bot」「bot」のいずれかを含む

上記の条件の①に関しては、ツイート文から位置情報を推定した結果が正しいかどうかを簡単に判定することを目的としている。また、プロフィールに位置情報を記載しているユーザは自分の住んでいる地域を表す語を含んでいるツイートが位置情報を記載していないツイートよりも多い傾向にある。②に関しては、Bot アカウントを除外することを目的としている。

4.2 実験結果

4.2.1 感染症に関する事実性解析

ここでは、対象となっている 7 つの感染症に対して SVM を用いた事実性解析の精度の評価結果を表 1 に示す。精度の算出式は式 (4.1) で行う。

事実性解析の精度 = (事実性解析の正解数)/(取得したツイート数) ... (4.1)

表 1: 感染症に関する事実性解析の実験結果

	精度
ヘルパンギーナ	83.75
おたふく風邪	66.25
感染性胃腸炎	79.75
流行性結膜炎	72.00
マイコプラズマ肺炎	84.00
プール熱	68.50
手足口病	67.75

表より、66～84[%] という結果になった。

4.3 位置推定に関する事実性解析

ここでは、評価用アカウントから取得したツイートに対して SVM を用いた事実性解析の精度の評価結果を表 2 に示す。精度の算出式は式 (4.2) を用いる。

事実性解析の精度 = (事実性解析の正解数)/(位置推定を行ったツイート数) ... (4.2)

位置推定を行ったツイートとは、例として東京という語が含まれているツイートの場合にそのツイートから投稿者が東京に住んでいると推定できたツイートを意味する。

表 2: 位置推定に関する事実性解析の実験結果

	精度
位置推定の事実性解析	78.00

表より、78.00[%] という結果になった。

次に、3.2.2 項で感染していると判定されたツイートの投稿者の位置推定に関する事実性解析を行った結果を示す。

表 3: 位置推定の事実性解析の実験結果

	精度
ヘルパンギーナ感染者	81.00
おたふく風邪感染者	79.00
感染性胃腸炎感染者	76.00
流行性結膜炎感染者	62.00
マイコプラズマ肺炎感染者	78.00
プール熱感染者	71.00
手足口病感染者	65.00

表より、62～81[%] という結果になった。

4.4 位置情報の視覚化

ここでは、ヘルパンギーナの感染者を位置推定した結果を元に作成した感染症マップを示す。地図上に示しているマークは、以下のような意味をする。

赤色のマーク:感染者が 1 人

青色のマークと数字:感染者が 2 人以上 10 人未満

黄色のマークと数字:感染者が 10 人以上



図 9: 位置情報に関する学習用データの例

5 考察

5.1 感染症に関する事実性解析

本研究では、SVM を用いて事実性解析を行ったが結果より感染症によって誤分類されている割合に変化があることがわかる。この要因として、「おたふく風邪」という感染症は基本的に子供がかかる感染症である。本研究では、ツイート文の投稿者本人ではなくその人物の身内が感染している場合でも正例と扱っている。例として「息子がおたふく風邪にかかった」というツイート文があった場合、このツイート文は正例となる。また、「子供が通っている学校でおたふく風邪が流行っている」、「娘と同じ幼稚園の子がおたふく風邪にかかっている」というように自分の子供が通っている学校などで流行していることを示しているツイートが多い傾向があった。このため、おたふく風邪や手足口病といった子供に流行る感染症に関しては精度が低くなったと考えられる。より精度を向上させるためには、学習用データの増加と偏りを少なくすることが有効であると考えられる。

5.2 位置推定に関する事実性解析

位置推定に関する事実性解析に関して、今回は2種類のデータを用いた評価を行った。誤分類される要因として、今回実験に用いたツイート文の多くは、一つのツイート文に一つの位置情報を表す語を含んでいるものが多い傾向にあった。しかし、一つのツイート文中に複数の位置情報を表す語が含まれているものもあり、誤分類されているツイートの多くは複数の位置情

報を表す語を含んでいるものであった。より精度を向上させるためには、学習用データの増加、ノイズ除去方法の改良あるいは追加、位置情報データベースを増加することが有効であると考えられる。現在行っているノイズ除去の方法は、音符などの特殊な記号のみとしている。これに加えてより多くの特殊な記号や顔文字などを除去することで精度が向上すると考えられる。位置情報データベースは、現在はスポット名に関しては「スカイツリー」や「鳥取砂丘」、イベント名に関しては「東京ゲームショウ」といったような全国的に知られているものが多いため、もう少し知名度の低いものも追加することで精度が向上すると考えられる。

6 結論

本研究では、感染症に関するツイートに感染しているか否かを判定する分類器と位置情報に関するツイートに対してそのツイートが住んでいる場所をつぶやかれているか否かを判定する分類器を作成し、その結果から感染症マップを作成した。結論として、実験よりTwitter ユーザの位置を推定することができたが、その中には正しく推定できなかったものもあった。そのため、事実性解析の際に用いる学習用データの増加や偏りを失くすことが今後の課題となる。

参考文献

- [1] 荒牧 英治, 増川 佐知子, 森田 瑞樹: Twitter Catches the Flu:事実性判定を用いたインフルエンザ流行予測, 情報処理学会研究報告, 2009
- [2] 杉谷 卓哉, 白川 真澄, 原 隆浩, 西尾 章治郎: 教師あり教科学習を用いたツイート投稿時のユーザ位置推定手法, 情報処理学会研究報告, 2013
- [3] 東京都感染症情報センター, <http://idsc.tokyo-eiken.go.jp/>
- [4] python の機械学習ライブラリ scikit-learn の紹介, <http://sucrose.hatenablog.com/entry/2013/05/25/133021>
- [5] サポートベクターマシン入門, <http://home.hiroshima-u.ac.jp/tkurita/lecture/svm.pdf>
- [6] インフルくん, http://mednlp.jp/influ_map/
- [7] 位置情報ダウンロードサービス, <http://nlftp.mlit.go.jp/isj/>

地理的な関連性を持つ時系列データの探索的分析を支援する 可視化システムに関する研究

Visualization System for Exploratory Analysis with Geographical Time-series Data

内藤 峻^{1*} 岩崎 有基² 松下 光範²
Shun Naito¹ Yuki Iwasaki² Mitsunori Matsushita²

¹ 関西大学大学院 総合情報学研究科
Graduate School of Informatics, Kansai University

² 関西大学 総合情報学部
Faculty of Informatics, Kansai University

Abstract: The goal of our study is to support a user's analysis of time-series data in an exploratory manner. Such exploratory analysis requires repeated access to various types of information related to the user's interests such as texts and numerical data. To support such the user's analysis, we have proposed a system that visualizes temporal changes in time-series data and presents the causes of those changes with the data. In this paper, we improve the system by (1) strengthening the linkage between graphs and articles (2) adding graph interactions. In addition, we conduct user experiments and examine the point of arrival of the system and the direction of future expansion.

1 はじめに

時系列データとは、熊本地震の被害者数や気温など時間の経過に伴って変化するデータである。このような時系列データは意思決定や問題解決の場面で役立てられている。意思決定や問題解決の場面では、時系列データの値の変化やその変化の要因を分析することで、有益な情報や新たな知見を得ることが重要である。しかし、このような時系列データの分析は仮説の生成や検証を探索的に繰り返す負荷の高い作業であるため、ユーザがこのような探索行為を円滑に行うことが難しいという問題がある。そこで本研究では、ユーザの興味や関心に応じて様々なモダリティの情報へのアクセスを繰り返しつつ時系列データを分析するための支援システムの実現を目指している。その端緒として、著者らはこれまでに、新聞記事と地図、統計データを対象に、ユーザが時系列データの経時的変化とその変化の要因の把握を可能にする可視化インタフェースを提案してきた [1]。先行インタフェースでは、(1) 容易に興味を持った時点の記事を参照することができない、(2) データの推移を把握しづらい、という問題があったため、本稿では、(1) グラフと記事との連携の強化、(2) グラ

フのインタラクションの追加、によりユーザの探索行為の円滑化を図る。加えて、ユーザ実験を行い、システムの到達点と今後の拡張の方向性について検討する。

2 関連研究

2.1 言語情報を用いた時系列データの探索的分析の支援

松下らは、動向情報テキストから統計量を抽出し、グラフとして可視化する手法を提案している [2]。動向情報テキストから単に統計量を抽出し、グラフの描画を行った場合、とびとびの値のみを抽出することしかできないため、ユーザの役に立つ可視化表現を得ることができない。この問題を解決するために、動向情報テキストの視覚的な情報要約を行うことを目的としている。提案手法では、比較表現と背景知識を用いて統計量を補填したり、テキスト中の定性表現を解釈し、グラフ上にその表現に当たる図形を描画したりしている。実験では、提案手法の補填によって多くの統計量を抽出できることと、提案手法によって生成された可視化表現がユーザの理解を高めることを明らかにしている。

山本らは、動向情報の変化とその変化の要因を視覚的に表示するシステムを研究している [3]。システムは、

*連絡先： 関西大学大学院総合情報学研究科知識情報学専攻
〒 569-1095 大阪府高槻市霊仙寺町 2-1-1
E-mail: mat@res.kut.ac.jp

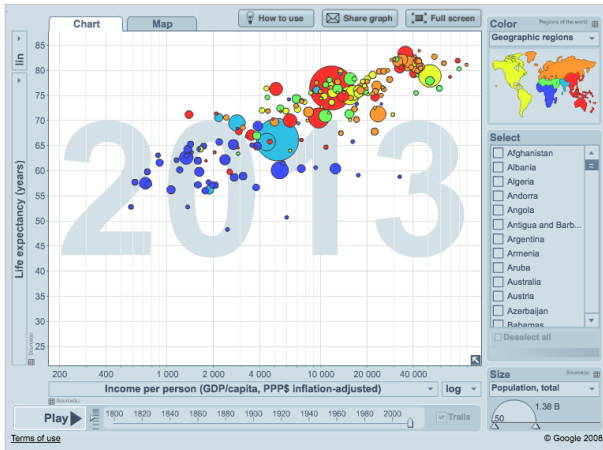


図 1: Gapminder (文献 [4] より図引用)

内閣支持率に関する新聞記事を入力することでユーザの興味を考慮し、変化の根拠となる要因をグラフ上に配置する。また要因の抽出手法では、動向情報が記載された新聞記事とコサイン距離が近い記事を要因としている。さらに、表示方法に関してユーザ実験を行い、ユーザが動向情報の要因を知りたい部分はグラフのどの部分なのか、要因をどのように表示すれば良いかを明らかにしている。

2.2 地理情報を用いた時系列データの探索的分析の支援

ハンス・ロスリングらは Gapminder を提案している [4]。Gapminder はユーザが瞬時の視覚分析に用いるデータを簡単にアクセスしやすくすることを手助けする。Gapminder は、国ごとの石油生産や国民一人あたりの収入、平均余命など様々な統計データを保持している。図 1 にシステムの外観を示す。ユーザはマッピングする統計データをグラフもしくは世界地図に切り替えることができる。グラフではユーザはこれらのデータの値の変化を 5 次元で一度に閲覧し、考察できる。5 次元は、x 軸、y 軸、時間、プロットサイズ、色で表現されている。ユーザはこのシステムを使うことで、国民一人あたりの収入と平均余命がどのような関係にあるのか調べることができる。

井上らは、感染症の調査データをよりわかりやすくするために WEB ベースのデータ可視化システムを提案している [5]。日本の感染症のサーベイランスレポートでは、全国の感染症の罹患率の傾向や周辺地域間の伝染を把握することが困難であることを解決するために、このシステムを開発している。システムは医療従事者が感染症の罹患率の傾向を素早く突き止め、一般住民に迫り来る病気の発生を警告することを手助けす

る。システムのデータベースは、2000 年から 2012 年までの 47 都道府県における 11 の小児科の感染症とインフルエンザの週報を含んでいる。それは約 300000 件に達する。

2.3 本研究の位置付け

時系列データの経時的変化と地理的な影響や規模の把握を容易にする可視化システムは提案されている [4] [5] が、時系列データの変化の理由や背景を知る用途には適していない。また、時系列データの経時的変化とその変化の理由や背景の把握を容易にする可視化システムは提案されている [3] が、地理的な影響や規模を知る用途には適していない。

本研究ではこれらのことを踏まえて、時系列データの経時的変化の把握を容易にするために折れ線グラフを、変化の理由や背景を知るために新聞記事を、地理的な関係性を把握するために地図を各々用いて、これらの異なるモダリティの情報にアクセスする方法を検討する。

3 システムの全体像とこれまでの取り組み

図 2 に本システムの目指す構成を示す。現状のシステムは新聞記事 DB、統計 DB、地図 DB を人手で作成している。しかし、人手での作成は、効率性や網羅性、リアルタイム性の点で問題がある。そのため、システムに用いるデータは WEB から抽出することを考えている。新聞記事 DB には、クローラを用いてあるトピックに関する記事を収集し、スクレイピング技術を用いて本文や見出しを抽出する。また、抽出された見出しや本文から自然言語処理技術を用いて日付や国名、出来事に関する文を抽出する。統計 DB や地図 DB は、オープンデータとの連携を検討している。

3.1 システムの全体像

図 3 に理想のシステムの全体像を示す。システムは、トピックを選択する選択ボックス (図 3-I)、地図を表示する地図ペイン (図 3-A) とグラフを表示するグラフペイン (図 3-B)、記事を表示する記事ペイン (図 3-C) で構成されている。選択ボックスには、「エボラ出血熱」や「台風」といったトピックがプルダウン形式で表示される。ユーザがトピックを選択すると、そのトピックに関連する統計量がグラフとしてグラフペインに描画され、同時に地図ペインには関連する地図を、記事ペインには関連する記事が表示される。また、地図とグ

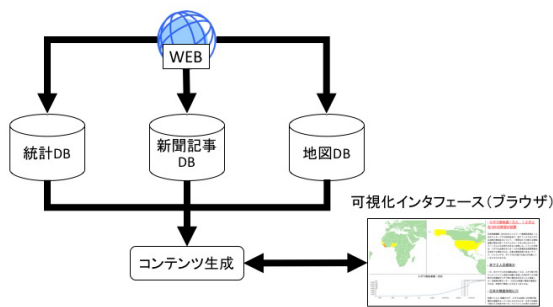


図 2: 実現を目指すシステムの構成 (文献 [6] より図引用)

ラフ上には、記事の有無を表すアノテーションとしてアイコンが付与されている。アイコンは文献 [7] を参考に、5 種類 (理由、背景、状況、出典、その他) を考えている。ユーザがアイコンをクリックすると、そのアイコンに対応する記事が表示される。さらに、グラフペインには、複数の統計グラフと凡例が提示されるようになっている。ユーザは凡例の左にあるチェックボックスに比較したい国を選択することでグラフペインに複数の統計グラフを描画することができる。これにより、ユーザは興味を持った国同士の統計量の変化を比較することで他国からの影響や規模を詳しく知ることができる。また、下のグラフを操作することで、上のグラフの表示する期間を変更したり、拡大や縮小を行うことができる。

現状のシステムには、地図上のアイコンや複数の統計グラフを描画する機能は実装されていないが、将来的にはこれらの機能を全て実装するつもりである。

3.2 これまでの取り組み

著者らは時系列データの経時的变化とその変化の要因の把握を可能にする可視化インタフェースを提案している [1]。グラフペインには、ユーザによって選択された日付を表示する青い線 (図 3-①) と記事の有無を表すアノテーションが表示されている。ユーザは青い線を左右にドラッグして日付を選択することができる。日付を選択すると、その時点の統計量が地図にマッピングされる。また、青い線をアノテーションに重ねると、その時点で起こった出来事について記述された記事がハイライトされる。さらに、地図にマッピングされた国をクリックすることができる。国をクリックすると、その国の統計量がグラフとしてグラフペインに描画される。同時に、その国名が含まれる記事がハイライトされる。加えて、記事をクリックすると、記事に含まれる日付がグラフ上に表示される。同時に、記事の本文に含まれている国がハイライトされる。

4 デザイン指針

4.1 対象とするデータの特徴

新聞記事は、ある時期における出来事やその出来事が起こった原因、場所、統計量、その統計量の具体的な値への言及や予測、記者の意見などが書かれている。そのため、出来事が起こった理由や背景を理解する上で有用である。しかし、新聞記事に書かれている統計量の値は近似値が用いられているため正確でなかったり、記者の観点で纏められていたりするため、客観性に欠ける。

統計データは、ある観測された場所において、ある時点の事象について測定された値である。例えば、人口統計や外国為替相場の推移などがこれに当たる。これらの統計データの値は、政府や国際連合の専門機関などが実施している厳密な環境において観測されていたり、センサを用いて取得される。そのため、これらのデータは正確である。また、折れ線グラフとして描画することでデータの変化や概要を容易に把握することができる。

地図は地球や地表、架空の世界の全部もしくは一部を平面上に縮尺表現したものである。例えば、地球全体もしくは大部分を表現している世界地図や統計データを地図上に表した統計地図などがこれに当たる。地図は地理的な位置関係や方向、距離、面積、形、高さを知る上で有用である。また、ある時点の統計量を地図にマッピングすることで、出来事の規模や地理的な広がりを把握できるといった特徴を持っている。

これらの情報はそれぞれ単体でも用いることができるが、ユーザの興味となる要素をトリガとしてインタラクティブに情報を提示することで、円滑な情報アクセスが可能なインタフェースを実現できると考えている。

4.2 データの相補的利用

図 4 に、本研究で対象とするデータの関係性を示す。地図はグラフの興味を持った時点の統計量をマッピングすることで、その事象の地理的な広がりを把握したり (図 4-①)、地理的な広がりの変化を見て興味を持った地域の統計グラフを参照することで、その地域の統計量の時間的な変化を把握したりできる (図 4-②)。また、そのような地理的な広がりの変化を見て興味を持った地域の新聞記事を参照することで、その変化の要因となる理由や背景を把握したり (図 4-③)、記事に言及されている統計量に対応するグラフを参照することで、その統計量の概況を把握したりできる (図 4-④)。さらに、記事で言及されている事象 (e.g., エボラ熱の患者数) に興味を持って、その事象に関する地理的影響や規模を把握したり (図 4-⑤)、グラフの特徴的な箇所の新

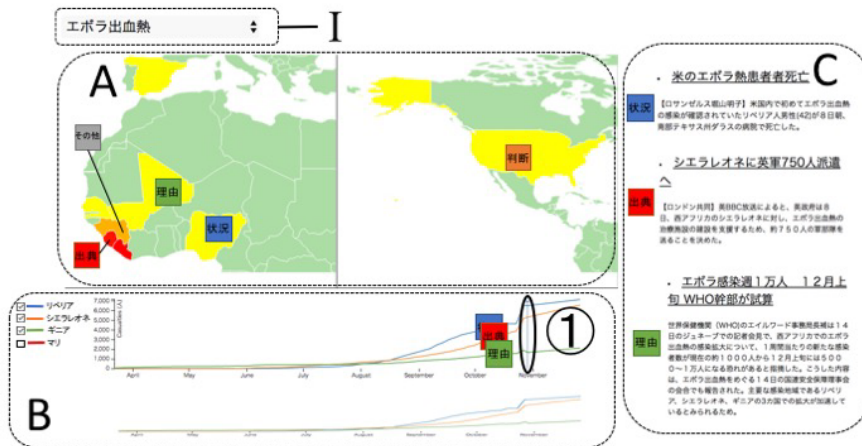


図 3: 理想のシステムの全体像

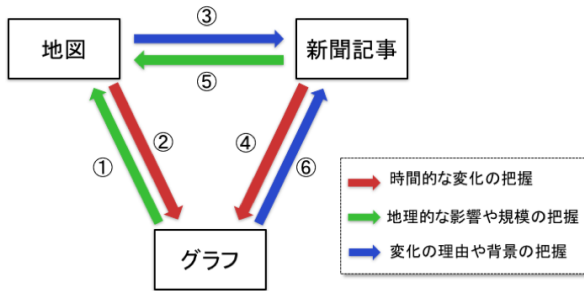


図 4: 対象とするデータの関係性 (文献 [6] より図引用)

聞記事を見ることでその変化の理由や背景を知ることができる (図 4-⑥)。本研究ではこのような情報アクセスを行えるようにシステムの機能をデザインした。これによって、ユーザの探索行為を円滑にする。

5 時系列データの探索的分析を支援する可視化システム

5.1 システムの構成

提案システムの構成を図 5 に示す。システムは、新聞記事と統計データ、地図の 3 つのデータベースとモジュール、入力判断部、コンテンツ生成部、画面生成部から構成されている。

システムはユーザの入力に応じて、これらのデータベースに格納されているデータを可視化表現 (e.g., グラフ, 地図, 記事) として生成する。ユーザの入力は、(1) 記事をクリックする, (2) グラフの日付を選択する, (3) 地図上の都道府県をクリックする, といった

直接操作を対象とする。次に、システムが実行する処理の流れを説明する。

まず、システムはユーザからの入力を入力判断部で判断する。入力判断部は、ユーザがどのような操作を行ったのかを解釈し、その結果をコンテンツ生成部へ伝える。コンテンツ生成部は、入力判断部から送られる結果をもとに、各モジュールに処理を伝達する。伝達を受けたモジュールは、データベースから必要なデータを受け取り、可視化表現を生成する。

地図モジュールでは、(1) 地理情報をもとに地図データベースから位置情報を取得し、地図を生成する、(2) 地図上の都道府県をハイライトする、という 2 つの処理を行う。都道府県のハイライトでは、あらかじめ統計モジュールから引き渡された統計量の最小値と最大値に基づいて、統計量を 0 から 255 に変換し、それを色の RGB 値に用いる。地図データベースは、地理情報とそれに関する位置情報を格納している。

記事モジュールでは、(1) トピックをもとに記事データベースから本文とその見出しを取り出し、記事のスニペットを生成する、(2) 記事のスニペットをハイライトする、という 2 つの処理を行う。記事データベースは、トピックごとに分類された新聞記事データを格納している。新聞記事データは本文とその見出し、記事の発行日を抽出したデータである。

統計モジュールでは、(1) 統計情報をもとに、統計データベースから統計量を取り出し、グラフとして描画する、(2) グラフ上の日付をハイライトする、という 2 つの処理を行う。また、グラフを描画する際に、記事へのアクセスが可能なトリガとしてアノテーションを付与する。統計データベースは、地域ごとの統計量を時系列順に格納している。

このように、各モジュールでは、(1) 可視化表現を

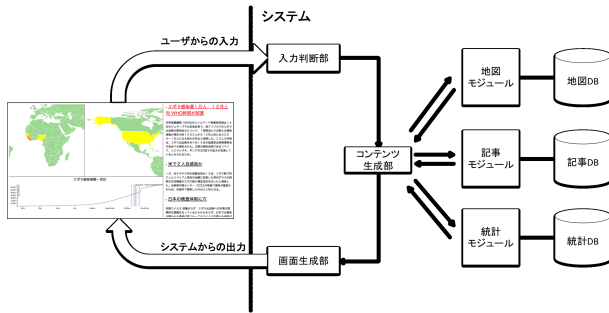


図 5: 提案システムの構成図 (文献 [1] より図引用)

生成する, (2) その可視化表現の一部をハイライトする, という 2 つの処理を行っている。

また, 各モジュールはコンテンツ生成部を介して, 他のモジュールと連携している。

記事モジュールは, データがタグ付けられた記事のスニペットから, コンテンツ生成部を介して, 日付を統計モジュールへ引き渡す。これは, グラフの描画やアノテーションの付与, 日付のハイライト, 都道府県のハイライトに用いられる。

統計モジュールでは, グラフ上のアノテーションの付与を行うために, 統計データの値の補間を行っている。本システムでは, 記事の有無を示すアノテーションとして円を折れ線グラフ上に配置している。この円を折れ線グラフ上に配置するためには, 記事が持つ日付に当たる統計データの値が必要になる。しかし, 統計データは, 1 週間毎に観測されたデータやところどころ抜けているデータがほとんどである。そのため, これらのデータの値を線形補間することで, 折れ線グラフ上に円を配置した。

統計モジュールは, コンテンツ生成部を介して, 統計量とその最大値と最小値, その地域名を地図モジュールへ引き渡す。これは, 地図のマッピングに用いられる。また, 日付情報をコンテンツ生成部を介して記事モジュールへ引き渡す。これは, 記事のハイライトに用いられる。

地図モジュールは, コンテンツ生成部を介して, 地理情報を記事モジュールや統計モジュールに引き渡す。これは, 記事のハイライトやグラフの描画に用いられる。

コンテンツ生成部は, 記事モジュールから引き渡された統計量とそれに関する日付を統計モジュールへ引き渡すといった処理を行う。このように, コンテンツ生成部と各モジュールがデータの受け渡しを繰り返すことにより, 可視化表現を生成する。そして, 生成された可視化表現は, コンテンツ生成部によって集められ, 画面生成部へ処理を移行する。画面生成部では, コンテンツ生成部で生成された可視化表現をユーザに提示する。

表 1: 統計データの例

日付	合計	岩手県	宮城県	福島県	その他
2011/3/18	6495	3370	1926	1139	60
2011/3/22	9415	3370	4846	1139	60
2011/3/31	11310	3370	6741	1139	60
...	...	...	...	...	...
2011/7/29	15645	4615	9364	1600	66
2011/3/18	15735	4649	9417	1603	66

5.2 対象データ

本システムでは, 統計データとして農林水産省の果樹生産出荷統計のみかんの出荷量¹と土地価格², 東日本大震災における死者の推移に関するデータ³を用いた。また, 統計データを都道府県別かつ時系列順に整理し, csv 形式にした。

統計データの例を表 1 に示す。東日本大震災の統計データには, 東日本大震災における死者・行方不明者の推移の wikipedia から死者の推移の統計データを人手で抽出したものを用いた。土地価格には, 土地価格が分かる土地代データという Web サイトから各都道府県の公示地価, 基準地価の総平均値を用いた。みかんの出荷量には, 農林水産省に掲載されている平成 15 年から 27 年までの果樹生産出荷統計の都道府県別の結果樹面積・10a 当たり収量・収穫量・出荷量のみかんの csv データの出荷量を用いた。

新聞記事データは, 野菜・果物と地価, 東日本大震災のニュース記事である。地価と東日本大震災のニュース記事には, Yahoo!ニュースの地価⁴と東日本大震災に関連するアーカイブ一覧⁵を用いた。アーカイブ一覧から各記事の見出しと本文, 記事の発行日を抽出し, csv 形式で保存するプログラムを作成した。記事の発行日は, 「年/月/日」のとなるようにフォーマットを統一した。

野菜・果物のニュース記事には, 農業協同組合新聞の野菜・果物のニュース⁶を用いた。ニュース一覧から 2013 年 1 月 7 日から 2016 年 12 月 1 日までの記事を前述したプログラムを用いて抽出した。

5.3 提案システムの概要

提案システムを図 6 に示す。選択ボックス (図 6-I) には, システムが提示しているトピック名が表示されており, ユーザは選択ボックスを操作することで分析

¹<http://www.maff.go.jp/j/tokei/kouhyou/sakumotu/>

²<http://www.tochidai.info/>

³<https://ja.wikipedia.org/wiki/東日本大震災における死者・行方不明者の推移>

⁴http://news.yahoo.co.jp/list/?t=2011sanrikuoki_eq

⁵http://news.yahoo.co.jp/list?t=price_of_land

⁶<http://www.jacom.or.jp/yasai/news/>

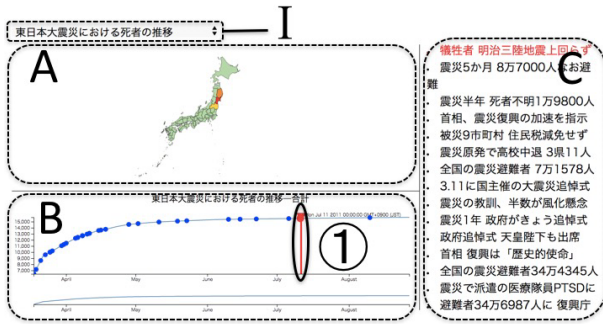


図 6: 提案システム

したいトピックを選択することができる。グラフペイン (図 6-B) には、2つのグラフが表示されている。上部のグラフ上には、赤色のスライダーと青色の円 (図 6-①) が表示されている。スライダーはマウスによる直接操作が可能であり、左右にドラッグすることで日付を選択し、それに併せて地図上にマッピングする統計データを切り替えることができる。青色の円は、その時点に記事があることを示している。スライダーがこの円に重なるとその時点のニュース記事がスクロールされ、スクロールされた記事と重なった円が赤くハイライトされる。また、下部の小さいグラフを操作することで上部のグラフの表示する期間を変更したり、拡大・縮小を行うことができる。記事ペイン (図 6-C) に表示されている記事はクリックすることが可能であり、クリックすると、その記事の本文を表示することができる。同時に、選択された記事に該当するグラフ上の円が赤くハイライトされる。

5.4 実装概要

5.2節で述べたプログラムの実装には、Python の version 3.5.2 と Python のライブラリとして BeautifulSoup の version 4.5.1 を用いた。5.3節で述べたシステムは、ウェブアプリケーションになっており、作成には HTML と CSS, JavaScript を用いた。また、JavaScript のライブラリとして、D3.js⁷ の version 3.4.13 と TopoJSON の version 1.6.18 を用いた。

6 実験

実験では、提案システムの問題点と改善点を明らかにするために、先行システムと提案システムとのユーザビリティの比較を行った。実験参加者には、シナリオの課題を達成するようにシステムを操作してもらい、アンケートに回答してもらった。これらの手順を1つ

表 2: 実験参加者と課題の対応表

実験参加者	先行システム	提案システム
A	東日本大震災	土地価格
B	みかんの出荷統計	東日本大震災
C	土地価格	みかんの出荷統計
D	東日本大震災	土地価格
E	みかんの出荷統計	東日本大震災
F	土地価格	みかんの出荷統計
G	土地価格	東日本大震災
H	土地価格	東日本大震災
I	みかんの出荷統計	土地価格
J	みかんの出荷統計	土地価格
K	東日本大震災	みかんの出荷統計
L	東日本大震災	みかんの出荷統計

のシステムで2回行った。実験参加者には、提案システムで2回、先行システムで2回の計4回アンケートに回答してもらった。シナリオ1はグラフから記事へアクセスする機能を、シナリオ2は記事からグラフへアクセスする機能を用いることが想定されている。アンケートの項目は、(1) シナリオの目的を達成できたか、(2) 機能は十分だったか、(3) 使う時に迷ったことなど不満な点はありませんでしたか、(4) こうした方が良いというような意見があれば教えてください、の4つである。項目1と2には、5段階のリッカート尺度を用いた。

実験参加者は筆者らが所属している研究室の大学生12人 (男性10人、女性2人) である。実験のシステムに用いたトピックは、「東日本大震災」と「土地価格」、「みかんの出荷統計」の3種類である。実験では、実験参加者が2つのシステムにおいて必ず異なるトピックを用いるようにした。また、2つのシステムの順序効果を統制した。実験参加者と実験に用いたシステムのトピックとの対応表を表2に示す。

アンケート結果のシナリオ1の目的の達成度を表3に、シナリオ2の目的の達成度を表4に示す。シナリオ1におけるシステムの機能の満足度を表5に、シナリオ2におけるシステムの機能の満足度を表6に示す。各システムのシナリオの目的の達成度と機能の満足度の平均スコアから、シナリオ1の目的の達成度を除いて提案システムのスコアが先行システムのスコアを上回っていることがわかる。シナリオ1において先行システムのスコアが提案システムのスコアを上回った要因としては、提案システムと比べて記事が見づらいこととグラフと記事との対応付けがわかりづらかったことが挙げられる。記事が見づらいことについては、先

⁷<https://d3js.org/>

表 3: シナリオ 1 の目的の達成度

実験参加者	先行システム	提案システム
A	4	5
B	2	4
C	4	2
D	4	4
E	4	4
F	5	4
G	5	3
H	4	2
I	4	4
J	4	5
K	4	4
L	4	4
平均	4.0	3.8

表 4: シナリオ 2 の目的の達成度

実験参加者	先行システム	提案システム
A	5	3
B	2	5
C	4	5
D	2	2
E	1	4
F	4	5
G	5	4
H	3	3
I	4	2
J	5	5
K	4	5
L	5	5
平均	3.7	4.0

行システムでは記事と記事の間隔が空いていたものの、提案システムではその間隔が空いていなかったため、記事が見つからなかったと 1 人の実験参加者が述べていた。グラフと記事との対応付けのわかりづらさについては、1 人の実験参加者がグラフから記事にアクセスするために円をクリックしたり、記事の発行日を見るために、グラフの表示期間外の記事をクリックしたり、していた。

7 議論

グラフから記事へアクセスする機能では、7 人の実験参加者が記事を見ようとしてグラフ上の円をクリックしていた。このことから、操作がわかりづらいことが明らかになった。この問題は、青い丸をクリックするとその記事にアクセスできる機能を設けることで解決できると考えている。

記事からグラフへアクセスする機能について、5 人の実験参加者から本機能を用いて記事がグラフのどの部分にあたるのか把握しようとする発話が確認された。

日本地図について、4 人の実験参加者から地図が小さいというコメントがあった。これを解決するためには、地図を拡大したり、縮小したりする機能が望ましい。

記事の選定について、提案システムでは、スクレイピングした記事をすべて用いている。そのため、提示している時系列データに関係がない記事や提示している時系列データの期間とは異なる記事も含まれている。また、時系列データの分析においてユーザは特徴的な変化に興味を持ち、その変化の原因を調べるために変化した付近の記事を参照する。このようなユーザにとって、変化した付近以外の記事や原因に関係がない記事

を参照することは時間や手間を要する。そのため、このような時系列データを分析するユーザにとってはグラフの特徴的な変化している時点付近に、グラフの変化について述べられている記事が対応付けられていることが望ましい。提案システムでは、このような対応付けを実現する方式について検討する必要がある。

新聞記事と時系列データのアラインメントについて、提案システムにおける記事と時系列データの対応付けは、記事の発行日を用いて行っている。しかし、記事には過去のある時点のデータの変化や過去の出来事について言及されているものが多くある。そのため、このような記事では記事の発行日を用いて対応付けるのではなく、記事の文章を読み取り、適切な箇所に対応付けることが望ましい。このような取り組みとして、著者らは記事と時系列データをアラインメントする手法について検討している [6]。この手法では、記事中の日付と値の変動に関する曖昧表現の 2 つに着目し、時系列データの該当箇所を特定している。この手法を提案システムに組み込むことで、より適切な記事と時系列データのアラインメントが可能となる。

8 おわりに

本研究の目的は、ユーザの興味や関心に応じて様々なモダリティの情報へのアクセスを繰り返しつつ時系列データを分析するための支援システムの実現である。著者らはこれまでに、ユーザが時系列データの経時的変化とその変化の要因の把握を可能にする可視化インタフェースを提案してきた [1]。先行インタフェースでは、(1) 容易に興味を持った時点の記事を参照するこ

表 5: シナリオ 1 におけるシステムの機能の満足度

実験参加者	先行システム	提案システム
A	4	5
B	2	4
C	4	3
D	3	4
E	3	4
F	3	5
G	4	4
H	4	2
I	4	4
J	2	5
K	4	4
L	4	4
平均	3.4	4.0

表 6: シナリオ 2 におけるシステムの機能の満足度

実験参加者	先行システム	提案システム
A	4	3
B	3	4
C	2	5
D	3	2
E	3	5
F	3	4
G	4	3
H	3	3
I	3	4
J	4	5
K	4	4
L	5	4
平均	3.4	3.8

とができない, (2) データの推移を把握しづらい, という問題があったため, 本稿では, (1) グラフと記事との連携の強化, (2) グラフのインタラクションの追加, によりユーザの探索行為の円滑化を図った. 加えて, ユーザ実験を行い, システムの到達点と今後の拡張の方向性について検討した. 今後は, (1) 記事と地図との連携の強化, (2) 分析に必要なデータの選定, を行うことでより効率的に時系列データを分析できるようにする.

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] Naito, S., Matsushita, M.: Supporting Consecutive Data Exploration by Visualizing Spatiotemporal Trend Information, in *Proceedings of the 2015 Conference on Technologies and Applications of Artificial Intelligence*, pp. 227–231 (2015)
- [2] 松下光範, 加藤恒昭: 数値情報の補填とグラフ概形の示唆による複数文書からの統計グラフ生成, 知能と情報, Vol. 18, No. 5, pp. 721–734 (2006)
- [3] 山本健一, 殿井加代子, 谷岡広樹: タグ付きコーパスを用いた動向情報とその要因の可視化, 言語処

理学会第 12 回年次大会ワークショップ「言語処理と情報可視化の接点」(2006)

- [4] Rosling, H.: Gapminder, (2011) <http://www.gapminder.org> (2017/2/19 確認)
- [5] Inoue, M., Hasegawa, S., Suyama, A., Kakehashi.: Development of a Web-based Data Visualization System for Comprehensible Ascertainment of the Spatiotemporal Extent of Infectious Diseases, 医療情報学, Vol. 33, No. 1, pp. 27–32 (2013)
- [6] 内藤峻, 古田遼樹, 松下光範: 時系列データの探索的分析を支援する可視化システム: 記事と時系列データのアラインメント方式の提案, 人工知能学会第 13 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. 1–6 (2016)
- [7] 松下光範, 加藤恒昭: 言語情報と数値情報の相補的利用を目指した可視化手法, 第 21 回人工知能学会全国大会, 3H8-3 (2007)

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 Nishikawa¹ Yukihiro Moriyama¹ Shun Naito¹ Mitsunori Matsushita¹

¹ 関西大学総合情報学部
Faculty of Informatics, Kansai University
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 appeared 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话题 (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を示唆される。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。

－ Phase 1 －

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。

－ Phase 2 －

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、被災者の苦勞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。

－ Phase 3 －

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。

－ Phase 4 －

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MeCab⁴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法を用いて各記事の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

表 2: TF-IDF 法の結果の一例

熊本地震	東京五輪	三菱自動車
搜索活動	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

⁴<https://sourceforge.net/projects/mecab/> (2016/12/12 確認)

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

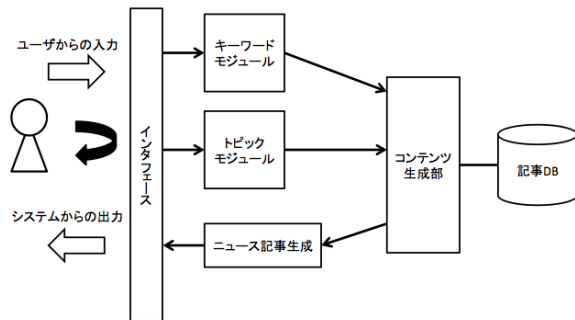


図 1: システム構成図

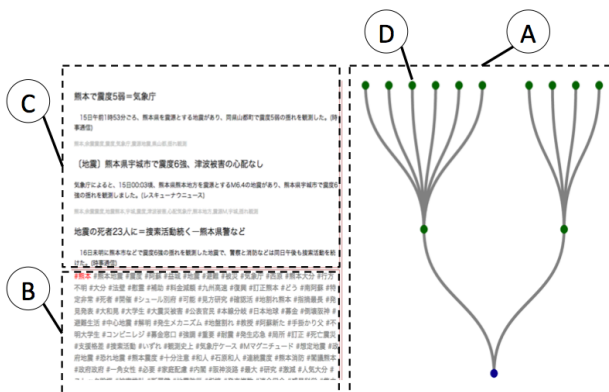


図 2: プロトタイプのスクリンショット

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生成する方法, (b) キーワードに関するニュース記事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在籍学生 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

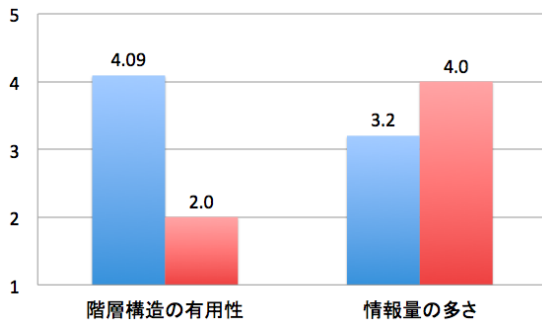


図 3: 5段階評価の平均点

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).